ECMA 31380 - Causal Machine Learning - Homework 1

Fernando Rocha Urbano

Autumn 2024

---

1 Identification of Variance in Random Experiments

In class we discussed how the assumptions of randomized treatment assignment, SUTVA, and consistency were sufficient to identify average treatment effects such as $\tau = \mathbb{E}[Y(1) - Y(0)]$ and conditional averages like $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$.

1.a Not identified variance of individual treatment

Prove that the variance of the individual treatment effects is not identified. That is, show that even though $\mathbb{E}[Y(1) - Y(0)]$ is identified, $\mathbb{V}[Y(1) - Y(0)]$ is not. Intuitively explain the source of the identification problem.

We aim to prove that the variance of the individual treatment effects $\mathbb{V}[Y(1) - Y(0)]$ is not identified under randomized treatment assignment, SUTVA, and consistency.

In a informal way, we can think that the variance is not identified because we never observe each individual at both states. Therefore, we cannot know for each individual the counterfactual. Below, we derive a formal solution.

Under the given assumptions, we can identify the average potential outcomes:

$$\mathbb{E}[Y(1)] = \mathbb{E}[Y \mid T = 1], \quad \mathbb{E}[Y(0)] = \mathbb{E}[Y \mid T = 0],$$

where T is the treatment indicator. Consequently, the average treatment effect (ATE) is identified:

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)].$$

The variance of the individual treatment effects is:

$$\mathbb{V}[Y(1) - Y(0)] = \mathbb{V}[Y(1)] + \mathbb{V}[Y(0)] - 2\,\mathbb{COV}[Y(1), Y(0)].$$

We can identify the variances $\mathbb{V}[Y(1)]$ and $\mathbb{V}[Y(0)]$ since:

$$\mathbb{V}[Y(1)] = \mathbb{V}[Y \mid T = 1], \quad \mathbb{V}[Y(0)] = \mathbb{V}[Y \mid T = 0].$$

However, the covariance $\mathbb{COV}[Y(1), Y(0)]$ is not identified. Again, this is because we never observe both $Y(1)$ and $Y(0)$ for the same individual; we only observe one potential outcome depending on the treatment assignment.

Assume, for contradiction, that $\mathbb{COV}[Y(1), Y(0)]$ is identified. Then, $\mathbb{V}[Y(1) - Y(0)]$ would be identified as well. However, consider two hypothetical scenarios with the same marginal distributions of $Y(1)$ and $Y(0)$ but different covariances:

Independent Potential Outcomes:

$$\mathbb{COV}[Y(1), Y(0)] = 0.$$

Thus,

$$\mathbb{V}[Y(1) - Y(0)] = \mathbb{V}[Y(1)] + \mathbb{V}[Y(0)].$$

Perfect Positive Correlation:

$$Y(1) = Y(0) + c,$$

for some constant c, implying

$$\mathbb{COV}[Y(1), Y(0)] = \mathbb{V}[Y(0)].$$

Therefore,

$$\mathbb{V}[Y(1) - Y(0)] = \mathbb{V}[Y(1)] + \mathbb{V}[Y(0)] - 2\,\mathbb{V}[Y(0)] = \mathbb{V}[Y(1)] - \mathbb{V}[Y(0)].$$

In both scenarios, the marginal distributions $\mathbb{V}[Y(1)]$ and $\mathbb{V}[Y(0)]$ are the same, but $\mathbb{V}[Y(1) - Y(0)]$ differs due to different covariances.

The core of the identification problem lies in the fundamental unobservability of individual-level treatment effects.

1.b Identified variance of conditional average treatment

In contrast, prove that the variance of the conditional average treatment effect, i.e., $\mathbb{V}[\tau(X)]$ is identified.

Under the assumptions of randomized treatment assignment, SUTVA (Stable Unit Treatment Value Assumption), and consistency, we can identify the conditional average treatment effect (CATE) $\tau(X)$ for each value of covariates X. The conditional average treatment effect is defined as:

$$\tau(X) = \mathbb{E}[Y(1) - Y(0) \mid X] = \mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X].$$

Since treatment is randomly assigned, we have:

$$\mathbb{E}[Y(1) \mid X] = \mathbb{E}[Y \mid T = 1, X], \quad \mathbb{E}[Y(0) \mid X] = \mathbb{E}[Y \mid T = 0, X].$$

Thus, $\tau(X)$ is identified from observable data:

$$\tau(X) = \mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X].$$

Now, the variance of $\tau(X)$ is:

$$\mathbb{V}[\tau(X)] = \mathbb{E}\left[\tau(X)^2\right] - \left(\mathbb{E}\left[\tau(X)\right]\right)^2.$$

To identify this variance, we need to show that both $\mathbb{E}\left[\tau(X)\right]$ and $\mathbb{E}\left[\tau(X)^2\right]$ are identifiable.
Since $\tau(X)$ is identified for each X, and X has a known distribution $f_X(x)$, we can compute:

$$\mathbb{E}\left[\tau(X)\right] = \int \tau(x) f_X(x)\, dx.$$

Similarly, we can compute:

$$\mathbb{E}\left[\tau(X)^2\right] = \int \tau(x)^2 f_X(x)\, dx.$$

Since both $\mathbb{E}\left[\tau(X)\right]$ and $\mathbb{E}\left[\tau(X)^2\right]$ are identified from the observed data, the variance $\mathbb{V}[\tau(X)]$ is also identified:

$$\mathbb{V}[\tau(X)] = \int \tau(x)^2 f_X(x)\, dx - \left(\int \tau(x) f_X(x)\, dx\right)^2.$$

In the discrete case where $X, T, Y \in \{0, 1\}$:

$$\tau(0) = \mathbb{E}[Y|T = 1, X = 0] - \mathbb{E}[Y|T = 0, X = 0]$$

$$\tau(1) = \mathbb{E}[Y|T = 1, X = 1] - \mathbb{E}[Y|T = 0, X = 1]$$

The respective variances are:

$$\begin{aligned}\mathbb{V}[\tau(0)] =& \mathbb{V}[Y|T = 1, X = 0] + \mathbb{V}[Y|T = 0, X = 0] \\ & - \mathbb{COV}([Y|T = 1, X = 0], [Y|T = 0, X = 0])\end{aligned}$$

$$\begin{aligned}\mathbb{V}[\tau(1)] =& \mathbb{V}[Y|T = 1, X = 1] + \mathbb{V}[Y|T = 0, X = 1] \\ & - \mathbb{COV}([Y|T = 1, X = 1], [Y|T = 0, X = 1])\end{aligned}$$

In these, all terms are identifiable.

In conclusion, the variance of the individual treatment effects $\mathbb{V}[Y(1) - Y(0)]$ is not identifiable because we cannot observe both $Y(1)$ and $Y(0)$ for the same individual. However, $\tau(X)$ is the average treatment effect conditional on covariates X, which can be identified because we observe both treated and control outcomes for each subpopulation defined by X.

1.c Real world context

Explain in a real world context why a decision would want to know $\mathbb{V}[Y(1) - Y(0)]$ and separately $\mathbb{V}[\tau(X)]$. For each $\mathbb{V}[Y(1) - Y(0)]$ and $\mathbb{V}[\tau(X)]$, explain how you would use this variance in research and in decision making.

In real-world decision-making and research, understanding not just the average effect of a treatment but also the variability of its effects across individuals or subpopulations is crucial. The variances $\mathbb{V}[Y(1) - Y(0)]$ and $\mathbb{V}[\tau(X)]$ capture different aspects of this heterogeneity.

Variance of Individual Treatment Effects: $\mathbb{V}[Y(1) - Y(0)]$

The variance $\mathbb{V}[Y(1) - Y(0)]$ measures the heterogeneity of the treatment effect at the individual level. In practical terms, it quantifies how much the effect of a treatment varies from one individual to another.

A high variance indicates that while some individuals may benefit greatly from the treatment, others may not benefit at all or may even be harmed.

- In healthcare, knowing this variance can guide decisions on whether to adopt a one-size-fits-all treatment approach or to tailor treatments to individual patients.
- For policymakers, by acknowledging the variability, strategies can be developed to mitigate risks for individuals who might experience adverse effects.
- High variability shows researchers that underlying factors may be causing differential treatment effects. In such case, if the research has not been done with the use of covariates, adding them can be benefitial to understand CATE.
- In clinical settings, patients can be better informed about the potential range of outcomes.

Variance of the Conditional Average Treatment Effect: $\mathbb{V}[\tau(X)]$

The variance $\mathbb{V}[\tau(X)]$ captures the variability of treatment effects inside each subpopulations defined by covariates X.

- Identifies which subgroups benefit in a more homogeneous way from the treatment.
- Identifies which subgroups should have further division due to relative big variance. Those groups can guide which new covariates should be added in a next study to improve efficiency of the estimates. In a healthcare example, acknowledging that CATE for young people is high can serve as a suggestion to collect data regarding weekly exercise hours as a covariate.
- Helps in optimizing the allocation of resources by focusing on groups that have high CATE but also small CATE variance. For instance, the subpopulation with the highest ATE (which is CATE) but big variance in the treatment effect might not be the ideal target of the treatment due to unexpected effects in the "fatter" tail of the conditional distribution.

Real-World Example: Healthcare Intervention

A new medication is developed to lower blood pressure.

Researchers want to know how blood pressure reduction varies among patients. A high variance may indicate that while some patients experience significant reductions, others see little to no effect (or even a negative effect). This relate to $\mathbb{V}[Y(1) - Y(0)]$.

Researchers can understand to who they should target the medication by investigating genetic or lifestyle factors contributing to the variability. With that, they can adjust clinical guidelines to account for patient-specific responses (check where CATE variance is low vs. high).

Finally, they can decide whether to prescribe the medication universally or only to patients in subgroups where the majority is likely to benefit.

Conclusion

Understanding both variances enhances the effectiveness of interventions by accounting for variability at both the individual and group levels. While $\mathbb{V}[Y(1) - Y(0)]$ remains theoretically important despite being unidentifiable, $\mathbb{V}[\tau(X)]$ offers practical insights that can be directly applied to improve outcomes in real-world settings.

For $\mathbb{V}[Y(1) - Y(0)]$:

$$\mathbb{V}[Y(1) - Y(0)] = \mathbb{E}\left[(Y(1) - Y(0) - \tau)^2\right], \quad \text{where } \tau = \mathbb{E}[Y(1) - Y(0)].$$

For $\mathbb{V}[\tau(X)]$:

$$\mathbb{V}[\tau(X)] = \mathbb{E}\left[(\tau(X) - \mathbb{E}[\tau(X)])^2\right].$$

2 Linear Regression in Randomized Experiments

Assume that $(y_i, x_i, t_i), i = 1, \ldots, n$ is an iid sample from $(Y, X, T) \in \mathbb{R}^2 \times \{0, 1\}$ (having a vector of covariates changes nothing but notation). Further assume this is an ideal randomized experiment in the sense that (i) T is randomized such that $\mathbb{P}[T = 1 \mid X = x] = p$, which is bounded inside $(0, 1)$, (ii) x_i is realized prior to randomization, (iii) SUTVA and consistency hold.

2.a Potential outcomes and treatment effect

Using potential outcomes notation and specific assumptions to prove that: (i) without loss of generality we can write $Y = Y_1 T + Y_0(1 - T) = \alpha + \beta T + \varepsilon$, (ii) the average treatment effect obeys $\tau := \mathbb{E}[Y_1 - Y_0] = \beta$, and (iii) the residuals ε are mean zero given the regressor.

We can use potential outcomes notation, where $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes for unit i under treatment and control, respectively. The observed outcome becomes:

$$Y = Y(1)T + Y(0)(1 - T)$$

This is possible because with the randomized experiment, SUTVA, overlap, and pre-treatment covariates:

$$\mathbb{E}[Y|T = 1] = Y(1)$$

$$\mathbb{E}[Y|T = 0] = Y(0)$$

Using the equalities stated above:

(i) Representation of Y as a Linear Model

We start by expressing Y in terms of its potential outcomes:

$$\begin{aligned} Y &= Y(1)T + Y(0)(1 - T) \\ &= Y(0) + (Y(1) - Y(0))\, T \end{aligned}$$

Defining:

- $\alpha := \mathbb{E}[Y(0)]$
- $\beta := \mathbb{E}[Y(1) - Y(0)]$

- $\varepsilon_i := (Y(0) - \mathbb{E}[Y(0)]) + (Y(1) - Y(0) - \mathbb{E}[Y(1) - Y(0)])\, T$

Adding and subtracting α, βT in the equation:

$$\begin{aligned}
Y &= Y(0) + (Y(1) - Y(0))\, T \\
&= \alpha + Y(0) + \beta T + (Y(1) - Y(0))\, T - \alpha - \beta T \\
&= \alpha + Y(0) + \beta T + (Y(1) - Y(0))\, T - \mathbb{E}[Y(0)] - \mathbb{E}[Y(1) - Y(0)] \\
&= \alpha + \beta T + Y(0) + (Y(1) - Y(0))\, T - \mathbb{E}[Y(0)] - \mathbb{E}[Y(1) - Y(0)]T \\
&= \alpha + \beta T + (Y(0) - \mathbb{E}[Y(0)]) + (Y(1) - Y(0) - \mathbb{E}[Y(1) - Y(0)])\, T \\
&= \alpha + \beta T + \varepsilon
\end{aligned}$$

Thus, without loss of generality, we can write $Y_i = \alpha + \beta T_i + \varepsilon_i$.
The same can be thought as defining the constants:

$$\alpha := \mathbb{E}[Y_i(0)], \quad \beta := \mathbb{E}[Y_i(1) - Y_i(0)]$$

Defining the residuals:

$$\varepsilon_i := \begin{cases} Y_i(0) - \alpha, & \text{if } T_i = 0, \\ Y_i(1) - (\alpha + \beta), & \text{if } T_i = 1. \end{cases}$$

Since T_i is independent of $Y_i(0)$ and $Y_i(1)$ due to randomization, ε_i is independent of T_i.
Now, express Y_i using these definitions:

$$\begin{aligned}
Y_i &= Y_i(1)T_i + Y_i(0)(1 - T_i) \\
&= [\alpha + \beta + \varepsilon_i]T_i + [\alpha + \varepsilon_i](1 - T_i) \\
&= \alpha + \beta T_i + \varepsilon_i.
\end{aligned}$$

This again shows that Y_i can indeed be written as:

$$Y_i = \alpha + \beta T_i + \varepsilon_i.$$

(ii) The Average Treatment Effect $\tau = \beta$
Using β and α:

$$\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\
&= (\alpha + \beta) - \alpha \\
&= \beta.
\end{aligned}$$

Thus, the average treatment effect τ equals β.
Also, by definition, we know that:

$$\tau := \mathbb{E}[Y(1) - Y(0)] = \beta.$$

This comes directly from our definition of β, so $\tau = \beta$.

(iii) Residuals are Mean Zero Given T

In other words, we need to show that $\mathbb{E}[\varepsilon \mid T] = 0$ for $T \in \{0, 1\}$.

Case 1: When $T = 0$

When $T = 0$, the residual simplifies to:

$$\varepsilon = Y(0) - \mathbb{E}[Y(0)].$$

Taking the conditional expectation:

$$\mathbb{E}[\varepsilon \mid T = 0] = \mathbb{E}[Y(0) - \mathbb{E}[Y(0)] \mid T = 0] = \mathbb{E}[Y(0) \mid T = 0] - \mathbb{E}[Y(0)].$$

Under randomization and since $Y(0)$ is independent of T:

$$\mathbb{E}[Y(0) \mid T = 0] = \mathbb{E}[Y(0)].$$

Therefore:

$$\mathbb{E}[\varepsilon \mid T = 0] = \mathbb{E}[Y(0)] - \mathbb{E}[Y(0)] = 0.$$

Case 2: When $T = 1$

When $T = 1$, the residual becomes:

$$\varepsilon = (Y(0) - \mathbb{E}[Y(0)]) + (Y(1) - Y(0) - \mathbb{E}[Y(1) - Y(0)]) .$$

Simplifying the residual:

$$\varepsilon = (Y(1) - \mathbb{E}[Y(1)]) .$$

Taking the conditional expectation:

$$\mathbb{E}[\varepsilon \mid T = 1] = \mathbb{E}[Y(1) - \mathbb{E}[Y(1)] \mid T = 1] = \mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(1)].$$

Again, due to randomization and independence:

$$\mathbb{E}[Y(1) \mid T = 1] = \mathbb{E}[Y(1)].$$

Therefore:

$$\mathbb{E}[\varepsilon \mid T = 1] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(1)] = 0.$$

Conclusion

Since $\mathbb{E}[\varepsilon \mid T = t] = 0$ for $t \in \{0, 1\}$, we have:

$$\mathbb{E}[\varepsilon \mid T] = 0.$$

Thus, the residuals ε are mean zero given the regressor T.

