Causal Mediation Analysis Assignment 1

Bin Yu

Apr 18, 2025

Question 1

Estimands

ITE

$$\text{ITE}_i \ = \ Y_i(1) - Y_i(0).$$

In our housing-voucher example, $Y_i(1)$ is the income five years later if i had received a voucher, and $Y_i(0)$ is the income if i had not. ITE is the difference in income for person i if they received a housing voucher ($Y_i(1)$) versus if they did not ($Y_i(0)$). In other words, ITE_i is the difference in individual i's potential incomeå if they had received a housing voucher ($a = 1$) versus if they had not ($a = 0$).

ATE

$$\text{ATE} = E\big[\text{ITE}_i\big] = E\big[Y_i(1) - Y_i(0)\big] = E\big[Y_i(1)\big] \ - \ E\big[Y_i(0)\big].$$

In our housing-voucher example, the ATE is the difference in average income five years later if all individuals had received a voucher versus if none had.

ATT

$$\text{ATT} = E\big[Y_i(1) - Y_i(0) \ \big| \ A_i = 1\big].$$

In our housing voucher example, the ATT represents the average income difference caused by the voucher for those people who actually got voucher. It is the difference of the mean of their actual income with income they would have had if they had not received the voucher. (Since we cannot observe the latter, we rely on assumptions to estimate this counterfactual)

The Fundamental Problem of Causal Inference

For any individual i, we only observe either $Y_i(1)$ (if they received the voucher) or $Y_i(0)$ (if they did not). We never observe both outcomes. For instance, if person i received the voucher, we see their income with the voucher ($Y_i(1)$) but cannot know what their income would have been without it ($Y_i(0)$). This missing data problem makes it fundamentally impossible to directly calculate causal effects for individuals.

Identification of ATE under Independence ($Y_i(0), Y_i(1) \perp A_i$)

If voucher assignment A_i is independent of potential incomes $(Y_i(0), Y_i(1))$, then:

$$E[Y_i(1)] = E[Y_i(1) \mid A_i = 1] = E[Y_i \mid A_i = 1],$$
$$E[Y_i(0)] = E[Y_i(0) \mid A_i = 0] = E[Y_i \mid A_i = 0].$$

Thus, the ATE simplifies to:

$$\text{ATE} = E[Y_i(1)] - E[Y_i(0)] = E[Y_i \mid A_i = 1] - E[Y_i \mid A_i = 0].$$

Interpretation in example: If voucher receipt is randomized, the income difference between voucher recipients and non-recipients directly estimates the ATE, which means that it could be estimated using observed data.

Identification of ATT under $Y_i(0) \perp A_i$

If the untreated outcome $Y_i(0)$ is independent of voucher receipt A_i, then:

$$E[Y_i(1) \mid A_i = 1] = E[Y_i \mid A_i = 1]$$

by definition.

and by $Y_i(0) \perp A_i$

$$E[Y_i(0) \mid A_i = 1] = E[Y_i(0)] = E[Y_i(0) \mid A_i = 0] = E[Y_i \mid A_i = 0].$$

The ATT becomes:

$$\text{ATT} = E\big[Y_i(1) - Y_i(0) \;\big|\; A_i = 1\big] = E[Y_i(1) \mid A_i = 1] - E[Y_i(0) \mid A_i = 1] = E[Y_i \mid A_i = 1] - E[Y_i \mid A_i = 0].$$

Interpretation in example: If the incomes of non-recipients ($A_i = 0$) reflect what recipients ($A_i = 1$) would have earned without the voucher, the observed income gap identifies the ATT.

Question 2

Decomposition of bias

$$\begin{aligned} E[Y_i \mid A_i = 1] \;-\; E[Y_i \mid A_i = 0] = {} & \underbrace{E[Y_i(1)] - E[Y_i(0)]}_{\text{ATE}} \\ & + \underbrace{\Big(E[Y_i(0) \mid A_i = 1] - E[Y_i(0) \mid A_i = 0]\Big)}_{\text{(i) baseline bias}} \\ & + \underbrace{\Big[\big(E[Y_i(1) - Y_i(0) \mid A_i = 1]\big) - \big(E[Y_i(1) - Y_i(0) \mid A_i = 0]\big)\Big]\, P(A_i = 0)}_{\text{(ii) heterogeneity bias}}. \end{aligned}$$

Explanation of Bias

- **Baseline bias:**

$$E[Y_i(0) \mid A_i = 1] - E[Y_i(0) \mid A_i = 0].$$

  This is the difference in what voucher-recipients would have earned without a voucher versus what non-recipients actually earned. In our example, individuals who successfully apply for vouchers may already be more motivated or skilled, so even they don't have a voucher, they would still have higher income.