2.b Heteroskedastic variance of residuals

Show that the variance of ε is heteroskedastic in general. Under what conditions will it be homoskedastic?

From the previous result, we have:

$$Y_i = \alpha + \beta T_i + \varepsilon_i,$$

where $T_i \in \{0, 1\}$.

The residuals are defined by:

$$\varepsilon_i = (Y_i(0) - \mu_0) + (Y_i(1) - Y_i(0) - \tau)\, T_i,$$

where:

- $\mu_0 = \mathbb{E}[Y_i(0)]$,
- $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$.

Under the assumptions of randomization and SUTVA, $Y_i(0)$ and $Y_i(1)$ are independent of T_i.

We will compute $\mathbb{V}(\varepsilon_i \mid T_i = t)$ for $t = 0$ and $t = 1$.

Case 1: When $T_i = 0$

When $T_i = 0$, the residual simplifies to:

$$\varepsilon_i = Y_i(0) - \mu_0.$$

Thus, the conditional variance is:

$$\mathbb{V}(\varepsilon_i \mid T_i = 0) = \mathbb{V}(Y_i(0) - \mu_0) = \mathbb{V}(Y_i(0)) = \sigma_0^2,$$

where σ_0^2 is the variance of the potential outcomes under control.

Case 2: When $T_i = 1$

When $T_i = 1$, the residual becomes:

$$\varepsilon_i = (Y_i(0) - \mu_0) + (Y_i(1) - Y_i(0) - \tau)\,.$$

Simplify the residual:

$$\varepsilon_i = (Y_i(1) - \mu_0 - \tau)\,.$$

Note that $\mu_0 + \tau = \mathbb{E}[Y_i(0)] + \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i(1)] = \mu_1$.

Therefore, the residual simplifies further to:

$$\varepsilon_i = Y_i(1) - \mu_1.$$

Thus, the conditional variance is:

$$\mathbb{V}(\varepsilon_i \mid T_i = 1) = \mathbb{V}(Y_i(1) - \mu_1) = \mathbb{V}(Y_i(1)) = \sigma_1^2,$$

where σ_1^2 is the variance of the potential outcomes under treatment.

Conclusion on Heteroskedasticity

When $T_i = 0$:

$$\mathbb{V}(\varepsilon_i \mid T_i = 0) = \sigma_0^2.$$

When $T_i = 1$:

$$\mathbb{V}(\varepsilon_i \mid T_i = 1) = \sigma_1^2.$$

In general, σ_0^2 and σ_1^2 are not expected to be equal. Therefore, the variance of ε_i depends on the value of T_i, indicating heteroskedasticity.

In this case, the residuals ε_i exhibit heteroskedasticity.

Conditions for Homoskedasticity

The residuals ε_i will be homoskedastic if and only if:

$$\mathbb{V}(\varepsilon_i \mid T_i = 0) = \mathbb{V}(\varepsilon_i \mid T_i = 1),$$

which implies:

$$\sigma_0^2 = \sigma_1^2.$$

Therefore, the residuals are homoskedastic when the variances of the potential outcomes under treatment and control are equal:

$$\mathbb{V}(Y_i(0)) = \mathbb{V}(Y_i(1)).$$

Heteroskedasticity arises because the variability of the outcomes differs between the treatment and control groups.

Under Randomization, individuals are randomly assigned to treatment or control, but their potential outcomes $Y_i(0)$ and $Y_i(1)$ may have different variances.

In Practice, heteroskedasticity affects the efficiency of estimators and the validity of standard errors in regression analysis. Thus, statistical Inference must account for heteroskedasticity to obtain correct standard errors and confidence intervals.

Under Homoskedasticity, standard Ordinary Least Squares (OLS) estimators are BLUE (Best Linear Unbiased Estimators), and standard errors are valid.

2.c Obtaining the model

Taking this model to data, we obtain

$$(\hat{\alpha}, \hat{\tau}) = \arg\min_{a,b} \sum_{i=1}^{n} (y_i - a - bt_i)^2.$$

Derive closed form solutions for the vector $\hat{\theta} = (\hat{\alpha}, \hat{\tau})'$ using matrix notation and for the individual components using scalar notation. Give conditions for the solutions to exist and be unique.

In our exercise, y_i is the observed outcome for individual i, $t_i \in \{0, 1\}$ is the treatment indicator, and ε_i is the error term. Our goal is to find the Ordinary Least Squares (OLS) estimators $\hat{\alpha}$ and $\hat{\tau}$ by minimizing the sum of squared residuals:

$$(\hat{\alpha}, \hat{\tau}) = \arg\min_{a,b} \sum_{i=1}^{n} (Y_i - a - bT_i)^2.$$

We will derive the estimators using both matrix notation and scalar notation.

Matrix notation

In matrix notation, X and Y are presented by:

$$X = \begin{pmatrix} 1 & T_1 \\ 1 & T_2 \\ \vdots & \vdots \\ 1 & T_n \end{pmatrix} \quad (\text{an } n \times 2 \text{ matrix}).$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad (\text{an } n \times 1 \text{ vector}).$$

The parameter vector is represented by:

$$\theta = \begin{pmatrix} \alpha \\ \tau \end{pmatrix}.$$

The minimization problem can be written as the minimization of the following euclidean norm:

$$\arg\min_{a,b} \sum_{i=1}^{n} (Y_i - a - bT_i)^2 = \arg\min_{\theta} ||Y - X\theta||_2^2$$

In a traditional multivariate calculus approach, minimizing the $f(\theta)$ requires us to derive with respect to the vector θ and set $f'(\theta) = 0$.

$$\begin{aligned} ||Y - \theta X||_2^2 &= (Y - X\theta)'(Y - X\theta) \\ &= Y'Y - 2\theta' X'Y + \theta' X'X\theta \end{aligned}$$

$$\nabla_\theta ||Y - \theta X||_2^2 = -2X'Y + 2X'X\theta$$

Setting $\nabla_\theta ||Y - \theta X||_2^2 = 0$:

$$-2X'Y + 2'X'X\theta = 0$$

$$2X'X\theta = 2X'Y$$

$$X'X\theta = X'Y$$

$$\hat{\theta} = (X'X)^{-1}X'Y$$

The matrix $X'X$ and vector $X'Y$ are computed as:

$$X'X = \begin{pmatrix} n & n_1 \\ n_1 & n_1 \end{pmatrix}, \quad X'Y = \begin{pmatrix} S_Y \\ S_{TY} \end{pmatrix},$$

where:

- n = total number of observations
- $n_1 = \sum_{i=1}^{n} T_i$ (number of treated units)
- $S_Y = \sum_{i=1}^{n} Y_i$
- $S_{TY} = \sum_{i=1}^{n} T_i Y_i$
- $\overline{Y}$ be the sample mean of Y

- $\overline{Y}_1$ be the sample mean of Y for observations with $T = 1$
- $\overline{Y}_0$ be the sample mean of Y for observations with $T = 0$

The determinant of $X'X$ is:

$$\begin{aligned}\det(X'X) &= nn_1 - n_1^2 \\ &= (n_1 + n_0)n_1 - n_1^2 \\ &= n_1 n_0 + n_1^2 - n_1^2 \\ &= n_1 n_0\end{aligned}$$

where $n_0 = n - n_1$ is the number of control units. The inverse of $X'X$ is:

$$(X'X)^{-1} = \frac{1}{n_1 n_0} \begin{pmatrix} n_1 & -n_1 \\ -n_1 & n \end{pmatrix}.$$

Using this, the OLS estimator is given by:

$$\hat{\theta} = \frac{1}{n_1 n_0} \begin{pmatrix} n_1 & -n_1 \\ -n_1 & n \end{pmatrix} \begin{pmatrix} S_Y \\ S_{TY} \end{pmatrix}.$$

This leads to the following components:

$$\begin{aligned}\hat{\alpha} &= \frac{S_Y - S_{TY}}{n_0} \\ &= \frac{n_0 \overline{Y}_0}{n_0} \\ &= \overline{Y}_0\end{aligned}$$

$$
\begin{aligned}
\hat{\tau} &= \frac{nS_{TY} - n_1 S_Y}{n_1 n_0} \\
&= \frac{nS_{TY}}{n_1 n_0} - \frac{n_1 S_Y}{n_1 n_0} \\
&= \frac{n\overline{Y}_1}{n_0} - \frac{S_Y}{n_0} \\
&= \frac{n\overline{Y}_1}{n_0} - \frac{(n_1 + n_0)S_Y}{(n_1 + n_0)n_0} \\
&= \frac{n\overline{Y}_1}{n_0} - \frac{(n_1 + n_0)\overline{Y}}{n_0} \\
&= \frac{n\overline{Y}_1}{n_0} - \frac{n\overline{Y}}{n_0} \\
&= \frac{n\overline{Y}_1}{n_0} - \frac{n_0\overline{Y}_0 + n_1\overline{Y}_1}{n_0} \\
&= \frac{n\overline{Y}_1}{n_0} - \overline{Y}_0 - \frac{n_1\overline{Y}_1}{n_0} \\
&= \frac{(n - n_1)\overline{Y}_1}{n_0} - \overline{Y}_0 \\
&= \frac{n_0\overline{Y}_1}{n_0} - \overline{Y}_0 \\
&= \overline{Y}_1 - \overline{Y}_0
\end{aligned}
$$

Scalar Notation

$$\text{SSR} = \sum_{i=1}^{n} (Y_i - \alpha - \tau T_i)^2$$

To minimize SSR, again we take the partial derivatives of SSR with respect to α and τ, set them to zero, and solve for α and τ.

With respect to α:

$$\frac{\partial \text{SSR}}{\partial \alpha} = -2 \sum_{i=1}^{n} (Y_i - \alpha - \tau T_i) = 0$$

Simplifying:

$$\sum_{i=1}^{n} Y_i = \alpha n + \tau \sum_{i=1}^{n} T_i$$

With respect to τ:

$$\frac{\partial \text{SSR}}{\partial \tau} = -2\sum_{i=1}^{n} T_i(Y_i - \alpha - \tau T_i) = 0$$

Simplifying:

$$\sum_{i=1}^{n} T_i Y_i = \alpha \sum_{i=1}^{n} T_i + \tau \sum_{i=1}^{n} T_i^2$$

Since T_i is binary (0 or 1), $T_i^2 = T_i$.
From the derivative with respect to α:

$$n\overline{Y} = \alpha n + \tau n_1 \quad \Rightarrow \quad \alpha = \overline{Y} - \tau\left(\frac{n_1}{n}\right)$$

From the derivative with respect to τ:

$$n_1\overline{Y}_1 = \alpha n_1 + \tau n_1 \quad \Rightarrow \quad \alpha = \overline{Y}_1 - \tau$$

Set the two expressions for α equal to each other:

$$\overline{Y} - \tau\left(\frac{n_1}{n}\right) = \overline{Y}_1 - \tau$$

Rearranging to solve for τ:

$$\overline{Y} - \overline{Y}_1 = \tau\left(\frac{n_1}{n} - 1\right) = -\tau\left(1 - \frac{n_1}{n}\right) = -\tau\left(\frac{n_0}{n}\right)$$

$$\tau = \frac{\overline{Y}_1 - \overline{Y}}{\frac{n_0}{n}} = \frac{\overline{Y}_1 - \overline{Y}}{\overline{T}(1-\overline{T})}$$

Since $\overline{Y} = \frac{n_1}{n}\overline{Y}_1 + \frac{n_0}{n}\overline{Y}_0$, substituting back gives:

$$\hat{\tau} = \overline{Y}_1 - \overline{Y}_0$$

Using the expression for τ in the equation for α:

$$\alpha = \overline{Y} - \tau\overline{T} = \overline{Y} - (\overline{Y}_1 - \overline{Y}_0)\overline{T}$$

More directly, from the second normal equation:

$$\hat{\alpha} = \overline{Y}_0$$

In this binary setting:

- The intercept α captures the expected value of Y when $T = 0$.

- The coefficient τ captures the treatment effect, i.e., how much the expected value of Y changes when T changes from 0 to 1.

In other words, when both Y and T are binary, the OLS estimates for α and τ correspond to the average outcomes in the control group and the difference in averages between the treatment and control groups, respectively.

The solutions $\hat{\alpha}$ and $\hat{\tau}$ exist and are unique if and only if both $n_1 > 0$ and $n_0 > 0$ meaning that both the treatment and control groups have observations. This directly implies that $n \geq p = 2$, where p is the number of parameters in the model.

2.d Proving estimator properties

Prove that $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$, where $\bar{Y}_1 = \frac{1}{n_1} \sum_{i=1}^{n} y_i t_i$ with $n_1 = \sum_{i=1}^{n} t_i$, and similarly for $\bar{Y}_0$.

We are given the linear model:

$$Y_i = \alpha + \tau T_i + \varepsilon_i,$$

From the derivation completed in (2.c), we arrive at:

$$\hat{\tau} = \frac{nS_{TY} - n_1 S_Y}{n_1 n_0}$$

Where:

- n: number of total observations.
- n_1: number of observatins for which $T = 1$.
- n_1: number of observatins for which $T = 0$.
- S_Y: $\sum_{i=1}^{n} Y_i$
- S_{YT}: $\sum_{i=1}^{n} Y_i T_i$
- $\overline{Y}$: $\frac{S_Y}{n}$
- $\overline{Y}_1$: $\frac{S_{YT}}{n_1}$
- $\overline{Y}_0$: $\frac{S_Y - S_{YT}}{n_0}$

With the defined variables, we reach the following result:

$$
\begin{aligned}
\hat{\tau} &= \frac{nS_{TY} - n_1 S_Y}{n_1 n_0} \\
&= \frac{nS_{TY}}{n_1 n_0} - \frac{n_1 S_Y}{n_1 n_0} \\
&= \frac{n\overline{Y}_1}{n_0} - \frac{S_Y}{n_0} \\
&= \frac{n\overline{Y}_1}{n_0} - \frac{(n_1 + n_0)S_Y}{(n_1 + n_0)n_0} \\
&= \frac{n\overline{Y}_1}{n_0} - \frac{(n_1 + n_0)\overline{Y}}{n_0} \\
&= \frac{n\overline{Y}_1}{n_0} - \frac{n\overline{Y}}{n_0} \\
&= \frac{n\overline{Y}_1}{n_0} - \frac{n_0\overline{Y}_0 + n_1\overline{Y}_1}{n_0} \\
&= \frac{n\overline{Y}_1}{n_0} - \overline{Y}_0 - \frac{n_1\overline{Y}_1}{n_0} \\
&= \frac{(n - n_1)\overline{Y}_1}{n_0} - \overline{Y}_0 \\
&= \frac{n_0\overline{Y}_1}{n_0} - \overline{Y}_0 \\
&= \overline{Y}_1 - \overline{Y}_0
\end{aligned}
$$

Therefore, the OLS estimator $\hat{\tau}$ is equal to the difference in sample means between the treated and control groups.

2.e Variance of estimator

Use least squares algebra to derive the variance of $\hat{\theta}$ conditional on $t_1, \ldots, t_n$ and find its probability limit. Give a consistent estimator. Is your estimator unbiased?

We consider the linear regression model:

$$Y_i = \alpha + \tau T_i + \varepsilon_i,$$

As derived in the previous question, the OLS estimator is:

$$\hat{\theta} = (\hat{\alpha}, \hat{\tau})' = (X'X)^{-1}X'Y,$$

The variance of $\hat{\theta}$ conditional on T is:

$$\mathbb{V}(\hat{\theta} \mid T) = (X'X)^{-1}X'\mathbb{V}(\varepsilon \mid T)X(X'X)^{-1}.$$

Let $n_1 = \sum_{i=1}^{n} T_i$ (number of treated units) and $n_0 = n - n_1$ (number of control units). Then,

$$X'X = \begin{pmatrix} n & n_1 \\ n_1 & n_1 \end{pmatrix}, \quad \text{and} \quad (X'X)^{-1} = \frac{1}{n_1 n_0}\begin{pmatrix} n_1 & -n_1 \\ -n_1 & n \end{pmatrix}.$$

Getting results from above, the error terms ε_i have variances conditional on T_i as follows:

$$\mathbb{V}(\varepsilon_i \mid T_i) = \begin{cases} \sigma_0^2, & \text{if } T_i = 0, \\ \sigma_1^2, & \text{if } T_i = 1. \end{cases}$$

Define the diagonal matrix $D = \text{diag}(v_1, \ldots, v_n)$ with $v_i = \mathbb{V}(\varepsilon_i \mid T_i)$. More explicitly:

$$D = \begin{bmatrix} \mathbb{V}(\varepsilon_1|T_1) & 0 & 0 & 0 & \ldots & 0 \\ 0 & \mathbb{V}(\varepsilon_2|T_2) & 0 & 0 & \ldots & 0 \\ 0 & 0 & \mathbb{V}(\varepsilon_3|T_3) & 0 & \ldots & 0 \\ 0 & 0 & 0 & \mathbb{V}(\varepsilon_4|T_4) & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & \mathbb{V}(\varepsilon_n|T_n) \end{bmatrix}$$

where $D_{i,j} = 0$ for all $i \neq j$ given that the errors are i.i.d.

Thus,

$$X'DX = \begin{pmatrix} n_0\sigma_0^2 + n_1\sigma_1^2 & n_1\sigma_1^2 \\ n_1\sigma_1^2 & n_1\sigma_1^2 \end{pmatrix}.$$

The results become simpler because $t_i, t_i^2 \in \{0, 1\}\ \forall\ i$

We compute:

$$\mathbb{V}(\hat{\theta} \mid T) = (X'X)^{-1}X'DX(X'X)^{-1}.$$

This results in:

$$\begin{aligned} \mathbb{V}(\hat{\theta} \mid T) &= \left[\frac{1}{n_1 n_0}\begin{pmatrix} n_1 & -n_1 \\ -n_1 & n \end{pmatrix}\right]\begin{pmatrix} n_0\sigma_0^2 + n_1\sigma_1^2 & n_1\sigma_1^2 \\ n_1\sigma_1^2 & n_1\sigma_1^2 \end{pmatrix}\left[\frac{1}{n_1 n_0}\begin{pmatrix} n_1 & -n_1 \\ -n_1 & n \end{pmatrix}\right] \\ &= \begin{pmatrix} \dfrac{\sigma_0^2}{n_0} & -\dfrac{\sigma_0^2}{n_0} \\ -\dfrac{\sigma_0^2}{n_0} & \dfrac{\sigma_0^2}{n_0} + \dfrac{\sigma_1^2}{n_1} \end{pmatrix} \end{aligned}$$

Probability Limit of the Variance

As $n \to \infty$, assuming the proportion of treated units $p = \lim_{n\to\infty} \dfrac{n_1}{n}$ exists and is bounded away from 0 and 1, we have:

- $n_1 \approx np$,
- $n_0 \approx n(1-p)$.

Thus, the probability limit of the variance-covariance matrix is:

$$\text{plim}_{n\to\infty} \mathbb{V}(\hat{\theta}) = \frac{1}{n} \begin{pmatrix} \frac{\sigma_0^2}{1-p} & -\frac{\sigma_0^2}{1-p} \\ -\frac{\sigma_0^2}{1-p} & \frac{\sigma_0^2}{1-p} + \frac{\sigma_1^2}{p} \end{pmatrix}.$$

Consistent Estimator of the Variance

We can estimate σ_0^2 and σ_1^2 using the sample variances from the control and treatment groups, respectively:

$$\hat{\sigma}_0^2 = s_0^2 = \frac{1}{n_0 - 1} \sum_{i:T_i=0} (Y_i - \bar{Y}_0)^2, \quad \hat{\sigma}_1^2 = s_1^2 = \frac{1}{n_1 - 1} \sum_{i:T_i=1} (Y_i - \bar{Y}_1)^2.$$

The consistent variance estimator is:

$$\hat{\mathbb{V}}(\hat{\theta} \mid T) = \begin{pmatrix} \frac{s_0^2}{n_0} & -\frac{s_0^2}{n_0} \\ -\frac{s_0^2}{n_0} & \frac{s_0^2}{n_0} + \frac{s_1^2}{n_1} \end{pmatrix}.$$

Unbiasedness of the Estimator

The sample variances s_0^2 and s_1^2 are unbiased estimators of σ_0^2 and σ_1^2, respectively. Therefore, the components of $\hat{\mathbb{V}}(\hat{\theta} \mid T)$ are unbiased estimators of the true variances and covariances. Hence, $\hat{\mathbb{V}}(\hat{\theta} \mid T)$ is both consistent and unbiased.

Now we will estimate the average treatment effect (ATE) using the plug-in principle. The ATE is $\tau = \mathbb{E}[Y(1) - Y(0)]$. The plug-in principle replaces unknown quantities with sample analogues (put hats on stuff) and replaces population averages with sample averages. Therefore we will consider

$$\hat{\tau}_{\text{pi}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_i,$$

for some estimate $\hat{\tau}_i = Y_i(1) \hat{-} Y_i(0)$.

2.f Least squares on demeaned covariates

First, we run least squares on demeaned covariates separately in each group to obtain $\hat{\theta}_t = (\hat{\alpha}_t, \hat{\beta}_t)'$ as

$$(\hat{\alpha}_t, \hat{\beta}_t) = \arg\min_{a,b} \sum_{i=1}^{n} 1\{t_i = t\} \left(y_i - a - b(x_i - \bar{x})\right)^2, \quad \text{where} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Give closed form solutions for the vector $\hat{\theta}_t = (\hat{\alpha}_t, \hat{\beta}_t)'$ using matrix notation and give conditions for the solutions to exist and be unique.

To arrive to a closed-form solutions for $\hat{\theta}_t = (\hat{\alpha}_t, \hat{\beta}_t)'$, we define for each treatment group t:

- Number of Observations:
$$n_t = \sum_{i=1}^{n} 1\{t_i = t\}.$$
- Response Vector:
$$Y_t = \begin{bmatrix} y_{i_1} \\ y_{i_2} \\ \vdots \\ y_{i_{n_t}} \end{bmatrix},$$
where i_j indexes observations with $t_{i_j} = t$.
- Design Matrix:
$$X_t = \begin{bmatrix} 1 & x_{i_1} - \bar{x} \\ 1 & x_{i_2} - \bar{x} \\ \vdots & \vdots \\ 1 & x_{i_{n_t}} - \bar{x} \end{bmatrix},$$
which is an $n_t \times 2$ matrix.

As derived in (2.c), the least squares estimator for $\hat{\theta}_t$ is:

$$\hat{\theta}_t = (X_t' X_t)^{-1} X_t' Y_t,$$

provided that $X_t' X_t$ is invertible.

Now, we compute $X_t' X_t$ and $X_t' Y_t$

Compute the components of $X_t' X_t$:

- Elements of $X_t' X_t$:
$$X_t' X_t = \begin{bmatrix} n_t & S_{x,t} \\ S_{x,t} & S_{xx,t} \end{bmatrix},$$
where $S_{x,t} = \sum_{i:t_i=t}(x_i - \bar{x})$ and $S_{xx,t} = \sum_{i:t_i=t}(x_i - \bar{x})^2$.
- Elements of $X_t' Y_t$:
$$X_t' Y_t = \begin{bmatrix} \sum_{i:t_i=t} y_i \\ S_{xy,t} \end{bmatrix},$$
where $S_{xy,t} = \sum_{i:t_i=t}(x_i - \bar{x}) y_i$.

From it, we derive the normal equations:

$$\begin{cases} n_t\hat{\alpha}_t + S_{x,t}\hat{\beta}_t = \sum_{i:t_i=t} y_i, \\ S_{x,t}\hat{\alpha}_t + S_{xx,t}\hat{\beta}_t = S_{xy,t}. \end{cases}$$

And solve for $\hat{\beta}_t$by first, expressing $\hat{\alpha}_t$ from the first equation:

$$\hat{\alpha}_t = \frac{1}{n_t}\left(\sum_{i:t_i=t} y_i - S_{x,t}\hat{\beta}_t\right).$$

Substitute $\hat{\alpha}_t$ into the second equation:

$$S_{x,t}\left(\frac{1}{n_t}\left(\sum_{i:t_i=t} y_i - S_{x,t}\hat{\beta}_t\right)\right) + S_{xx,t}\hat{\beta}_t = S_{xy,t}.$$

Simplify:

$$\frac{S_{x,t}}{n_t}\sum_{i:t_i=t} y_i - \frac{(S_{x,t})^2}{n_t}\hat{\beta}_t + S_{xx,t}\hat{\beta}_t = S_{xy,t}.$$

Multiply both sides by n_t:

$$S_{x,t}\sum_{i:t_i=t} y_i - (S_{x,t})^2\hat{\beta}_t + n_t S_{xx,t}\hat{\beta}_t = n_t S_{xy,t}.$$

Rewriting:

$$\left(-(S_{x,t})^2 + n_t S_{xx,t}\right)\hat{\beta}_t = n_t S_{xy,t} - S_{x,t}\sum_{i:t_i=t} y_i.$$

Define the determinant D_t:

$$D_t = n_t S_{xx,t} - (S_{x,t})^2.$$

Then, the solution for $\hat{\beta}_t$ is:

$$\hat{\beta}_t = \frac{n_t S_{xy,t} - S_{x,t}\sum_{i:t_i=t} y_i}{D_t}.$$

We now aim to simplify $\hat{\beta}_t$.
Express $S_{x,t}$ and $\sum_{i:t_i=t} y_i$ in terms of sample means:

- Group Means:

$$\bar{x}_t = \frac{1}{n_t}\sum_{i:t_i=t} x_i, \quad \bar{y}_t = \frac{1}{n_t}\sum_{i:t_i=t} y_i.$$

- Overall Mean:
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$
- Centered Sums:
$$S_{x,t} = n_t(\bar{x}_t - \bar{x}), \quad \sum_{i:t_i=t} y_i = n_t\bar{y}_t.$$

Substitute these into $\hat{\beta}_t$:

$$\hat{\beta}_t = \frac{n_t S_{xy,t} - n_t(\bar{x}_t - \bar{x})n_t\bar{y}_t}{n_t S_{xx,t} - n_t^2(\bar{x}_t - \bar{x})^2}.$$

Simplify numerator and denominator:

- Numerator:
$$n_t S_{xy,t} - n_t^2(\bar{x}_t - \bar{x})\bar{y}_t = n_t\left(\sum_{i:t_i=t}(x_i - \bar{x})y_i - n_t(\bar{x}_t - \bar{x})\bar{y}_t\right).$$
- Denominator:
$$D_t = n_t S_{xx,t} - n_t^2(\bar{x}_t - \bar{x})^2 = n_t\left(\sum_{i:t_i=t}(x_i - \bar{x})^2 - n_t(\bar{x}_t - \bar{x})^2\right).$$

The denominator simplifies to $n_t\sum_{i:t_i=t}(x_i - \bar{x}_t)^2$, and the numerator becomes:

$$n_t\left(\sum_{i:t_i=t}(x_i - \bar{x})(y_i - \bar{y}_t)\right) = n_t\sum_{i:t_i=t}(x_i - \bar{x}_t)(y_i - \bar{y}_t).$$

Therefore, $\hat{\beta}_t$ becomes:

$$\hat{\beta}_t = \frac{\sum_{i:t_i=t}(x_i - \bar{x}_t)(y_i - \bar{y}_t)}{\sum_{i:t_i=t}(x_i - \bar{x}_t)^2}.$$

Solve for $\hat{\alpha}_t$

Using the expression for $\hat{\alpha}_t$:

$$\hat{\alpha}_t = \bar{y}_t - (\bar{x}_t - \bar{x})\,\hat{\beta}_t.$$

Final Closed-Form Solutions:

- Estimate of $\hat{\beta}_t$:
$$\hat{\beta}_t = \frac{\sum_{i:t_i=t}(x_i - \bar{x}_t)(y_i - \bar{y}_t)}{\sum_{i:t_i=t}(x_i - \bar{x}_t)^2} = \frac{\mathrm{Cov}_t(x_i, y_i)}{\mathrm{Var}_t(x_i)},$$

where:

$$\text{Cov}_t(x_i, y_i) = \frac{1}{n_t} \sum_{i:t_i=t} (x_i - \bar{x}_t)(y_i - \bar{y}_t),$$

$$\text{Var}_t(x_i) = \frac{1}{n_t} \sum_{i:t_i=t} (x_i - \bar{x}_t)^2.$$

- Estimate of $\hat{\alpha}_t$:

$$\hat{\alpha}_t = \bar{y}_t - (\bar{x}_t - \bar{x})\hat{\beta}_t.$$

Conditions for Existence and Uniqueness of Solutions:

- Positive Variance of Covariates within Each Group:

  The denominator in $\hat{\beta}_t$ must be positive:

$$\text{Var}_t(x_i) = \frac{1}{n_t} \sum_{i:t_i=t} (x_i - \bar{x}_t)^2 > 0.$$

  This condition ensures that the covariate x_i varies within group t. If all x_i are identical within a group, $\text{Var}_t(x_i) = 0$, and $\hat{\beta}_t$ is undefined.

- Sufficient Number of Observations:

  Each group must have at least two observations ($n_t \geq 2$) to compute variances and covariances.

- Full Rank of the Design Matrix:

  The matrix $X_t' X_t$ must be invertible. This requires that the columns of X_t are linearly independent, which is guaranteed when $\text{Var}_t(x_i) > 0$.

In conclusion, we have derived closed-form solutions for the parameters $\hat{\theta}_t = (\hat{\alpha}_t, \hat{\beta}_t)'$ when performing least squares regression on demeaned covariates within each treatment group t. The solutions are:

- Slope Estimate:

$$\hat{\beta}_t = \frac{\text{Cov}_t(x_i, y_i)}{\text{Var}_t(x_i)}.$$

- Intercept Estimate:

$$\hat{\alpha}_t = \bar{y}_t - (\bar{x}_t - \bar{x})\hat{\beta}_t.$$

This estimation approach allows us to compute individual treatment effect estimates $\hat{\tau}_i = \hat{Y}_i(1) - \hat{Y}_i(0)$ and, subsequently, the average treatment effect (ATE) using the plug-in principle:

$$\hat{\tau}_{\text{pi}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_i.$$

2.g Predicted values for counterfactuals

Given $\hat{\theta}_0$ and $\hat{\theta}_1$, form predicted values for each potential outcome, i.e., give expressions for $\hat{Y}_i(1)$ and $\hat{Y}_i(0)$. Combine these counterfactual predictions with each observation's factual (observed) outcome to obtain estimates of the individual causal effect $\tau_i = Y_i(1) - Y_i(0)$. Denote these predicted values as $\hat{\tau}_i$.

Given the estimated parameters $\hat{\theta}_0 = (\hat{\alpha}_0, \hat{\beta}_0)'$ and $\hat{\theta}_1 = (\hat{\alpha}_1, \hat{\beta}_1)'$ from the least squares regressions on demeaned covariates within each treatment group derived in (2.f), we aim to form predicted values for each potential outcome $Y_i(1)$ and $Y_i(0)$ for every observation i.

For each observation i, we consider the following:

- Observed Potential Outcome: The potential outcome corresponding to the treatment actually received is observed.
  - If $T_i = 1$, then $Y_i(1) = Y_i$ is observed.
  - If $T_i = 0$, then $Y_i(0) = Y_i$ is observed.
- Counterfactual Potential Outcome: The potential outcome corresponding to the treatment not received is unobserved and must be estimated using the regression model from the opposite treatment group (possible due to the stated assumptions).

Therefore, the predicted potential outcomes $\hat{Y}_i(1)$ and $\hat{Y}_i(0)$ are defined as:

- For Observations with $T_i = 1$:
  - Observed Outcome:
    $$\hat{Y}_i(1) = Y_i.$$
    Since the individual received the treatment ($T_i = 1$), their outcome under treatment is observed.
  - Predicted Counterfactual Outcome:
    $$\hat{Y}_i(0) = \hat{\alpha}_0 + \hat{\beta}_0(x_i - \bar{x}).$$
    This is the predicted outcome under control, using the estimated parameters from the control group regression.
- For Observations with $T_i = 0$:
  - Observed Outcome:
    $$\hat{Y}_i(0) = Y_i.$$
    Since the individual is in the control group ($T_i = 0$), their outcome under control is observed.

– Predicted Counterfactual Outcome:

$$\hat{Y}_i(1) = \hat{\alpha}_1 + \hat{\beta}_1(x_i - \bar{x}).$$

This is the predicted outcome under treatment, using the estimated parameters from the treatment group regression.

With the predicted potential outcomes, we estimate the individual causal effect τ_i for each observation i as:

$$\hat{\tau}_i = \hat{Y}_i(1) - \hat{Y}_i(0).$$

This calculation differs depending on the treatment assignment T_i:

- For Observations with $T_i = 1$:

$$\hat{\tau}_i = Y_i - \left(\hat{\alpha}_0 + \hat{\beta}_0(x_i - \bar{x})\right).$$

Here, Y_i is the observed outcome under treatment, and $\hat{\alpha}_0 + \hat{\beta}_0(x_i - \bar{x})$ is the predicted outcome under control.