- **Heterogeneity bias:**

$$\big(E[Y_i(1) - Y_i(0) \mid A_i = 1] - E[Y_i(1) - Y_i(0) \mid A_i = 0]\big)\, P(A_i = 0).$$

  This captures variation in the treatment effect itself: if those who receive vouchers gain more (or less) on average than those who do not, that difference distorts the mean comparison. In our example, the effect of a housing voucher may be larger for individuals who actually receive it (the treated group) than for those who do not (the control group), because treated individuals may be better able to take advantage of the opportunities provided by the voucher because they are systematically more skilled or well educated.

If ATT (the average treatment effect on the treated) exceeds ATC (the average treatment effect on the untreated), this heterogeneity induces additional bias, particularly when the proportion of untreated individuals, $P(A_i = 0)$, is large.

Question 3

Directed Acyclic Graph (DAG)

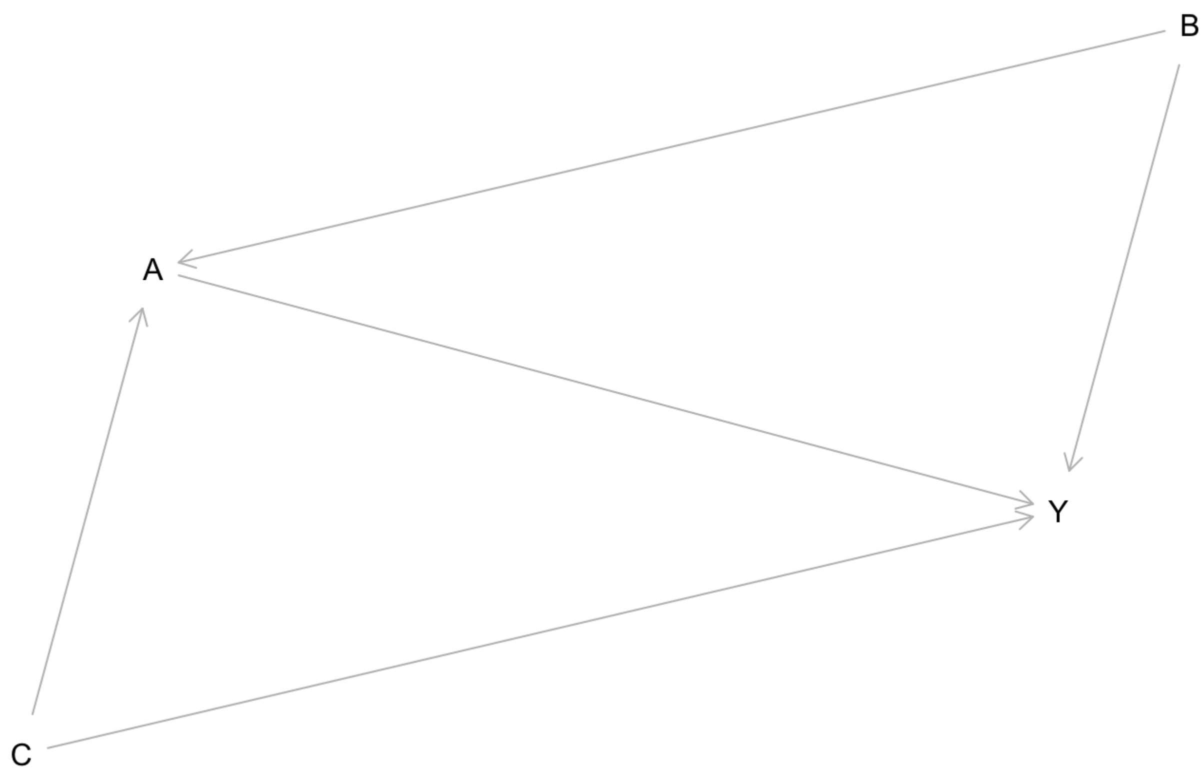

Figure 1: DAG

Identification of ATE

Using the following code:

```
adjustmentSets(housingDAG, "A","Y")
{ B, C }
```

In this DAG, all back-door paths from A to Y run through the common causes B and C, and both B and C are not descendants of A. Since both B and C are observed, the backdoor paths from A to Y are blocked by conditioning on (B, C). Hence the ATE $E[Y_i(1) - Y_i(0)]$ can be nonparametrically identified.