- For Observations with $T_i = 0$:

$$\hat{\tau}_i = \left(\hat{\alpha}_1 + \hat{\beta}_1(x_i - \bar{x})\right) - Y_i.$$

Here, Y_i is the observed outcome under control, and $\hat{\alpha}_1 + \hat{\beta}_1(x_i - \bar{x})$ is the predicted outcome under treatment.

In conclusion, for all $i = 1, \ldots, n$:

$$\hat{Y}_i(1) = \begin{cases} Y_i, & \text{if } T_i = 1, \\ \hat{\alpha}_1 + \hat{\beta}_1(x_i - \bar{x}), & \text{if } T_i = 0. \end{cases}$$

$$\hat{Y}_i(0) = \begin{cases} \hat{\alpha}_0 + \hat{\beta}_0(x_i - \bar{x}), & \text{if } T_i = 1, \\ Y_i, & \text{if } T_i = 0. \end{cases}$$

And the estimated Individual Causal Effects:

$$\hat{\tau}_i = \begin{cases} Y_i - \left(\hat{\alpha}_0 + \hat{\beta}_0(x_i - \bar{x})\right), & \text{if } T_i = 1, \\ \left(\hat{\alpha}_1 + \hat{\beta}_1(x_i - \bar{x})\right) - Y_i, & \text{if } T_i = 0. \end{cases}$$

Again, the stated assumptions are key for the result to work:

- SUTVA (Stable Unit Treatment Value Assumption): Assumes that the potential outcomes for any unit are unaffected by the treatment assignments of other units.

- Consistency: The observed outcome corresponds to the potential outcome under the treatment actually received.
- Randomization: The treatment assignment T_i is independent of the potential outcomes given X_i, ensuring unbiased estimation.

If they hold, the estimated ATE using the plug-in principle is:

$$\hat{\tau}_{\text{pi}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_i.$$

By averaging the estimated individual causal effects $\hat{\tau}_i$, we obtain an estimate of the overall effect of the treatment across the population.

2.h Sample average treatment effect

Show that $\hat{\tau}_{\text{pi}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_i = \hat{\alpha}_1 - \hat{\alpha}_0$, the latter being the intercepts from (2).

We aim to show that the average of the estimated individual causal effects $\hat{\tau}_{\text{pi}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_i$ is equal to the difference of the estimated intercepts from the group-specific regressions, i.e., $\hat{\tau}_{\text{pi}} = \hat{\alpha}_1 - \hat{\alpha}_0$.

We begin by defining the estimated individual causal effects $\hat{\tau}_i$:

- For observations with $T_i = 1$:

$$\hat{\tau}_i = Y_i - \left(\hat{\alpha}_0 + \hat{\beta}_0 (x_i - \bar{x}) \right).$$

- For observations with $T_i = 0$:

$$\hat{\tau}_i = \left(\hat{\alpha}_1 + \hat{\beta}_1 (x_i - \bar{x}) \right) - Y_i.$$

Now, express $\hat{\tau}_{\text{pi}}$ as:

$$\hat{\tau}_{\text{pi}} = \frac{1}{n} \left(\sum_{i \in I_1} \hat{\tau}_i + \sum_{i \in I_0} \hat{\tau}_i \right),$$

where I_1 and I_0 denote the sets of indices where $T_i = 1$ and $T_i = 0$, respectively, and $n_1 = |I_1|$, $n_0 = |I_0|$ ($n = n_1 + n_0$).

Substituting the expressions for $\hat{\tau}_i$ into $\hat{\tau}_{\text{pi}}$:

$$\hat{\tau}_{\text{pi}} = \frac{1}{n} \left(\sum_{i \in I_1} \left[Y_i - \left(\hat{\alpha}_0 + \hat{\beta}_0 (x_i - \bar{x}) \right) \right] + \sum_{i \in I_0} \left[\left(\hat{\alpha}_1 + \hat{\beta}_1 (x_i - \bar{x}) \right) - Y_i \right] \right).$$

Rewriting the sums:

$$\hat{\tau}_{\text{pi}} = \frac{1}{n} \left(\sum_{i \in I_1} Y_i - n_1 \hat{\alpha}_0 - \hat{\beta}_0 \sum_{i \in I_1} (x_i - \bar{x}) + n_0 \hat{\alpha}_1 + \hat{\beta}_1 \sum_{i \in I_0} (x_i - \bar{x}) - \sum_{i \in I_0} Y_i \right).$$

Since $\bar{x}$ is the overall mean, the sum of demeaned covariates over all observations is zero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \implies \sum_{i\in I_1}(x_i - \bar{x}) + \sum_{i\in I_0}(x_i - \bar{x}) = 0.$$

Thus,

$$\sum_{i\in I_1}(x_i - \bar{x}) = -\sum_{i\in I_0}(x_i - \bar{x}).$$

Define:

$$S_{X,1} = \sum_{i\in I_1}(x_i - \bar{x}), \quad S_{X,0} = \sum_{i\in I_0}(x_i - \bar{x}) = -S_{X,1}.$$

Similarly, define the sums of Y_i:

$$S_{Y,1} = \sum_{i\in I_1} Y_i = n_1\bar{y}_1, \quad S_{Y,0} = \sum_{i\in I_0} Y_i = n_0\bar{y}_0.$$

Substitute back into $\hat{\tau}_{\text{pi}}$:

$$\hat{\tau}_{\text{pi}} = \frac{1}{n}\left(S_{Y,1} - S_{Y,0} - n_1\hat{\alpha}_0 + n_0\hat{\alpha}_1 - \hat{\beta}_0 S_{X,1} + \hat{\beta}_1(-S_{X,1})\right).$$

Simplify the terms involving $\hat{\beta}_t$:

$$-\hat{\beta}_0 S_{X,1} + \hat{\beta}_1(-S_{X,1}) = -(\hat{\beta}_0 + \hat{\beta}_1)S_{X,1}.$$

So,

$$\hat{\tau}_{\text{pi}} = \frac{1}{n}\left(S_{Y,1} - S_{Y,0} - n_1\hat{\alpha}_0 + n_0\hat{\alpha}_1 - (\hat{\beta}_0 + \hat{\beta}_1)S_{X,1}\right).$$

To simplify this further, we use the fact that the group means of the covariates, $\bar{x}_1$ and $\bar{x}_0$, are related to the overall mean $\bar{x}$ by:

$$\bar{x}_1 - \bar{x} = \frac{n_0}{n}(\bar{x}_1 - \bar{x}_0), \quad \bar{x}_0 - \bar{x} = -\frac{n_1}{n}(\bar{x}_1 - \bar{x}_0).$$

This gives:

$$S_{X,1} = n_1\left(\frac{n_0}{n}(\bar{x}_1 - \bar{x}_0)\right) = \frac{n_1 n_0}{n}(\bar{x}_1 - \bar{x}_0).$$

Substituting this back into $\hat{\tau}_{\text{pi}}$:

$$\hat{\tau}_{\text{pi}} = \frac{1}{n}\left(n_1\bar{y}_1 - n_0\bar{y}_0 - n_1\hat{\alpha}_0 + n_0\hat{\alpha}_1 - (\hat{\beta}_0 + \hat{\beta}_1)\left(\frac{n_1 n_0}{n}(\bar{x}_1 - \bar{x}_0)\right)\right).$$

Next, observe that the intercepts $\hat{\alpha}_t$ in each group are related to the group means:

$$\hat{\alpha}_t = \bar{y}_t - \hat{\beta}_t(\bar{x}_t - \bar{x}).$$

Thus, we can rewrite:

$$n_1(\bar{y}_1 - \hat{\alpha}_1) - n_0(\bar{y}_0 - \hat{\alpha}_0) = n_1\hat{\beta}_1(\bar{x}_1 - \bar{x}) - n_0\hat{\beta}_0(\bar{x}_0 - \bar{x}).$$

Substituting the expressions for $\bar{x}_t - \bar{x}$ into this:

$$n_1(\bar{y}_1 - \hat{\alpha}_1) - n_0(\bar{y}_0 - \hat{\alpha}_0) = \frac{n_1 n_0}{n}(\hat{\beta}_1 + \hat{\beta}_0)(\bar{x}_1 - \bar{x}_0).$$

Finally, we substitute this back into the expression for $\hat{\tau}_{\text{pi}}$, giving:

$$\hat{\tau}_{\text{pi}} = \frac{1}{n}\left(n(\hat{\alpha}_1 - \hat{\alpha}_0) + \frac{n_1 n_0}{n}(\hat{\beta}_1 + \hat{\beta}_0)(\bar{x}_1 - \bar{x}_0) - \frac{n_1 n_0}{n}(\hat{\beta}_1 + \hat{\beta}_0)(\bar{x}_1 - \bar{x}_0)\right).$$

The terms involving $\hat{\beta}_0$ and $\hat{\beta}_1$ cancel out, and we are left with:

$$\hat{\tau}_{\text{pi}} = \hat{\alpha}_1 - \hat{\alpha}_0.$$

Thus, we have shown that:

$$\hat{\tau}_{\text{pi}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_i = \hat{\alpha}_1 - \hat{\alpha}_0.$$

2.i Least squares for ATE

Show how to obtain $\hat{\tau}_{\text{pi}}$ using a single least squares regression and, if needed, by taking linear combinations of its coefficients.

The combined regression model that accounts for differences in both intercepts and slopes between the treatment and control groups is:

$$Y_i = \gamma_0 + \gamma_1 T_i + \gamma_2(x_i - \bar{x}) + \gamma_3 T_i(x_i - \bar{x}) + \varepsilon_i,$$

where:

- γ_0 is the intercept for the control group,
- γ_1 represents the difference in intercepts between the treatment and control groups,
- γ_2 is the slope for the control group,
- γ_3 represents the difference in slopes between the treatment and control groups.

The model allows us to interpret the coefficients as follows:

- For the control group ($T_i = 0$):

$$Y_i = \gamma_0 + \gamma_2(x_i - \bar{x}) + \varepsilon_i.$$

- For the treatment group ($T_i = 1$):

$$Y_i = (\gamma_0 + \gamma_1) + (\gamma_2 + \gamma_3)(x_i - \bar{x}) + \varepsilon_i.$$

Here, γ_1 represents the difference in intercepts between the treatment and control groups.

Using the combined data from both groups, we estimate the regression coefficients $\hat{\gamma}_0$, $\hat{\gamma}_1$, $\hat{\gamma}_2$, and $\hat{\gamma}_3$ by minimizing the sum of squared residuals:

$$(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3) = \arg \min_{\gamma_0, \gamma_1, \gamma_2, \gamma_3} \sum_{i=1}^{n} \left(Y_i - \gamma_0 - \gamma_1 T_i - \gamma_2 (x_i - \bar{x}) - \gamma_3 T_i (x_i - \bar{x})\right)^2 .$$

Now, recall that from the separate regressions for each group, we have:

- $\hat{\alpha}_0$ = intercept estimate from the control group regression,
- $\hat{\alpha}_1$ = intercept estimate from the treatment group regression,
- $\hat{\tau}_{\text{pi}} = \hat{\alpha}_1 - \hat{\alpha}_0$.

From the combined regression, we can identify that:

- $\hat{\alpha}_0 = \hat{\gamma}_0$,
- $\hat{\alpha}_1 = \hat{\gamma}_0 + \hat{\gamma}_1$.

Thus, the difference in intercepts is:

$$\hat{\alpha}_1 - \hat{\alpha}_0 = (\hat{\gamma}_0 + \hat{\gamma}_1) - \hat{\gamma}_0 = \hat{\gamma}_1.$$

Therefore, we conclude that:

$$\hat{\tau}_{\text{pi}} = \hat{\alpha}_1 - \hat{\alpha}_0 = \hat{\gamma}_1.$$

To summarize, the plug-in estimator for the average treatment effect is:

$$\hat{\tau}_{\text{pi}} = \hat{\gamma}_1.$$

This combined regression method efficiently estimates the ATE, avoiding the need for separate regressions for each group. The regression captures the differential effects of the treatment through both intercept and slope terms, and the coefficient $\hat{\gamma}_1$ directly provides the estimated ATE.

2.j Convergence in probability

Prove that $\hat{\tau}_{\text{pi}}$ converges in probability to the average treatment effect.

We aim to show that the plug-in estimator for the average treatment effect (ATE),

$$\hat{\tau}_{\text{pi}} = \hat{\alpha}_1 - \hat{\alpha}_0,$$

converges in probability to the true ATE,

$$\tau = \mathbb{E}[Y(1) - Y(0)].$$

To establish this result, it is required that we:

- Define the Population Regression Parameters.
- Show that the Sample Estimates Converge to the Population Parameters.
- Demonstrate that $\hat{\tau}_{\text{pi}} \xrightarrow{p} \tau$.

First, we define the population regression parameters.
In the context of our randomized experiment, we have:

- Potential Outcomes: $Y_i(0)$ and $Y_i(1)$ are the potential outcomes under control and treatment, respectively.
- Observed Outcome: $Y_i = Y_i(T_i) = Y_i(0)(1 - T_i) + Y_i(1)T_i$.

We run separate least squares regressions within each treatment group $t \in \{0, 1\}$:

$$Y_i = \alpha_t + \beta_t(x_i - \bar{x}) + \varepsilon_i, \quad \text{for } T_i = t,$$

where $\bar{x}$ is the overall sample mean of x_i.
Let us define the population counterparts of the regression parameters:

- Population Intercept:
$$\alpha_t = \mathbb{E}[Y_i|T_i = t] - \beta_t \left(\mathbb{E}[x_i|T_i = t] - \bar{x}\right),$$
where β_t is the population slope.
- Population Slope:
$$\beta_t = \frac{\text{Cov}[x_i, Y_i|T_i = t]}{\text{Var}[x_i|T_i = t]}.$$

By randomization and the independence of T_i from x_i and $Y_i(t)$, we have:

$$\mathbb{E}[x_i|T_i = t] = \mathbb{E}[x_i] = \bar{x}, \quad \forall\ t \in \{0, 1\}.$$

Thus, the population intercept simplifies to:

$$\alpha_t = \mathbb{E}[Y_i|T_i = t], \quad \forall\ t \in \{0, 1\}.$$

Next, we establish the convergence of sample estimates to their population counterparts.

For $\hat{\alpha}_t \xrightarrow{p} \alpha_t$, we use the law of large numbers (LLN). Under the assumption of independent and identically distributed (i.i.d.) samples and finite second moments, sample means and variances converge in probability to their population counterparts.

The sample intercept is:

$$\hat{\alpha}_t = \bar{Y}_t - \hat{\beta}_t(\bar{x}_t - \bar{x}),$$

where:

- $\bar{Y}_t = \frac{1}{n_t} \sum_{i:T_i=t} Y_i$,
- $\bar{x}_t = \frac{1}{n_t} \sum_{i:T_i=t} x_i$.

By LLN (and again by randomized experiment), we have:

$$\bar{Y}_t \xrightarrow{p} \mathbb{E}[Y_i|T_i = t], \quad \bar{x}_t \xrightarrow{p} \mathbb{E}[x_i|T_i = t] = \bar{x}.$$

Moreover, the slope estimate converges:

$$\hat{\beta}_t = \frac{\sum_{i:T_i=t}(x_i - \bar{x})(Y_i - \bar{Y}_t)}{\sum_{i:T_i=t}(x_i - \bar{x})^2} \xrightarrow{p} \beta_t.$$

Thus, we conclude that:

$$\hat{\alpha}_t = \bar{Y}_t - \hat{\beta}_t(\bar{x}_t - \bar{x}) \xrightarrow{p} \mathbb{E}[Y_i|T_i = t] - \beta_t(\bar{x} - \bar{x}) = \mathbb{E}[Y_i|T_i = t] = \alpha_t.$$

Which concludes that $\hat{\alpha}_t$ converges in probability to $\alpha_t = \mathbb{E}[Y_i|T_i = t]$.

Finally, we demonstrate that $\hat{\tau}_{\text{pi}} \xrightarrow{p} \tau$.

Since we have:

$$\hat{\alpha}_1 \xrightarrow{p} \mathbb{E}[Y_i|T_i = 1], \quad \hat{\alpha}_0 \xrightarrow{p} \mathbb{E}[Y_i|T_i = 0],$$

we can compute the difference of intercepts:

$$\hat{\tau}_{\text{pi}} = \hat{\alpha}_1 - \hat{\alpha}_0 \xrightarrow{p} \mathbb{E}[Y_i|T_i = 1] - \mathbb{E}[Y_i|T_i = 0].$$

By randomization, we know that:

$$\mathbb{E}[Y_i|T_i = t] = \mathbb{E}[Y_i(t)], \quad \forall\, t \in \{0, 1\}.$$

Thus:

$$\hat{\tau}_{\text{pi}} \xrightarrow{p} \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = \tau.$$

Therefore, we have shown that:

$$\hat{\tau}_{\text{pi}} \xrightarrow{p} \tau = \mathbb{E}[Y(1) - Y(0)].$$

The assumptions used include randomization, the Stable Unit Treatment Value Assumption (SUTVA), and the consistency of the observed outcome with the potential outcome under the received treatment.

The consistency of $\hat{\alpha}_t$ is ensured by the independence of T_i from x_i and $Y_i(t)$, as well as the convergence of sample means and variances to their population counterparts.

In conclusion, by the law of large numbers and randomization, $\hat{\tau}_{\text{pi}}$ converges in probability to the true ATE, confirming that the regression-based plug-in estimator is valid for estimating causal effects in randomized experiments.

2.k Asymptotic distribution

Carefully derive the asymptotic distribution of $\hat{\tau}_{\text{pi}}$ and provide an estimator of the asymptotic variance. State your assumptions carefully and completely.

Necessary assumptions are:

1. Random Sampling and Independence: The data $\{(Y_i, X_i, T_i)\}_{i=1}^n$ are independent and identically distributed (i.i.d.) draws from the joint distribution of (Y, X, T).

2. Randomized Treatment Assignment: Treatment assignment T_i is independent of potential outcomes and covariates, with $\mathbb{P}[T_i = 1 \mid X_i] = p$, where $0 < p < 1$.

3. Stable Unit Treatment Value Assumption (SUTVA) and Consistency: No interference between units - the potential outcomes for one individual do not depend on the treatment status of individuals. The observed outcome corresponds to the potential outcome under the treatment received $[Y_i = Y_i(0)(1 - T_i) + Y_i(1)T_i]$.

4. Finite Moments: The potential outcomes $Y_i(0)$ and $Y_i(1)$ and the covariate X_i have finite second moments:
$$\mathbb{E}[Y_i(t)^2] < \infty, \quad \mathbb{E}[X_i^2] < \infty, \quad t \in \{0, 1\}.$$

5. Linear Model within Groups: For each treatment group $t \in \{0, 1\}$, the outcome model is correctly specified as:
$$Y_i = \alpha_t + \beta_t(X_i - \mu_X) + \varepsilon_i, \quad \text{for } T_i = t,$$
where $\mu_X = \mathbb{E}[X_i]$, and ε_i satisfies:
$$\mathbb{E}[\varepsilon_i \mid X_i] = 0, \quad \text{Var}[\varepsilon_i \mid X_i] = \sigma_t^2 < \infty.$$

6. Positive Variance of Covariates within Groups: The variance of X_i within each treatment group is positive:
$$\text{Var}[X_i \mid T_i = t] > 0, \quad t \in \{0, 1\}.$$

We begin by considering the separate ordinary least squares (OLS) regressions within each treatment group $t \in \{0, 1\}$:

$$y_i = \hat{\alpha}_t + \hat{\beta}_t(x_i - \bar{x}) + \hat{\varepsilon}_i, \quad \text{for } t_i = t,$$

Asymptotic Distribution of $\hat{\alpha}_t$

As derived in previous questions, for each treatment group t, the OLS estimator $\hat{\alpha}_t$ is given by:

$$\hat{\alpha}_t = \bar{Y}_t - \hat{\beta}_t(\bar{x}_t - \bar{x}),$$

Under the stated assumptions, we have:

- The sample means $\bar{Y}_t$ and $\bar{X}_t$ converge in probability to their population counterparts:

$$\bar{Y}_t \xrightarrow{p} \mathbb{E}[Y_i \mid T_i = t], \quad \bar{X}_t \xrightarrow{p} \mathbb{E}[X_i \mid T_i = t] = \mu_X.$$

- The OLS slope estimator $\hat{\beta}_t$ converges in probability to the population slope β_t:

$$\hat{\beta}_t \xrightarrow{p} \beta_t.$$

- The term $(\bar{X}_t - \bar{X})$ converges in probability to zero:

$$\bar{X}_t - \bar{X} \xrightarrow{p} \mu_X - \mu_X = 0.$$

Therefore, the asymptotic distribution of $\hat{\alpha}_t$ is primarily determined by the behavior of $\bar{Y}_t$.

By the Central Limit Theorem (CLT), we have:

$$\sqrt{n_t}(\bar{Y}_t - \mathbb{E}[Y_i \mid T_i = t]) \xrightarrow{d} \mathcal{N}(0, \sigma_t^2),$$

where $\sigma_t^2 = \text{Var}[Y_i \mid T_i = t]$.

Since $\hat{\alpha}_t$ differs from $\bar{Y}_t$ by a term that converges to zero in probability (due to $(\bar{X}_t - \bar{X}) \xrightarrow{p} 0$), we have:

$$\sqrt{n_t}(\hat{\alpha}_t - \alpha_t) = \sqrt{n_t}(\bar{Y}_t - \mathbb{E}[Y_i \mid T_i = t]) + o_p(1),$$

implying:

$$\sqrt{n_t}(\hat{\alpha}_t - \alpha_t) \xrightarrow{d} \mathcal{N}(0, \sigma_t^2).$$

Asymptotic Distribution of $\hat{\tau}_{\text{pi}}$

Recall that $\hat{\tau}_{\text{pi}} = \hat{\alpha}_1 - \hat{\alpha}_0$. Thus:

$$\sqrt{n}(\hat{\tau}_{\text{pi}} - \tau) = \sqrt{n}\left((\hat{\alpha}_1 - \alpha_1) - (\hat{\alpha}_0 - \alpha_0)\right) + \sqrt{n}(\alpha_1 - \alpha_0 - \tau).$$

Since $\alpha_t = \mathbb{E}[Y_i \mid T_i = t]$ and $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$, and by randomization:

$$\alpha_t = \mathbb{E}[Y_i(t)],$$

we have $\alpha_1 - \alpha_0 = \tau$, and thus:

$$\sqrt{n}(\alpha_1 - \alpha_0 - \tau) = 0.$$

Therefore:

$$\sqrt{n}(\hat{\tau}_{\text{pi}} - \tau) = \sqrt{n}\left((\hat{\alpha}_1 - \alpha_1) - (\hat{\alpha}_0 - \alpha_0)\right).$$

Since $n_t = p_t n$, where $p_t = \mathbb{P}[T_i = t]$, we can write:

$$\sqrt{n}(\hat{\alpha}_t - \alpha_t) = \sqrt{p_t n}(\hat{\alpha}_t - \alpha_t) \times \frac{1}{\sqrt{p_t}} \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_t^2}{p_t}\right).$$

Independence of $\hat{\alpha}_1$ and $\hat{\alpha}_0$

Due to random sampling and independent treatment assignment, the estimators $\hat{\alpha}_1$ and $\hat{\alpha}_0$ are asymptotically independent.

Combining the above results, we have:

$$\sqrt{n}(\hat{\tau}_{\text{pi}} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_1^2}{p_1} + \frac{\sigma_0^2}{p_0}\right).$$

Estimator of the Asymptotic Variance

To estimate the asymptotic variance, we need consistent estimators of σ_t^2.

Within each group t, we estimate σ_t^2 using the residuals from the OLS regression:

$$\hat{\sigma}_t^2 = \frac{1}{n_t - 2} \sum_{i:T_i=t} \left(Y_i - \hat{\alpha}_t - \hat{\beta}_t(X_i - \bar{X})\right)^2.$$

The adjustment in this case is by $n_t - 2$ as a consequence of the linear model with two parameters.

The asymptotic variance of $\hat{\tau}_{\text{pi}}$ is estimated by:

$$\widehat{\text{Var}}(\hat{\tau}_{\text{pi}}) = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}.$$

Since $n_t = p_t n$, we have:

$$\widehat{\text{Var}}(\hat{\tau}_{\text{pi}}) = \frac{\hat{\sigma}_1^2}{p_1 n} + \frac{\hat{\sigma}_0^2}{p_0 n}.$$

As $n \to \infty$, $\hat{\sigma}_t^2 \xrightarrow{p} \sigma_t^2$, ensuring that $\widehat{\text{Var}}(\hat{\tau}_{\text{pi}})$ is a consistent estimator of the asymptotic variance.

Under the stated assumptions, the plug-in estimator $\hat{\tau}_{\text{pi}}$ has the following asymptotic distribution:

$$\sqrt{n}(\hat{\tau}_{\text{pi}} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_1^2}{p_1} + \frac{\sigma_0^2}{p_0}\right),$$

and the asymptotic variance can be consistently estimated using:

$$\widehat{\text{Var}}(\hat{\tau}_{\text{pi}}) = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}.$$

This allows us to construct confidence intervals and perform hypothesis tests regarding the average treatment effect τ.

This is only possible because:

- The independence of treatment assignment ensures that $\hat{\alpha}_1$ and $\hat{\alpha}_0$ are asymptotically independent.
- The finite variance assumption for the errors ε_i guarantees the applicability of the Central Limit Theorem.
- The correct specification of the linear model within each group is crucial for the consistency and asymptotic normality of the OLS estimators.

2.1 Efficiency gain

Assume that, for $t = \{0, 1\}$, $Y(t) = \alpha_t + \beta_t X + \varepsilon_t$, with $\mathbb{E}[\varepsilon_t \mid X] = 0$ and $\mathbb{E}[\varepsilon_t^2 \mid X] = \sigma_t^2$. Prove that the asymptotic variance from part (k) is weakly smaller than that of (e), i.e., prove that $\hat{\tau}_{\text{pi}}$ is at least as efficient as the difference in means. Under what conditions is $\hat{\tau}_{\text{pi}}$ strictly more efficient?

We are given that, for $t \in \{0, 1\}$,

$$Y_i(t) = \alpha_t + \beta_t X_i + \varepsilon_{i,t},$$

with $\mathbb{E}[\varepsilon_{i,t} \mid X_i] = 0$ and $\mathbb{E}[\varepsilon_{i,t}^2 \mid X_i] = \sigma_t^2$. Our goal is to prove that the asymptotic variance of the plug-in estimator $\hat{\tau}_{\text{pi}}$ is weakly smaller than that of the difference-in-means estimator $\hat{\tau}_{\text{DM}}$.

Asymptotic Variance of $\hat{\tau}_{\text{DM}}$:

The difference-in-means estimator is defined as:

$$\hat{\tau}_{\text{DM}} = \bar{Y}_1 - \bar{Y}_0,$$

where $\bar{Y}_t = \frac{1}{n_t} \sum_{i:T_i=t} Y_i$, and n_t is the number of observations in group t.

Since treatment assignment is random and independent of X_i, the distribution of X_i is the same across treatment groups. Therefore, we can compute the variance of Y_i within each group:

$$\begin{aligned} \operatorname{Var}(Y_i \mid T_i = t) &= \operatorname{Var}(\alpha_t + \beta_t X_i + \varepsilon_{i,t} \mid T_i = t) \\ &= \beta_t^2 \operatorname{Var}(X_i) + \operatorname{Var}(\varepsilon_{i,t} \mid T_i = t) \\ &= \beta_t^2 \operatorname{Var}(X_i) + \sigma_t^2, \end{aligned}$$

since $\varepsilon_{i,t}$ is independent of X_i and has variance σ_t^2.

The asymptotic variance of $\hat{\tau}_{\text{DM}}$ is then:

$$\operatorname{Var}(\hat{\tau}_{\text{DM}}) = \frac{\operatorname{Var}(Y_i \mid T_i = 1)}{n_1} + \frac{\operatorname{Var}(Y_i \mid T_i = 0)}{n_0} = \frac{\beta_1^2 \operatorname{Var}(X_i) + \sigma_1^2}{n_1} + \frac{\beta_0^2 \operatorname{Var}(X_i) + \sigma_0^2}{n_0}.$$

Asymptotic Variance of $\hat{\tau}_{\text{pi}}$:

From previous results, the asymptotic variance of the plug-in estimator is:

$$\operatorname{Var}(\hat{\tau}_{\text{pi}}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}.$$

Comparison of Variances:

Compute the difference between the asymptotic variances:

$$\begin{aligned}\mathrm{Var}(\hat{\tau}_{\mathrm{DM}}) - \mathrm{Var}(\hat{\tau}_{\mathrm{pi}}) &= \left(\frac{\beta_1^2 \,\mathrm{Var}(X_i) + \sigma_1^2}{n_1} + \frac{\beta_0^2 \,\mathrm{Var}(X_i) + \sigma_0^2}{n_0}\right) - \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}\right) \\ &= \frac{\beta_1^2 \,\mathrm{Var}(X_i)}{n_1} + \frac{\beta_0^2 \,\mathrm{Var}(X_i)}{n_0}.\end{aligned}$$

Since $n_t > 0$ and $\mathrm{Var}(X_i) \geq 0$, we have:

$$\mathrm{Var}(\hat{\tau}_{\mathrm{DM}}) - \mathrm{Var}(\hat{\tau}_{\mathrm{pi}}) \geq 0.$$

Therefore,

$$\mathrm{Var}(\hat{\tau}_{\mathrm{pi}}) \leq \mathrm{Var}(\hat{\tau}_{\mathrm{DM}}).$$

Conditions for Strict Inequality

The inequality is strict unless both $\beta_1 = 0$ and $\beta_0 = 0$. Specifically:

- If at least one of β_1 or β_0 is non-zero and $\mathrm{Var}(X_i) > 0$, then $\mathrm{Var}(\hat{\tau}_{\mathrm{DM}}) > \mathrm{Var}(\hat{\tau}_{\mathrm{pi}})$.
- If $\beta_1 = \beta_0 = 0$, then the difference becomes zero, and the variances are equal.

The plug-in estimator $\hat{\tau}_{\mathrm{pi}}$ adjusts for the covariate X_i by estimating and removing the linear relationship between X_i and Y_i within each treatment group. This adjustment reduces the variability of the estimator when there is an association between X_i and Y_i (i.e., when $\beta_t \neq 0$).

In conclusion, Under the specified model and assumptions, $\hat{\tau}_{\mathrm{pi}}$ is at least as efficient as $\hat{\tau}_{\mathrm{DM}}$ and $\hat{\tau}_{\mathrm{pi}}$ is strictly more efficient than $\hat{\tau}_{\mathrm{DM}}$ when there is variability in X_i ($\mathrm{Var}(X_i) > 0$) and at least one of the slope coefficients β_t is non-zero.

Adjusting for covariates that are predictive of the outcome can reduce the variance of the estimated treatment effect.

3 Multiple Treatments

Consider studying two different binary treatments, $S \in \{0,1\}$ and $T \in \{0,1\}$. Assume we have access to experimental data where both S and T have been randomly assigned with constant (possibly different) probabilities, and an outcome Y and a pre-treatment covariate X are measured. The sample data are thus $(t_i, s_i, y_i, x_i), i = 1, \ldots, n$.

3.a Defining potential outcomes

Carefully define the potential outcomes in this setting. Make any assumptions explicit.

In this setting, we consider two binary treatments $S \in \{0,1\}$ and $T \in \{0,1\}$, an outcome Y, and a pre-treatment covariate X. For each unit $i = 1, \ldots, n$, we define the potential outcomes as:

$$Y_i(s,t), \quad \text{for } s \in \{0,1\},\ t \in \{0,1\}.$$

These potential outcomes represent the value of the outcome Y that would be observed for unit i under each possible combination of the treatments S and T. Specifically, there are four potential outcomes for each unit:

$$Y_i(0,0), \quad Y_i(0,1), \quad Y_i(1,0), \quad Y_i(1,1).$$

Assumptions

- Stable Unit Treatment Value Assumption (SUTVA):
  - No interference: The potential outcome for unit i depends only on its own treatments S_i and T_i, not on the treatments assigned to other units.
  - Consistency: The observed outcome Y_i corresponds to the potential outcome under the observed treatments:
$$Y_i = Y_i(s_i, t_i).$$
- Random Assignment:
  - The treatments S and T are randomly assigned independently of each other, the potential outcomes, and the pre-treatment covariate X:
$$(S_i, T_i) \perp\!\!\!\perp (Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), X_i)\,.$$
  - The probabilities of assignment to treatments S and T are constant across units but may differ between S and T.

Under these definitions and assumptions, the potential outcomes framework allows us to analyze the causal effects of treatments S and T on the outcome Y while accounting for the pre-treatment covariate X.

3.b Linear regression model for potential outcomes

Write down a linear regression model to recover the average of each potential outcome.

To recover the average of each potential outcome $E[Y(s,t)]$ for $s,t \in \{0,1\}$, we specify the following linear regression model:

$$Y_i = \beta_0 + \beta_S S_i + \beta_T T_i + \beta_{ST}(S_i \times T_i) + \varepsilon_i,$$

where:

- Y_i is the observed outcome for unit i.
- S_i and T_i are binary indicators of the treatments S and T, respectively.
- $S_i \times T_i$ is the interaction term between S_i and T_i.
- $\beta_0, \beta_S, \beta_T, \beta_{ST}$ are parameters to be estimated.
- ε_i is the error term with $E[\varepsilon_i | S_i, T_i] = 0$.