In other words, conditioning on B and C d-separates A and Y, blocking all backdoor paths (e.g., $A \leftarrow B \rightarrow Y$ and $A \leftarrow C \rightarrow Y$). This ensures:

$$Y(a) \perp\!\!\!\perp A \mid B, C \quad \text{for } a = 0, 1.$$

Nonparametric Expression for the ATE

The ATE can be estimated nonparametrically using the observed data:

$$\text{ATE} \;= E[Y(1)] - E[Y(0)] \;=\; E_{B,C}\Big[E\big[Y_i \mid A_i = 1,\, B_i = b,\, C_i = c\big] - E\big[Y_i \mid A_i = 0,\, B_i = b,\, C_i = c\big]\Big].$$

Since both B and C are discrete:

$$\text{ATE} = \sum_b \sum_c \Big(E[Y \mid A_i = 1, B_i = b, C_i = c] - E[Y \mid A_i = 0, B_i = b, C_i = c]\Big)\, P(B_i = b, C_i = c).$$

In words, for each combination of education b and family size c, we compare the observed mean income among voucher-recipients versus non-recipients, then average those differences according to the joint distribution of (B, C) in the population.

Question 4

DAG

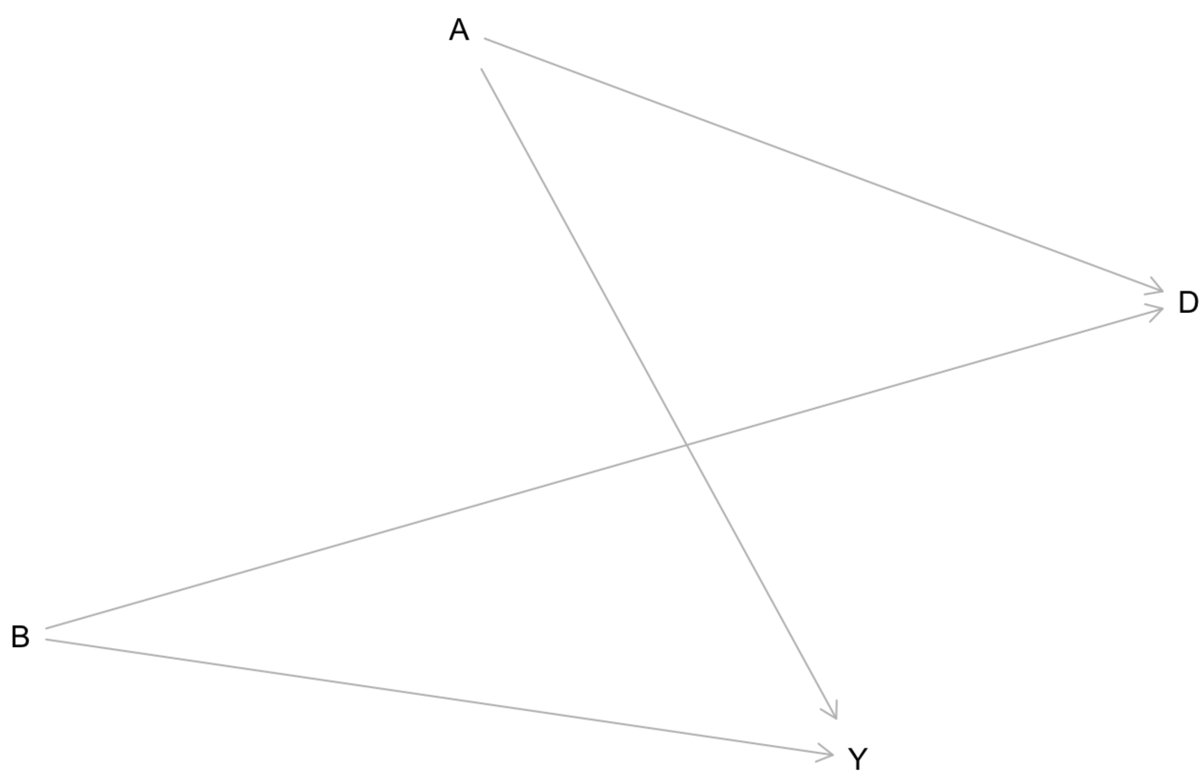

Figure 2: DAG

DAG Explanation

- **Nodes:**
  - A_i: Randomly assigned voucher receipt (treatment).
  - B_i: Education level (observed confounder).
  - D_i: Whether the individual was located for follow-up (missing data indicator).
  - Y_i: Income (outcome).
- **Edges:**
  - $A_i \to D_i$: Voucher receipt affects the likelihood of being located.
  - $B_i \to D_i$: Higher education increases the probability of moving away (reducing $D_i = 1$).
  - $A_i \to Y_i$: Voucher receipt directly affects income.
  - $B_i \to Y_i$: Education directly affects income.

Conditioned on D

The ATE cannot be identified after conditioning on D

Because vouchers were randomly assigned,

$$Y_i \perp A_i \quad \Longrightarrow \quad E[Y_i \mid A_i = 1] - E[Y_i \mid A_i = 0] \;=\; \text{ATE}.$$

However, D_i is a collider on the path

$$A_i \;\longrightarrow\; D_i \;\longleftarrow\; B_i \;\longrightarrow\; Y_i.$$

By the rules of d-separation:

Without conditioning on D_i, the path A_i—D_i—B_i—Y_i is blocked at D_i, so A_i and Y_i remain independent.

After conditioning on D_i, A_i and Y_i are not independent:

$$A_i \;\not\perp\; Y_i \mid D_i = 1$$

so

$$E[Y_i \mid A_i = 1, D_i = 1] - E[Y_i \mid A_i = 0, D_i = 1] \;\neq\; \text{ATE}.$$

Conditioned on B and D

Conditioned on both B and D can identify the ATE.

For the only back door path:

$$A_i \;\longrightarrow\; D_i \;\longleftarrow\; B_i \;\longrightarrow\; Y_i.$$

Without any conditioning, D_i is a collider and this path is blocked, so $A_i \perp Y_i$.

Conditioning on D_i makes A_i and Y_i not independent.

However, if we now also condition on B_i, we block the path$A_i \;\longrightarrow\; D_i \;\longleftarrow\; B_i \;\longrightarrow\; Y_i$ at B_i, because B_i is a confounder on the segment $D_i \leftarrow B_i \to Y_i$. Thus, by d-separation, conditioning on B_i blocks every path through B_i, making all back-door paths from A_i to Y_i blocked, and once we condition on $D_i = 1$ and $B_i = b$, A_i and Y_i become independent:

$$A_i \;\perp\; Y_i \;\Big|\; D_i = 1,\; B_i = b.$$

Thus,

$$ATE = \sum_b \Big(E[Y_i \mid A_i = 1,\, D_i = 1,\, B_i = b] - E[Y_i \mid A_i = 0,\, D_i = 1,\, B_i = b] \Big) P(B_i = b)$$

In other words, averaging these differences over the distribution of B_i identify the ATE.

Question 5

True ATE

$$Y_i(1) = 1 + B_i + U_{Y_i}$$
$$Y_i(0) = 0 + B_i + U_{Y_i}$$
$$\text{ATE} = E\big[Y_i(1) - Y_i(0)\big] = E\big[(1 + B_i + U_{Y_i}) - (B_i + U_{Y_i})\big] = E[1] = 1.$$

Modeling

Using the following code:

```
set.seed(42)

nSim <- 10000  # sampling time
n     <- 500   # sample size
est    <- numeric(nSim)

true_ATE <- 1

for(s in seq_len(nSim)) {

  U_A <- runif(n)
  U_B <- runif(n)
  U_D <- runif(n)
  U_Y <- rnorm(n)


  A <- as.integer(U_A <= 0.5)
  B <- as.integer(U_B <= 0.2)
  D <- as.integer(U_D <= (1 - 0.4*B - 0.2*A))
  Y <- A + B + U_Y

  # Conditioned on D = 1
  yt <- Y[A == 1 & D == 1]
  yc <- Y[A == 0 & D == 1]
  est[s] <- mean(yt) - mean(yc)
}

mean_est <- mean(est)
sd_est   <- sd(est)
bias     <- mean_est - true_ATE

cat("After", nSim, "times simulation\n")
cat("  mean ATE =", round(mean_est, 3), "\n")
cat("  std     =", round(sd_est,   3), "\n")
cat("  bias   =", round(bias,     3), "\n")
```