Under the assumptions of random assignment and SUTVA, the coefficients directly relate to the average potential outcomes:
Average Potential Outcome for $S = 0, T = 0$:

$$E[Y_i(0,0)] = \beta_0.$$

Average Potential Outcome for $S = 0, T = 1$:

$$E[Y_i(0,1)] = \beta_0 + \beta_T.$$

Average Potential Outcome for $S = 1, T = 0$:

$$E[Y_i(1,0)] = \beta_0 + \beta_S.$$

Average Potential Outcome for $S = 1, T = 1$:

$$E[Y_i(1,1)] = \beta_0 + \beta_S + \beta_T + \beta_{ST}.$$

By estimating β_0, β_S, β_T, and β_{ST} from the regression model, we can recover the averages of all four potential outcomes.

Optionally, to improve the precision of our estimates by accounting for the pre-treatment covariate X_i, we can extend the model:

$$Y_i = \beta_0 + \beta_S S_i + \beta_T T_i + \beta_{ST}(S_i \times T_i) + \gamma X_i + \varepsilon_i,$$

where γ captures the effect of X_i on Y_i. Nonetheless, we still expect to get average of each potential outcome without the addition of covariates because of the described assumptions above.

3.c Estimating the average treatment effect

Use estimates of the parameters in that model to estimate the average effect of treatment T. That is, for an individual drawn at random from the superpopulation, what is the estimated expected difference in their outcome under $T = 1$ versus $T = 0$?

To estimate the average effect of treatment T, denoted as ATE_T, we analyze the expected difference in the outcome Y when $T = 1$ versus $T = 0$ for an individual randomly drawn from the superpopulation. Utilizing the previously specified linear regression model:

$$Y_i = \beta_0 + \beta_S S_i + \beta_T T_i + \beta_{ST}(S_i \times T_i) + \varepsilon_i,$$

Under the potential outcomes framework, the expected outcomes conditional on treatment assignments are:

$$\begin{aligned}
\mathbb{E}[Y_i(T = 1) \mid S_i = s] &= \beta_0 + \beta_S s + \beta_T + \beta_{ST} s, \\
\mathbb{E}[Y_i(T = 0) \mid S_i = s] &= \beta_0 + \beta_S s.
\end{aligned}$$

The individual treatment effect of T for a unit with $S_i = s$ is the difference between these expected outcomes:

$$\mathbb{E}[Y_i(T = 1) - Y_i(T = 0) \mid S_i = s] = \beta_T + \beta_{ST} s.$$

To obtain the average treatment effect across the entire population, we take the expectation of the individual treatment effect over the distribution of S:

$$\text{ATE}_T = \mathbb{E}[\beta_T + \beta_{ST} S_i] = \beta_T + \beta_{ST}\mathbb{E}[S_i].$$

In practice, $\mathbb{E}[S_i]$ can be estimated by the sample mean $\overline{S}$. Therefore, the estimated average treatment effect of T is:

$$\hat{\text{ATE}}_T = \hat{\beta}_T + \hat{\beta}_{ST} \cdot \overline{S},$$

where:

- $\hat{\beta}_T$ is the estimated coefficient for treatment T,

- $\hat{\beta}_{ST}$ is the estimated coefficient for the interaction term $S \times T$,
- $\overline{S} = \frac{1}{n}\sum_{i=1}^{n} S_i$ is the sample mean of the treatment S.

The term $\hat{\beta}_T$ captures the average effect of treatment T when $S = 0$, while $\hat{\beta}_{ST} \cdot \overline{S}$ adjusts this effect based on the prevalence of $S = 1$ in the population. Together, they provide an estimate of the overall average effect of T across all units, accounting for the interaction between treatments S and T.

Under these assumptions, the estimator $\hat{\text{ATE}}_T$ provides an unbiased estimate of the true average treatment effect of T.

3.d Monte Carlo simulations

Using Monte Carlo simulations, compare this estimator to the simple difference in means estimator for treatment T. Compare in terms of bias, consistency, and efficiency. Explain what you find.

We aim to compare two estimators for the average treatment effect (ATE) of treatment T in the presence of another binary treatment S. The estimators under consideration are:

1. Regression-Based Estimator: Derived from a linear regression model that includes both treatments and their interaction.
2. Difference-in-Means Estimator: A simple comparison of the average outcomes between treated and control groups for T.

We will employ Monte Carlo simulations to assess these estimators in terms of bias, consistency, and efficiency.

Data Generating Process

We assume the following data generating process based on the potential outcomes framework:

$$Y_i = \beta_0 + \beta_S S_i + \beta_T T_i + \beta_{ST}(S_i \times T_i) + \gamma X_i + \varepsilon_i,$$

where:

- $S_i, T_i \sim \text{Bernoulli}(p_S), \text{Bernoulli}(p_T)$ are independently assigned binary treatments.
- $X_i \sim \mathcal{N}(0, 1)$ is a pre-treatment covariate.
- $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the error term.

For the simulation, we set the parameters as follow:

$$\beta_0 = 0, \quad \beta_S = 1, \quad \beta_T = 2, \quad \beta_{ST} = 0.5, \quad \gamma = 1, \quad \sigma = 1.$$

Estimators

1. **Regression-Based Estimator ($\hat{\text{ATE}}_T$)**:

$$\hat{\text{ATE}}_T = \hat{\beta}_T + \hat{\beta}_{ST} \cdot \overline{S},$$

   where $\overline{S}$ is the sample mean of S.

2. **Difference-in-Means Estimator ($\hat{\text{DIM}}_T$)**:

$$\hat{\text{DIM}}_T = \overline{Y}_{T=1} - \overline{Y}_{T=0},$$

   where $\overline{Y}_{T=1}$ and $\overline{Y}_{T=0}$ are the sample means of Y for treated and control groups, respectively.

R Code

```r
# Set seed for reproducibility
set.seed(123)

# Simulation parameters
n_sim <- 10000        # Number of simulations
n <- 1000             # Sample size per simulation
p_S <- 0.5            # Probability of S=1
p_T <- 0.5            # Probability of T=1

# True parameters
beta_0 <- 0
beta_S <- 1
beta_T <- 2
beta_ST <- 0.5
gamma <- 1
sigma <- 1

# Initialize vectors to store estimates
ATE_T_reg <- numeric(n_sim)
ATE_T_dim <- numeric(n_sim)

for (i in 1:n_sim) {
  # Generate treatments
  S <- rbinom(n, 1, p_S)
  T <- rbinom(n, 1, p_T)

  # Generate covariate
  X <- rnorm(n, 0, 1)

  # Generate outcome
  epsilon <- rnorm(n, 0, sigma)
  Y <- beta_0 + beta_S * S + beta_T * T + beta_ST * (S * T) + gamma * X + epsilon

  # Regression-based estimator
  model <- lm(Y ~ S + T + S:T)
  beta_hat <- coef(model)
  S_bar <- mean(S)
  ATE_T_reg[i] <- beta_hat["T"] + beta_hat["S:T"] * S_bar

  # Difference-in-means estimator
  Y_T1 <- Y[T == 1]
  Y_T0 <- Y[T == 0]
```

```r
  ATE_T_dim[i] <- mean(Y_T1) - mean(Y_T0)
}

# Calculate true ATE_T
# E[Y(T=1)] - E[Y(T=0)] = (beta_T + beta_ST * E[S]) - (0) = beta_T + beta_ST * p_S
true_ATE_T <- beta_T + beta_ST * p_S

# Calculate Bias
bias_reg <- mean(ATE_T_reg) - true_ATE_T
bias_dim <- mean(ATE_T_dim) - true_ATE_T

# Calculate Variance
var_reg <- var(ATE_T_reg)
var_dim <- var(ATE_T_dim)

# Calculate Mean Squared Error
mse_reg <- mean((ATE_T_reg - true_ATE_T)^2)
mse_dim <- mean((ATE_T_dim - true_ATE_T)^2)

# Summary of results
results <- data.frame(
  Estimator = c("Regression-Based", "Difference-in-Means"),
  Bias = c(bias_reg, bias_dim),
  Variance = c(var_reg, var_dim),
  MSE = c(mse_reg, mse_dim)
)

print(results)
```

Listing 1: Monte Carlo Simulation in R

Summary of Simulation Results

The simulation was run multiple times with different random seeds, yielding the following results:

Estimator	Bias	Variance	MSE
Regression-Based	0.002111760	0.008025591	0.008029248
Difference-in-Means	0.001988131	0.009521444	0.009524444
Regression-Based	0.0002000581	0.008124515	0.008123742
Difference-in-Means	0.0003602820	0.009840552	0.009839698
Regression-Based	-0.0006589664	0.008154055	0.008153674
Difference-in-Means	-0.0010183309	0.009717103	0.009717168
Regression-Based	-0.0004873721	0.008102807	0.008102964
Difference-in-Means	-0.0005380001	0.009653952	0.009654145

Bias

$$\text{Bias} = \mathbb{E}[\hat{\theta}] - \theta$$

Regression-Based Estimator: The biases observed are very close to zero (ranging from approximately -0.00066 to 0.00211), indicating that this estimator is unbiased or exhibits negligible bias across different simulations.

Difference-in-Means Estimator: Similarly, biases are near zero (ranging from approximately -0.00102 to 0.00199), confirming that this estimator is also unbiased under the given simulation setup.

Variance

$$\text{Variance} = \mathbb{V}(\hat{\theta})$$

Regression-Based Estimator: The variances are consistently around 0.0081, slightly lower across all simulations.

Difference-in-Means Estimator: The variances are consistently higher, around 0.0095 to 0.0098.

Mean Squared Error (MSE)

$$\text{MSE} = \text{Bias}^2 + \text{Variance}$$

Regression-Based Estimator: MSE values align closely with the variances due to the minimal bias, averaging around 0.0081.

Difference-in-Means Estimator: MSE values are slightly higher, around 0.0095 to 0.0098, reflecting both the higher variance and similar bias.

Consistency An estimator is considered consistent if, as the sample size increases, the estimator converges in probability to the true value of the parameter being estimated. In order to test that, we do simulations of different sizes.

```r
n_max <- 2450

ATE_T_dim <- numeric(n_max)
ATE_T_reg <- numeric(n_max)

for (n in 1:n_max){
    # Generate treatments
    S <- rbinom(n*10, 1, p_S)
    T <- rbinom(n*10, 1, p_T)

    # Generate covariate
    X <- rnorm(n*10, 0, 1)

    # Generate outcome
    epsilon <- rnorm(n*10, 0, sigma)
    Y <- beta_0 + beta_S * S + beta_T * T + beta_ST * (S * T) + gamma * X + epsilon

    # Regression-based estimator
    model <- lm(Y ~ S + T + S:T)
    beta_hat <- coef(model)
    S_bar <- mean(S)
    ATE_T_reg[n] <- beta_hat["T"] + beta_hat["S:T"] * S_bar

    # Difference-in-means estimator
    Y_T1 <- Y[T == 1]
    Y_T0 <- Y[T == 0]
    ATE_T_dim[n] <- mean(Y_T1) - mean(Y_T0)
}
```

Listing 2: Consistency Monte Carlo Simulation in R

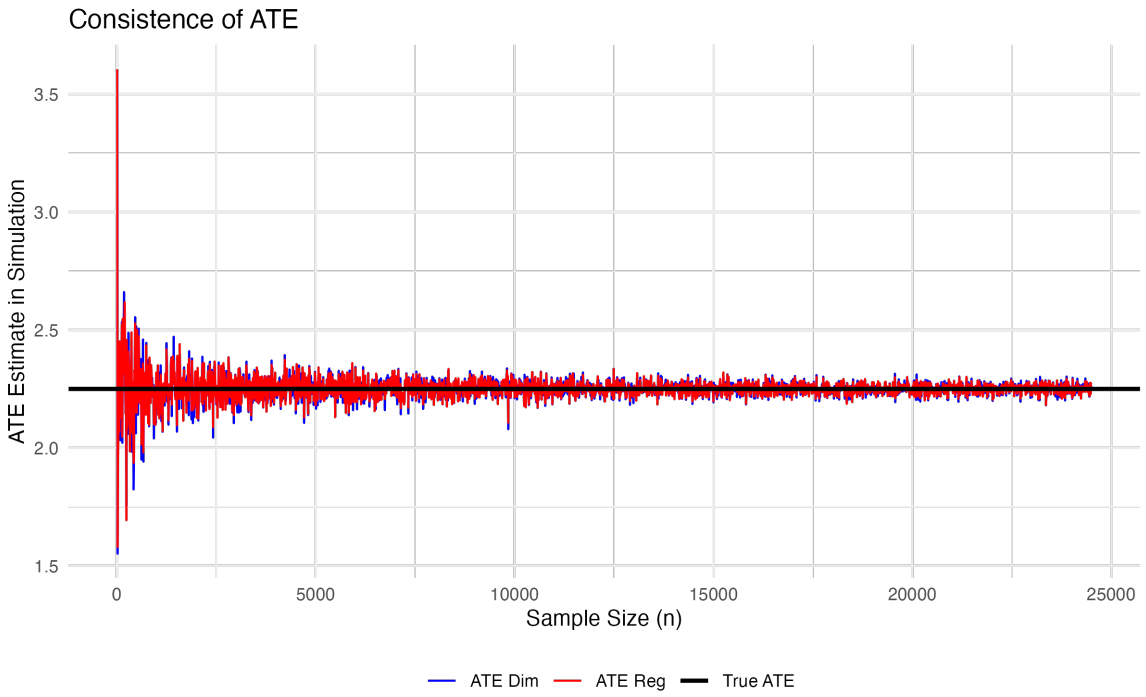

Figure 1: Consistency (ATE Estimate vs. Sample Size)

The simulation agrees with our theoretical conclusion that both estimates are consistents.

Conclusion

Both estimators exhibit negligible bias across all simulation runs, validating their unbiased nature under random treatment assignment and the model specifications.

The Regression-Based Estimator consistently shows lower variance compared to the Difference-in-Means Estimator. This suggests that the regression-based approach is more efficient, providing estimates with less variability across different samples.

Since both estimators are unbiased, the MSE primarily reflects the variance. The regression-based estimator's lower MSE further corroborates its higher efficiency.

This enhanced efficiency of the regression-based approach can be attributed to its utilization of additional information from the covariates and interaction terms. These findings align with theoretical expectations that leveraging a more comprehensive model structure, including interactions and covariates, can yield more efficient estimators in causal inference settings.

Researchers should consider modeling interactions and adjusting for relevant covariates to achieve more reliable and efficient estimates.

4 Linear Models and Interactions

We are studying a marketing campaign for a website. We have data from 200 different markets (think of billboards advertising the website placed in different cities) and we observe:

- **visitors** = Total visitors to the website (in hundreds of thousands) coming from that market,
- **spend** = Total spent (in thousands of dollars) on advertising in that market.

This data is on Canvas in the file `marketing_data.csv`. Assume throughout that changes in spending *cause* changes in visitor numbers. The question is whether the spending on advertising is helping or hurting website traffic and by how much.

We start with the following (heteroskedastic) regression model with independent observations:

$$visitors_i = \beta_0 + \beta_T \times spend_i + \varepsilon_i, \quad \varepsilon \sim \mathcal{N}(0, \sigma_i^2), \quad \text{correlation}(spend_i, \varepsilon_i) = 0.$$

4.a Scatter plot and interpretation

Show a scatter plot of the data along with the fitted regression line from this model. Give a precise and numerical causal interpretation of the results of running this linear regression. How does spending cause visitors to change? Include a discussion of the statistical significance.