Result:

```
After 10000 times simulation
  mean ATE = 0.982
```

```
std     = 0.104
bias    = -0.018
```

Running 10000 simulations of size $n = 500$ under the specified data-generating process yielded the following summary for the estimator

$$\hat{\Delta} = \{\bar{Y}_i \mid A_i = 1,\, D_i = 1\} - \{\bar{Y}_i \mid A_i = 0,\, D_i = 1\}.$$

$$\begin{aligned} E[\hat{\Delta}] &\approx 0.982, \\ \mathrm{SD}(\hat{\Delta}) &\approx 0.104, \\ \mathrm{Bias}(\hat{\Delta}) &= E[\hat{\Delta}] - \mathrm{ATE} \;=\; 0.982 - 1 = -0.018. \end{aligned}$$

As we have proved in Question 4, only conditioned on D can not identify the true ATE. The simulation confirms that conditioning on being located ($D_i = 1$) induces bias and hence

$$E[Y_i \mid A_i = 1, D_i = 1] - E[Y_i \mid A_i = 0, D_i = 1]$$

is not an unbiased estimator for the true ATE of 1.

Question 6

Check the Covariates

To verify that random assignment balanced pre-treatment covariates, we conduct two-sample t-tests of each covariate by treatment status:

Covariate	Mean (Control)	Mean (Treated)	p-value
age	24.45	24.63	0.7216
educ	10.19	10.38	0.1443
black	0.800	0.801	0.9645
hisp	0.113	0.094	0.4155
married	0.158	0.168	0.7028
re74	3672.49	3571.00	0.8257
re75	3026.68	3066.10	0.9172

Table 1: Balance of Pre-treatment Covariates by Treatment Status

All p-values exceed 0.14, and the group means are nearly identical. This confirms that random assignment achieved balance on age, education, race/ethnicity, marital status, and pre-treatment earnings.

Experimental ATT Estimate

We estimate the treatment effect by the unadjusted regression

$$\mathrm{re78}_i = \beta_0 + \beta_1\, \mathrm{treated}_i + \varepsilon_i,$$

where treated $= 1$ for NSW participants in the experimental group.

The output is:

```
Call:
lm(formula = re78 ~ treated, data = nsw)
```

```
Residuals:
   Min     1Q Median     3Q    Max
 -5976  -5090  -1519   3361  54332

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5090.0      302.8  16.811   <2e-16 ***
treated        886.3      472.1   1.877   0.0609 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6242 on 720 degrees of freedom
  (18482 observations deleted due to missingness)
Multiple R-squared:  0.004872, Adjusted R-squared:  0.003489
F-statistic: 3.525 on 1 and 720 DF,  p-value: 0.06086
```

Relationship between ATE and ATT under Randomization

Under complete random assignment,

$$A_i \perp \{Y_i(0), Y_i(1)\} \quad \Longrightarrow \quad E[Y_i(0) \mid A_i = 1] = E[Y_i(0) \mid A_i = 0] \quad \text{and} \quad E[Y_i(1) \mid A_i = 1] = E[Y_i(1)].$$

Hence

$$\text{ATE} = E[Y_i(1) - Y_i(0)] = E[Y_i(1) - Y_i(0) \mid A_i = 1] = \text{ATT}.$$

Interpretation of ATT

By definition,

$$\text{ATT} = E\big[Y_i(1) - Y_i(0) \mid A_i = 1\big],$$

the ATT is the effect of the treatment on those actually treated. In the NSW example, $Y_i(1)$ is the 1978 earnings of individual i if assigned to job training, and $Y_i(0)$ is their (counterfactual) earnings if had not been trained. Our estimate $\hat{\beta}_1 = \$886.3$ therefore means:

On average, among those participants who received the NSW job-training program, their incomes in 1978 were \$886 higher than the incomes they would have if not received the training. This ATT estimate suggests a meaningful positive effect of the job-training intervention.

Question 7

Non-Experimental ATT Estimates

Unadjusted Model gives

```
> # Unadjusted non-experimental ATT
> non_exp_model <- lm(re78 ~ ntreated, data = nsw)
> summary(non_exp_model)

Call:
lm(formula = re78 ~ ntreated, data = nsw)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
 -15750   -9191    1264    9814  105423

Coefficients:
            Estimate  Std. Error t value Pr(>|t|)
(Intercept) 15750.30     79.65   197.74   <2e-16 ***
ntreated    -9773.95    633.36   -15.43   <2e-16 ***
---
Residual standard error: 10830 on 18777 DF
Multiple R-squared: 0.01252, Adjusted R-squared: 0.01247
F-statistic: 238.1 on 1 and 18777 DF, p-value: < 2.2e-16
```

Interpretation

We use the CPS/PSID samples to approximate the counterfactual mean $E[Y_i(0) \mid A_i = 1]$.

The unadjusted coefficient $\hat{\beta}_{\text{ntreated}} = -9773.95$ implies

$$E[Y \mid A = 1] \; - \; E[Y \mid A = 0] \; \approx \; -\$9{,}774.$$

that compare the NSW participants who actually received the training to the CPS/PSID comparison group—the treated group appears to earn \$9,774 less on average in 1978. If we believe this is the ATT, then on average, among those participants who received the NSW job-training program, their incomes in 1978 were \$9,774 lower than the incomes they would have if not received the training. However, this estimation does not reflect a real negative causal effect of the program.

This contrasts the experimental ATT of +\$886. The large negative bias arises because the CPS/PSID group is not exchangeable with NSW participants (maybe baseline incomes and composition differ).

The reason is that: in order to get unbiased ATT, for $E[Y \mid A = 0]$ from the CPS/PSID sample to equal the unobserved $E[Y(0) \mid A = 1]$, we would need the conditional independence:

$$Y_i(0) \;\perp\; A_i \quad \Longrightarrow \quad E[Y_i(0) \mid A_i = 1] = E[Y_i(0) \mid A_i = 0].$$

This key assumption fails here because:

Baseline bias: NSW participants differ systematically from CPS/PSID in education, motivation, and pre-treatment earnings, so $E[Y(0) \mid A = 1] \neq E[Y \mid \text{CPS/PSID}]$.

Heterogeneity bias: The true treatment gain $\tau_i = Y_i(1) - Y_i(0)$ also varies across these groups.

In fact, the CPS/PSID comparison group differs substantially from the NSW-treated group on key baseline covariates (e.g. education, pre-treatment earnings, demographic composition). Such imbalance implies that the difference $E[Y \mid A = 1] - E[Y \mid A = 0]$ is a mix of both the causal effect and the pre-existing differences in expected outcomes, thus it leads to bias.

As a result, the non-experimental ATT is heavily biased, both in sign and magnitude, comparing to the experimental estimate.

Adjusted ATT

Adjusted Model gives

```
> # Covariate-adjusted non-experimental ATT
```

```
> non_exp_model_1 <- lm(
+   re78 ~ ntreated + age + educ + black + hisp + married + re74 + re75,
+   data = nsw
+ )
> summary(non_exp_model_1)

Call:
lm(formula = re78 ~ ntreated + age + educ + black + hisp + married +
    re74 + re75, data = nsw)

Residuals:
    Min      1Q  Median      3Q     Max
 -60246   -3463     721    3541  110770

Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)  5.519e+03  3.281e+02  16.821   <2e-16 ***
ntreated    -1.036e+03  4.690e+02  -2.208   0.0272 *
age         -1.088e+02  5.874e+00 -18.524   <2e-16 ***
educ         1.746e+02  2.040e+01   8.561   <2e-16 ***
black       -3.481e+02  1.894e+02  -1.838   0.0661 .
hisp        -7.025e+01  2.270e+02  -0.309   0.7570
married      1.751e+02  1.438e+02   1.217   0.2235
re74         2.944e-01  1.111e-02  26.501   <2e-16 ***
re75         5.102e-01  1.112e-02  45.897   <2e-16 ***
---
Residual standard error: 7590 on 18770 DF
Multiple R-squared: 0.5151, Adjusted R-squared: 0.5149
F-statistic: 2492 on 8 and 18770 DF, p-value: <2.2e-16
```

Interpretation

Covariate adjustment rests on the conditional ignorability assumption:

$$\{Y_i(0),\, Y_i(1)\} \;\perp\; A_i \;\bigm|\; \mathbf{C}_i,$$

where

$$\mathbf{C}_i = \{\text{age}_i,\ \text{educ}_i,\ \text{black}_i,\ \text{hisp}_i,\ \text{married}_i,\ \text{re74}_i,\ \text{re75}_i\}.$$

This means that once we condition on each individual's baseline covariates $\mathbf{C}_i$, within any stratum defined by $\mathbf{C}_i$, the assignment A_i is as good as random, so treated and control units have the same distribution of $\{Y_i(0), Y_i(1)\}$.