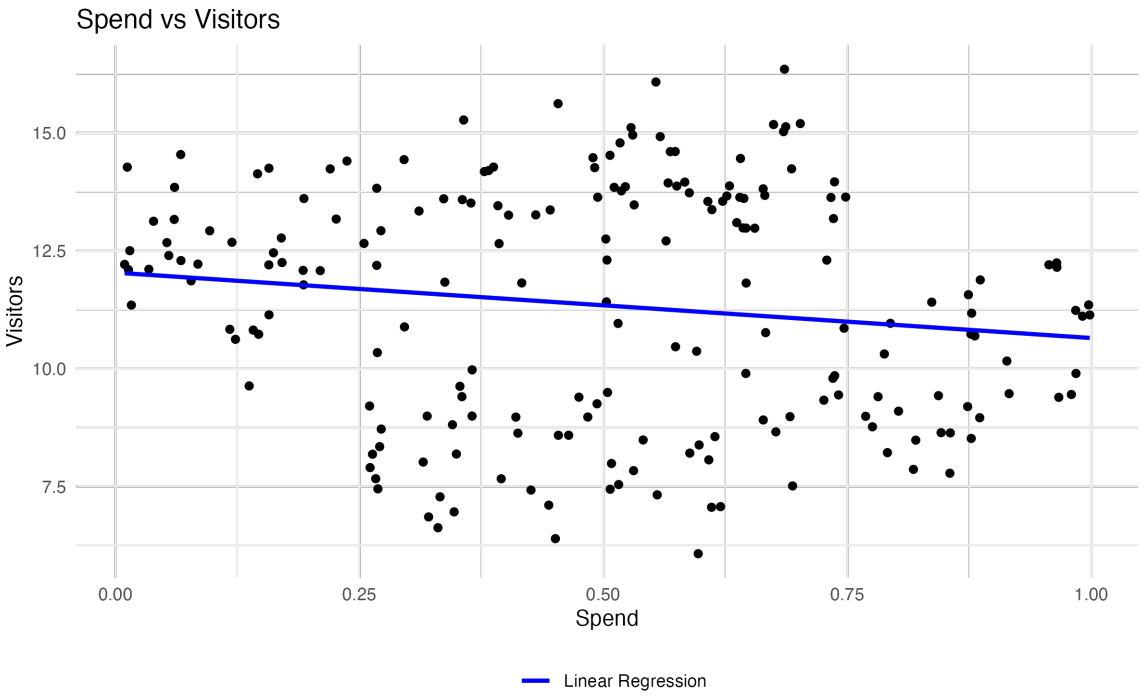

Figure 2: Spending vs. Visitors

The scatter plot illustrates the relationship between advertising spending (`spend`) and the number of website visitors (`visitors`) across 200 different markets. The fitted regression line shows the estimated linear relationship derived from the model:

$$\text{visitors}_i = \beta_0 + \beta_T \times \text{spend}_i + \varepsilon_i$$

```r
library(tidyverse)
library(readxl)

marketing_data <- read.csv("marketing_data.csv", sep=",") %>%
  as_tibble() %>%
  select(-X)

# General Regression and Plot
ols_visitors_spending <- lm(visitors ~ spend, data = marketing_data)
summary(ols_visitors_spending)

# Call:
# lm(formula = visitors ~ spend, data = marketing_data)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -5.1353 -2.1523  0.3405  2.1699  5.2667
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
```

```r
# (Intercept)  12.0359     0.3813   31.57   <2e-16 ***
# spend        -1.3877     0.6768   -2.05   0.0416 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 2.477 on 198 degrees of freedom
# Multiple R-squared:  0.02079,	Adjusted R-squared:  0.01585
# F-statistic: 4.205 on 1 and 198 DF,  p-value: 0.04163

# By City Regression and Plot
marketing_data %>%
  ggplot(aes(x=spend, y=visitors)) +
  geom_point() +
  labs(
    title="Spend vs Visitors",
    x="Spend",
    y="Visitors"
  ) +
  geom_smooth(aes(color = "Linear Regression"), method = "lm", se = FALSE) +
  theme_minimal() +
  scale_color_manual(name = "", values = c("Linear Regression" = "blue")) +
  theme(legend.position = "bottom")
ggsave("marketing_spending_vs_visitors.png", width = 8, height = 5, dpi = 300)
```

Listing 3: Linear Model for Visitors

From the regression output, the estimated coefficients are:

$$\hat{\beta}_0 = 12.0359 \quad (\text{Std. Error: } 0.3813,\ t = 31.57,\ p < 2 \times 10^{-16})$$
$$\hat{\beta}_T = -1.3877 \quad (\text{Std. Error: } 0.6768,\ t = -2.05,\ p = 0.0416)$$

The intercept $\hat{\beta}_0 = 12.0359$ represents the estimated number of visitors (in hundreds of thousands) when advertising spending is zero. This estimate is statistically significant with a very low p-value ($p < 2 \times 10^{-16}$), indicating strong evidence against the null hypothesis that $\beta_0 = 0$.

The coefficient for `spend`, $\hat{\beta}_T = -1.3877$, suggests that for each additional thousand dollars spent on advertising, the number of website visitors decreases by approximately 1.3877 hundred thousand (i.e., 138,770 visitors). The p-value associated with $\hat{\beta}_T$ is 0.0416, which is below the conventional significance level of 0.05. Therefore, we reject the null hypothesis that $\beta_T = 0$, concluding that advertising spending has a statistically significant negative effect on website visitors.

Assuming that the model is correctly specified and that all relevant confounders are controlled for, the negative coefficient $\hat{\beta}_T$ can be interpreted causally. Specifically, increasing advertising spending by one thousand dollars is expected to result in a decrease of approximately 138,770 website visitors, holding all else constant.

Nonetheless, for this causal relationship to hold, assumptions regarding the data generating process and the data must be met. In this scenario we might have:

1. Observational data: we should suppose that this is not a randomized experiment. Therefore, it is hard to assume that:
$$Y_i(s) \perp S_i$$
2. Omitted variables: the data generating process can have other features which are not being represented here. If those features are post-treatment, they have influence in the estimated

effect of spending on visitors. Even if we had other variables, the causal relationship would only hold if:

$$Y_i(s) \perp S_i|X$$

Meaning that: $\mathbb{E}[Y(s)|S = s, X = x] = \mathbb{E}[Y(s)|X = x]$

3. Linear model: the data generating process might not be linear, which would state that for different levels of spending the impact on visitors might be different. Even if the model is not linear, if the previous two assumptions hold, we can still recover ATE. Nonetheless, in this last case, assuming linear model when the data generating process would say otherwise does not allow us to recover CATE.

When looking at the presented results above, a researcher should be cautious in stating that those assumptions hold. The small R^2 and the scatter plot suggest that collecting data about covariates should help develop a more reliable analysis.

4.b Increasing visitors

Using the results above and your answer in (a), if the goal is to increase the visitors to our website, should we raise or lower our spending?

If a researcher believes that the previous assumptions hold, the β_T represents the ATE. If that is the case, lowering the spending should increase the visitors in the website.

Nonetheless, as stated in the previous answer, one should be cautious to say that the assumptions above hold. The low R^2, observational data, possibility of omitted covariates pos-treatment, counterintuitive answer are valid concerns (all stated more clearly in 4.a) that should be considered before making a definitive decision.

4.c Group Regressions

Run a linear model of *visitors* on *spend* separately for each group separately. What do you find and how does this differ from what you found before? In your explanation, remember that spending causes visitors in this data.

---

city $\in \{0, 1\}$

We define $\text{city}_i = \text{city}_i - 1$. Therefore, the values that were originally 2 become 1 and the values that were originally 1 become 0. This provides more intuitive results in the following questions.

The analysis of the linear regression models conducted separately for small cities and big cities yields contrasting results compared to the initial regression performed on the entire dataset.

Prior to running the regression, we define $city_i = city_i - 1$, making $city \in \{0, 1\}$, which facilitates further analysis.

```r
ols_visitors_spending_small_cities <- lm(
  visitors ~ spend, data = marketing_data_small_cities
)
summary(ols_visitors_spending_small_cities)

# Call:
#   lm(formula = visitors ~ spend, data = marketing_data_small_cities)
#
# Residuals:
#   Min      1Q  Median      3Q     Max
# -3.2205 -0.6139 -0.0231  0.8664  2.3934
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)  12.2342     0.2196  55.704  < 2e-16 ***
#   spend         2.6232     0.4873   5.383 4.37e-07 ***
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 1.151 on 107 degrees of freedom
# Multiple R-squared:  0.2131,	Adjusted R-squared:  0.2057
# F-statistic: 28.97 on 1 and 107 DF,  p-value: 4.37e-07


ols_visitors_spending_big_cities <- lm(
  visitors ~ spend, data = marketing_data_big_cities
)
summary(ols_visitors_spending_big_cities)

# Call:
#   lm(formula = visitors ~ spend, data = marketing_data_big_cities)
#
# Residuals:
#   Min       1Q   Median       3Q      Max
# -2.84717 -0.85156 -0.05242  0.85707  2.60109
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)   6.7741     0.3517  19.259  < 2e-16 ***
#   spend         3.5916     0.5219   6.882 7.95e-10 ***
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 1.153 on 89 degrees of freedom
# Multiple R-squared:  0.3473,	Adjusted R-squared:   0.34
# F-statistic: 47.37 on 1 and 89 DF,  p-value: 7.952e-10
```

Listing 4: Linear Model for Visitors separeted by City Size

Regression Results for Small Cities

$$\text{visitors}_i = \beta_{0\text{S}} + \beta_{TS} \times \text{spend}_i + \varepsilon_i$$

$$\begin{aligned}
\text{visitors}_i &= 12.2342 + 2.6232 \times \text{spend}_i + \varepsilon_i, \\
\text{Std. Error} &= [0.2196,\ 0.4873], \\
t\text{-values} &= [55.704,\ 5.383], \\
p\text{-values} &= [< 2 \times 10^{-16},\ 4.37 \times 10^{-07}].
\end{aligned}$$

- The intercept $\hat{\beta}_0 = 12.2342$ represents the estimated number of visitors (in hundreds of thousands) when advertising spending is zero in small cities. This estimate is highly significant ($p < 2 \times 10^{-16}$).
- The coefficient for `spend`, $\hat{\beta}_{T\text{S}} = 2.6232$, indicates that for each additional thousand dollars spent on advertising in small cities, the number of website visitors increases by approximately 262,320 visitors. This positive relationship is highly statistically significant ($p = 4.37 \times 10^{-07}$).
- The model explains approximately 21.31% of the variance in visitor numbers ($R^2 = 0.2131$), suggesting a moderate fit.

Regression Results for Big Cities

$$\text{visitors}_i = \beta_{0\text{B}} + \beta_{T\text{B}} \times \text{spend}_i + \varepsilon_i$$

$$\begin{aligned} \text{visitors}_i &= 6.7741 + 3.5916 \times \text{spend}_i + \varepsilon_i, \\ \text{Std. Error} &= [0.3517,\ 0.5219], \\ t\text{-values} &= [19.259,\ 6.882], \\ p\text{-values} &= [< 2 \times 10^{-16},\ 7.95 \times 10^{-10}]. \end{aligned}$$

- The intercept $\hat{\beta}_0 = 6.7741$ represents the estimated number of visitors (in hundreds of thousands) when advertising spending is zero in big cities. This estimate is highly significant ($p < 2 \times 10^{-16}$).
- The coefficient for `spend`, $\hat{\beta}_{T\text{B}} = 3.5916$, suggests that for each additional thousand dollars spent on advertising in big cities, the number of website visitors increases by approximately 359,160 visitors. This positive relationship is highly statistically significant ($p = 7.95 \times 10^{-10}$).
- The model explains approximately 34.73% of the variance in visitor numbers ($R^2 = 0.3473$), indicating a better fit compared to the small cities model.

Comparison with the Initial Regression

In the initial regression conducted on the entire dataset, the coefficient for `spend` was negative ($\hat{\beta}_T = -1.3877$) and statistically significant ($p = 0.0416$), suggesting that increased advertising spending was associated with a decrease in website visitors, a highly contraintuitive result.

The difference can possibly be explained by:

- Omitted variable bias in the aggregate model: the aggregate model does not account for inherent differences between small and big cities. If we assume that the variable city size (X) allow us to state that
$$Y_i(s) \perp S_i | X,$$
we can recover the ATE and the regression with the aggregate dataset without considering X does not provide us the ATE.
- Lack of interaction term: the initial model does not include interaction term between spending and city size, not allowing for the possibility that the impact on different city sizes are different ($\beta_{T\text{B}} \neq \beta_{T\text{S}}$). Adding an interaction effect can help reconcile the differing effects observed.

- Interaction Effects: If big cities generally have higher advertising spending, the positive relationship within big cities might dominate the overall trend. However, if small cities have lower spending with less impact, the aggregation might reflect a complex interplay that results in a negative coefficient.

In conclusion, if separating the dataset by city makes $Y_i(s) \perp S_i$, meaning that $\mathbb{E}[Y(s)|S = k] = \mathbb{E}[Y(s)]$ for $s, k \in \Omega$, we can now retrieve the ATE for each kind of city. If that is the case, the causal relationship identified in the previous question is invalid and should be disconsidered. On the other hand, other omitted variables might still exist, not allowing us to do causal inference, even thought we are aware that `spending` causes `visitors`.

If this model is well specified, the causal relationship between spending and visitors is positive and the company should hope for an increase in visitors after having increased spending.

4.d Scatter plot with two regression lines

Show a scatter plot of the data with these two fitted regression lines, with the different market types shown in different colors.

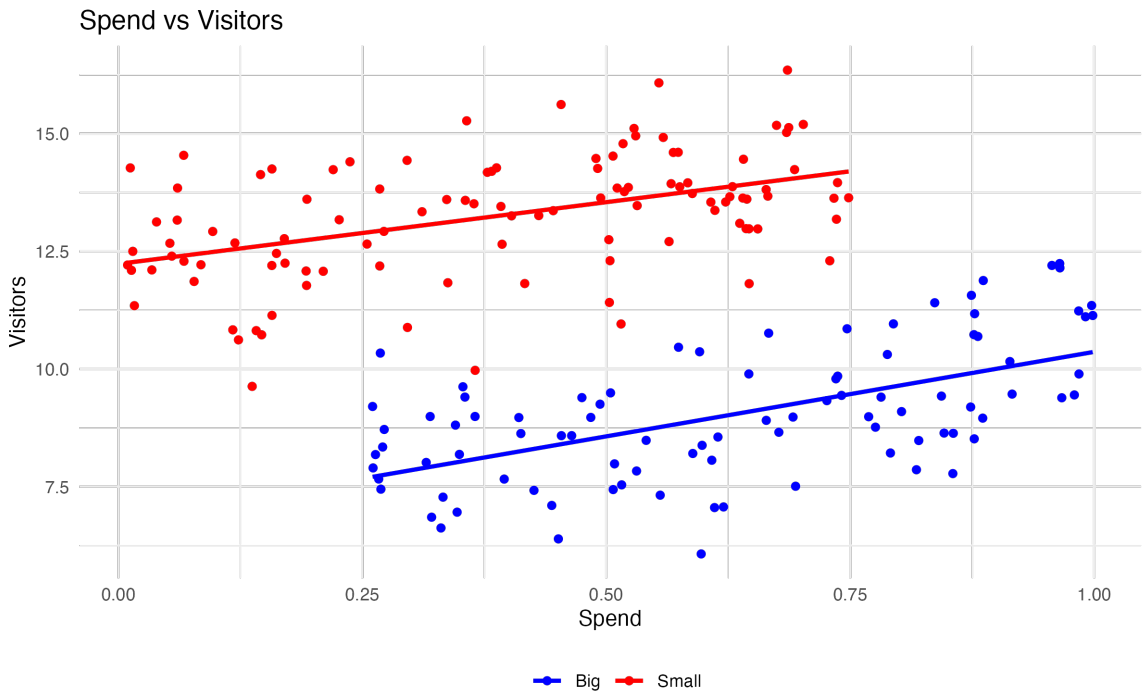

Figure 3: Spending vs. Visitors

4.e Single regression with interaction

Write down a single regression model using *visitors*, *spend*, and *city*, that when taken to the data would yield exactly the four coefficient estimates in part (c). Match up your coefficients to the estimates in (c).

$$\text{visitors}_i = \beta_{0\text{I}} + \beta_{1\text{I}} \times \text{spend}_i + \beta_{2\text{I}} \times \text{city}_i + \beta_{3\text{I}} \times (\text{spend}_i \times \text{city}_i) + \varepsilon_i$$

city $\in \{0, 1\}$

Again, we define $\text{city}_i = \text{city}_i - 1$. Therefore, the values that were originally 2 become 1 and the values that were originally 1 become 0. This provides more intuitive results.

```r
# Multivariate Regression
marketing_data_with_interaction <- marketing_data %>%
  mutate(interaction_spend_city = (spend) * city)

ols_visitors_with_interaction <- lm(
  visitors ~ ., data = marketing_data_with_interaction
)
summary(ols_visitors_with_interaction)

# Call:
#   lm(formula = visitors ~ ., data = marketing_data_with_interaction)
#
# Residuals:
#   Min      1Q  Median      3Q     Max
# -3.2205 -0.7589 -0.0294  0.8690  2.6011
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)             12.2342     0.2199  55.643  < 2e-16 ***
#   spend                    2.6232     0.4879   5.377 2.14e-07 ***
#   city                    -5.4601     0.4144 -13.176  < 2e-16 ***
#   interaction_spend_city   0.9683     0.7139   1.356    0.177
# ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 1.152 on 196 degrees of freedom
# Multiple R-squared:  0.7903,	Adjusted R-squared:  0.7871
# F-statistic: 246.2 on 3 and 196 DF,  p-value: < 2.2e-16
```

Listing 5: Linear Model for Visitors separeted by City Size

$$\hat{\text{visitors}}_i = 12.2342 + 2.6232 \times \text{spend}_i - 5.4601 \times \text{city}_i + 0.9684 \times (\text{spend}_i \times \text{city}_i)$$

- $\hat{\beta}_{0\text{I}} = 12.2342$: This is the intercept term representing the estimated number of visitors (in hundreds of thousands) when advertising spending is zero and the market is a small city ($\text{city}_i = 0$).

- $\hat{\beta}_{1\text{I}} = 2.6232$: This coefficient represents the effect of advertising spending on visitors in small cities. Specifically, for each additional thousand dollars spent on advertising in a small city, the number of website visitors increases by approximately 262,320.
- $\hat{\beta}_{2\text{I}} = -5.4601$: This coefficient captures the difference in the intercept between big cities and small cities. It indicates that, holding advertising spending constant, big cities have an estimated 546,010 fewer visitors compared to small cities when advertising spending is zero.
- $\hat{\beta}_{3\text{I}} = 0.9684$: This interaction term represents the additional effect of advertising spending on visitors in big cities. For each additional thousand dollars spent on advertising in a big city, the number of website visitors increases by approximately 96,840 beyond the effect observed in small cities.

Matching Coefficients to Group Regressions

We recover the regression fitted model for small cities is:

$$\widehat{\text{visitors}}_i = 12.2342 + 2.6232 \times \text{spend}_i$$

And the one for big cities:

$$\widehat{\text{visitors}}_i = 6.7741 + 3.5916 \times \text{spend}_i$$

The models can be combined into a single model:

$$\widehat{\text{visitors}}_i = c_i(6.7741 + 3.5916 \times \text{spend}_i) + (1 - c_i)(12.2342 + 2.6232 \times \text{spend}_i)$$

Where $c_i = 1$ if the city is big and $c_i = 0$ if the city is small. The result can be written as:

$$\begin{aligned}
\widehat{\text{visitors}}_i &= c_i(6.7741 + 3.5916 \times \text{spend}_i) + (1 - c_i)(12.2342 + 2.6232 \times \text{spend}_i) \\
&= c_i(6.7741 + 3.5916 \times \text{spend}_i) + 12.2342 + 2.6232 \times \text{spend}_i - c_i(12.2342 + 2.6232 \times \text{spend}_i) \\
&= 12.2342 + c_i(6.7741 - 12.2342) + 2.6232 \times \text{spend}_i + c_i(3.5916 \times \text{spend}_i - 2.6232 \times \text{spend}_i) \\
&= 12.2342 + c_i(6.7741 - 12.2342) + 2.6232 \times \text{spend}_i + c_i \times \text{spend}_i(3.5916 - 2.6232) \\
&= 12.2342 - 5.4601 c_i + 2.6232 \times \text{spend}_i + 0.9684 c_i \times \text{spend}_i \\
&= 12.2342 - 5.4601 \times \text{city}_i + 2.6232 \times \text{spend}_i + 0.9684 \times \text{city}_i \times \text{spend}_i \quad \text{(rewriting city)}
\end{aligned}$$

The final line is the model in part (4.e).

In other words, we combine the separate models by city into a single model with interaction.

- $\beta_{0\text{I}} = \beta_{0\text{S}} = 12.2342$
- $\beta_{1\text{I}} = \beta_{1\text{S}} = 2.6232$
- $\beta_{2\text{I}} = \beta_{0\text{B}} - \beta_{0\text{S}} = 6.7741 - 12.2342 = -5.4601$: represents the difference between the visitors in big and small cities considering zero spending.

- $\beta_{3\text{I}} = \beta_{1\text{B}} - \beta_{1\text{S}} = 3.5916 - 2.6232 = 0.9684$: represents the difference between the generated increased of spend in visitors for big and small cities.

city $\in \{1, 2\}$ The solution without modifying the city variable follows the same approach.

```
# Call:
# lm(formula = visitors ~ ., data = marketing_data_with_interaction)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -3.2205 -0.7589 -0.0294  0.8690  2.6011
#
# Coefficients:
#                         Estimate Std. Error t value Pr(>|t|)
# (Intercept)              17.6943     0.5628  31.439   <2e-16 ***
# spend                     1.6549     1.1062   1.496    0.136
# city                     -5.4601     0.4144 -13.176   <2e-16 ***
# interaction_spend_city    0.9683     0.7139   1.356    0.177
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 1.152 on 196 degrees of freedom
# Multiple R-squared:  0.7903,	Adjusted R-squared:  0.7871
# F-statistic: 246.2 on 3 and 196 DF,  p-value: < 2.2e-16
```

Listing 6: Combined Model without Modifications in City

$$\hat{\text{visitors}}_i = 17.6943 + 1.6549 \times \text{spend}_i - 5.4601 \times \text{city}_i + 0.9683 \times (\text{spend}_i \times \text{city}_i)$$

Again, we merge the models separated by city type to get the combined model:
Small cities:

$$\hat{\text{visitors}}_i = \hat{\beta}_{0,\text{S}} + \hat{\beta}_{1,\text{S}} \times \text{spend}_i$$

Big cities:

$$\hat{\text{visitors}}_i = \hat{\beta}_{0,\text{B}} + \hat{\beta}_{1,\text{B}} \times \text{spend}_i$$

The combined model becomes:

$$\hat{\text{visitors}}_i = (2 - \text{city}_i) \left[\hat{\beta}_{0,\text{S}} + \hat{\beta}_{1,\text{S}} \times \text{spend}_i\right] + (\text{city}_i - 1) \left[\hat{\beta}_{0,\text{B}} + \hat{\beta}_{1,\text{B}} \times \text{spend}_i\right]$$

Which can be viewed as:

$$\begin{aligned}\hat{\text{visitors}}_i = (2 - \text{city}_i) &\left[\hat{\beta}_{0,\text{S}} + \hat{\beta}_{1,\text{S}} \times \text{spend}_i\right] \quad \text{(becomes 0 when } city_i = 2 \text{ - big)} \\ &+ (\text{city}_i - 1) \left[\hat{\beta}_{0,\text{B}} + \hat{\beta}_{1,\text{B}} \times \text{spend}_i\right] \quad \text{(becomes 0 when } city_i = 1 \text{ - small)}\end{aligned}$$

The model can be rearranged to:

$$\begin{aligned}\hat{\text{visitors}}_i = &\left[2\hat{\beta}_{0,\text{S}} - \hat{\beta}_{0,\text{B}}\right] + \left[2\hat{\beta}_{1,\text{S}} - \hat{\beta}_{1,\text{B}}\right] \text{spend}_i \\ &+ \left[\hat{\beta}_{0,\text{B}} - \hat{\beta}_{0,\text{S}}\right] \text{city}_i + \left[\hat{\beta}_{1,\text{B}} - \hat{\beta}_{1,\text{S}}\right] \text{city}_i \times \text{spend}_i\end{aligned}$$

Here, we can rewrite (4.e) as parts of (4.c) again:

$$2\hat{\beta}_{0,\text{S}} - \hat{\beta}_{0,\text{B}} = 2(12.2342) - 6.7741 = 17.6943 = \hat{\beta}_{0,\text{I}}$$

$$2\hat{\beta}_{1,\text{S}} - \hat{\beta}_{1,\text{B}} = 2(2.6232) - 3.5916 = 1.6742 = \hat{\beta}_{1,\text{I}}$$

$$\hat{\beta}_{0,\text{B}} - \hat{\beta}_{0,\text{S}} = 12.2342 - 6.7741 = -5.4601 = \hat{\beta}_{2,\text{I}}$$

$$\hat{\beta}_{1,\text{B}} - \hat{\beta}_{1,\text{S}} = 3.5916 - 2.6232 = 0.9614 = \hat{\beta}_{3,\text{I}}$$

4.f Expression for β_T

Give an expression for β_T, from the model shown before part (a), in terms of the pieces of your model in part (e) and other features of the relationships between the three observed variables. Assume more things if you need to but be explicit. Use this expression to highlight what features of the data (i.e. the data generating process in the population) are leading to the differences between the separate fits in parts (c) and (e) compared to the combined fit in part (a). Demonstrate these using the data.

(For future reference, this phenomenon is called "Simpson Paradox")

---

city $\in \{0, 1\}$

Again, we define $\text{city}_i = \text{city}_i - 1$. Therefore, the values that were originally 2 become 1 and the values that were originally 1 become 0. This provides more intuitive results.

To express β_T from Model (4.a) in terms of the parameters from the model in (4.e), we begin by considering the definitions of both models.

Model 4.a:

$$\text{visitors}_i = \beta_0 + \beta_T \times \text{spend}_i + \varepsilon_i,$$

where β_T represents the average effect of spend_i on visitors_i across all markets.

Model 4.e:

$$\text{visitors}_i = \beta_{0\text{I}} + \beta_{1\text{I}} \times \text{spend}_i + \beta_{2\text{I}} \times \text{city}_i + \beta_{3\text{I}} \times \text{spend}_i \times \text{city}_i + \varepsilon_i.$$

Our goal is to derive an expression for β_T in terms of $\beta_{1\text{I}}$, $\beta_{2\text{I}}$, $\beta_{3\text{I}}$, and relevant statistical properties of the data.

We start by computing the covariance between visitors_i and spend_i in Model (4.e):

$$\mathbb{COV}(\text{visitors}_i, \text{spend}_i) = \mathbb{COV}(\beta_{0\text{I}} + \beta_{1\text{I}} \times \text{spend}_i + \beta_{2\text{I}} \times \text{city}_i + \beta_{3\text{I}} \times \text{spend}_i \times \text{city}_i + \varepsilon_i,\ \text{spend}_i)$$

$$\begin{aligned}\mathbb{COV}(\text{visitors}_i, \text{spend}_i) = {} & \beta_{1\text{I}} \times \mathbb{V}(\text{spend}_i) + \beta_{2\text{I}} \times \mathbb{COV}(\text{city}_i, \text{spend}_i) \\ & + \beta_{3\text{I}} \times \mathbb{COV}(\text{spend}_i \times \text{city}_i, \text{spend}_i) + \mathbb{COV}(\varepsilon_i, \text{spend}_i).\end{aligned}$$

Since ε_i is uncorrelated with spend_i, $\mathbb{COV}(\varepsilon_i, \text{spend}_i) = 0$. Therefore:

$$\mathbb{COV}(\text{visitors}_i, \text{spend}_i) = \beta_{1\text{I}} \times \mathbb{V}(\text{spend}_i) + \beta_{2\text{I}} \times \mathbb{COV}(\text{city}_i, \text{spend}_i) + \beta_{3\text{I}} \times \mathbb{COV}(\text{spend}_i \times \text{city}_i, \text{spend}_i).$$

In model (4.a), we have:

$$\beta_T = \frac{\mathbb{COV}(\text{visitors}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)}.$$

Substituting the expression for $\mathbb{COV}(\text{visitors}_i, \text{spend}_i)$, we get:

$$\beta_T = \beta_{1\text{I}} + \beta_{2\text{I}} \times \frac{\mathbb{COV}(\text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)} + \beta_{3\text{I}} \times \frac{\mathbb{COV}(\text{spend}_i \times \text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)}.$$

Using statistics calculated from data:

- $\beta_{1\text{I}} = 2.6232$
- $\beta_{2\text{I}} = -5.4601$
- $\beta_{3\text{I}} = 0.9683$
- $\mathbb{COV}(\text{city}_i, \text{spend}_i) = 0.06058889$
- $\mathbb{COV}(\text{spend}_i \times \text{city}_i, \text{spend}_i) = 0.0628975$
- $\mathbb{V}(\text{spend}_i) = 0.06729443$

Compute the ratios:

$$\frac{\mathbb{COV}(\text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)} = \frac{0.06058889}{0.06729443} \approx 0.9,$$

$$\frac{\mathbb{COV}(\text{spend}_i \times \text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)} = \frac{0.0628975}{0.06729443} \approx 0.935.$$

Now, calculate β_T:

$$\beta_T = 2.6232 + (-5.4601) \times 0.9 + 0.9683 \times 0.935,$$

$$\beta_T = 2.6232 - 4.91409 + 0.906,$$

$$\beta_T = (-2.29089) + 0.906 = -1.38489.$$

Thus, the expression for β_T in terms of the parameters from the model (4.e) is:

$$\beta_T = \beta_{1\text{I}} + \beta_{2\text{I}} \times \frac{\mathbb{COV}(\text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)} + \beta_{3\text{I}} \times \frac{\mathbb{COV}(\text{spend}_i \times \text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)}.$$

This result demonstrates how the covariance between city_i and spend_i, as well as the covariance between the interaction term $\text{spend}_i \times \text{city}_i$ and spend_i, influence the overall effect of spend_i on visitors_i in model (4.a).

Features of the Data Leading to Differences Between Models

The differences between the fits in Model E and model (4.a) arise due to the following features of the data:

- Correlation between city and spend: The covariance $\mathbb{COV}(\text{city}_i, \text{spend}_i) = 0.06058889$ indicates that advertising spend is higher in big cities ($\text{city}_i = 1$) compared to small cities. This positive correlation means that city markets tend to have higher advertising spend.
- Interaction effect between spend and city: The covariance $\mathbb{COV}(\text{spend}_i \times \text{city}_i, \text{spend}_i) = 0.0628975$ reflects how the effect of advertising spend on visitors differs between big and small cities. The positive covariance reinforces the interaction between spend_i and city_i significantly affects the total impact of spend_i on visitors_i.
- Variation in spend across cities: From the data, the average spend in big cities is higher than in small cities:
  $$\hat{\mu}_S : 0.390, \quad \hat{\mu}_B : 0.633.$$
  This difference contributes to the covariance terms and thus affects the estimation of β_T.
- Omission of city effects in model (4.a): model (4.a) does not account for the city-specific effects and interactions present in the data. As a result, it provides an average effect that blends the differing relationships in big and small cities, leading to a different estimate of β_T.

By incorporating the covariances and variances from the data into the expression for β_T, we see how the city-related variables influence the overall effect of advertising spend:

- The term $\beta_{2\text{I}} \times \frac{\mathbb{COV}(\text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)}$ adjusts for the fact that spend is higher in big cities and that big cities have a different baseline number of visitors (captured by $\beta_{2\text{I}}$).
- The term $\beta_{3\text{I}} \times \frac{\mathbb{COV}(\text{spend}_i \times \text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)}$ captures the differential effect of spend on visitors in big cities versus small cities.
- The combined effect shows that, despite $\beta_{1\text{I}}$ being positive (2.6232), the overall β_T becomes negative due to the substantial negative adjustment from the city-related terms.

In conclusion, this analysis highlights how the differences in advertising spend between big cities and small cities, along with the interaction effects, lead to the differences observed between the separate fits in model (4.a) and the combined fit in model (4.a).

Model without modification, city $\in \{1, 2\}$

If we consider city $\in \{1, 2\}$, the theorical solution is still the same:

$$\beta_T = \beta_{1\text{I}} + \beta_{2\text{I}} \times \frac{\mathbb{COV}(\text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)} + \beta_{3\text{I}} \times \frac{\mathbb{COV}(\text{spend}_i \times \text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)}.$$

We only find difference in the data:

- $\beta_{1\text{I}} = 1.6549$
- $\beta_{2\text{I}} = -5.4601$
- $\beta_{3\text{I}} = 0.9683$
- $\mathbb{COV}(\text{city}_i, \text{spend}_i) = 0.06058889$
- $\mathbb{COV}(\text{spend}_i \times \text{city}_i, \text{spend}_i) = 0.1301919$
- $\mathbb{V}(\text{spend}_i) = 0.06729443$

The modifications occur in $\beta_{1\text{I}}$ and $\mathbb{COV}(\text{spend}_i \times \text{city}_i, \text{spend}_i)$.

$$\beta_T = \beta_{1\text{I}} + \beta_{2\text{I}} \times \frac{\mathbb{COV}(\text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)} + \beta_{3\text{I}} \times \frac{\mathbb{COV}(\text{spend}_i \times \text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)}.$$

Compute the new ratios:

$$\frac{\mathbb{COV}(\text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)} = \frac{0.06058889}{0.06729443} \approx 0.9,$$

$$\frac{\mathbb{COV}(\text{spend}_i \times \text{city}_i, \text{spend}_i)}{\mathbb{V}(\text{spend}_i)} = \frac{0.1301919}{0.06729443} \approx 1.935.$$

$$\beta_T = 1.6549 + (-5.4601) \times 0.9 + 0.9683 \times 1.935 = -1.385529$$

Successfully, the result again matches.