In other word, this assumption means that once we "slice" our sample into groups that share the same values of the covariates $\mathbf{C}_i$, the treated and untreated units within each slice look alike in terms of their potential outcomes. In other words, within each covariate stratum (for example, 30-year-olds with a high-school diploma who were unmarried and earned \$3,000 in 1974 and \$2,500 in 1975), the distribution of what they would have earned without treatment, $Y_i(0)$, is the same whether or not they actually received the treatment:

$$P\bigl(Y_i(0) \le y \mid A_i = 1, \mathbf{C}_i = c\bigr) = P\bigl(Y_i(0) \le y \mid A_i = 0, \mathbf{C}_i = c\bigr).$$

This "balance" of the potential outcomes on the covariates $\mathbf{C}_i$ is precisely what allows us to use the observed outcomes Y_i and regression adjustment on $\mathbf{C}_i$ to recover an unbiased estimate of the causal effect. If such covariate balance fails (i.e. if $Y_i(0)$ or $Y_i(1)$ still depends on A_i after adjusting for $\mathbf{C}_i$), then no amount of regression or matching on $\mathbf{C}_i$ alone can fully remove the bias.

In practice, we fit

$$\widehat{Y}_i = \hat{\alpha} + \hat{\tau}\, A_i + \hat{\beta}_1\, \text{age}_i + \cdots + \hat{\beta}_7\, \text{re75}_i,$$

so that $\hat{\tau}$ consistently estimates $E[Y_i(1) - Y_i(0) \mid A_i = 1]$, provided the model for $f(\mathbf{C}_i)$ is correctly specified and no important confounders are omitted.

Comparison

After adjusting, the coefficient is $\hat{\beta}_{\text{ntreated}} \approx -\$1{,}036$. This estimate is still below zero but much closer to the experimental ATT than the unadjusted value.

- Comparing with unadjusted non-experimental estimate: from $-\$9{,}774$ to $-\$1{,}036$, much closer to the experimental ATT, shows that age, education, race/ethnicity, marital status and pre-treatment earnings explain most of the initial discrepancy.
- Comparing with experimental estimate: There is still a remaining gap between adjusted non-experimental estimate and experimental estimate of ATT ($-\$1{,}036$ vs. $+\$886$) reflects that there is still confounding by unobserved factors not captured by $\mathbf{C}_i$.

In sum, covariate adjustment substantially attenuates the non-experimental bias but does not fully recover the true causal effect without randomization since voucher applicants and control group members differ systematically on both observed and unobserved factors.

Question 8

Propensity-Score Weighted ATT Estimate

The IPT ATT estimate is given below:

```
Call:
svyglm(formula = re78 ~ ntreated, design = design_ipw)

Survey design:
svydesign(ids = ~1, data = nsw, weights = ~ipw_att)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6601.6      227.3  29.041   <2e-16 ***
ntreated      -625.3      461.0  -1.356    0.175
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 49900434)

Number of Fisher Scoring iterations: 2
```

Interpretation and Comparison

The IPT-weighted estimate of the ATT is $\hat{\tau}_{\text{IPW}} = -\$625.3$, which implies that, after re-weighting the CPS/PSID controls to match NSW-treated participants on age, education, race/ethnicity, marital status, and pre-treatment earnings, the treated group appears to earn \$625 less in 1978 than they would have if they had not received the training if we believe this is true ATT.

The IPT estimate is closer the experimental ATT (+$886) than the unadjusted non-experimental estimator ($9,774) and the covariate-adjusted estimator ($1,036), but still differs in sign and magnitude with experimental ATT.

General Conclusions

- By up-weighting control-group individuals whose covariates match the NSW-treated distribution, the IPT weights balance age, education, race/ethnicity, marital status, and prior earnings across groups, thereby mitigating selection bias.
- Re-weighting by these observed covariates reduces bias relative to the naive comparison, confirming that much of the non-experimental bias arose from baseline covariate imbalance.
- However, the remaining discrepancy with the true experimental ATT suggests confounding by unmeasured factors or possible model misspecification.
- This underscores that, even with careful weighting or regression adjustment, the observational data of CPS/PSID alone cannot fully replicate the unbiased inference of ATT afforded by random assignment.