Causal Mediation Analysis Assignment 3

Bin Yu

May 15, 2025

Question 1

Formal definition

Let $Y(d, m)$ denote the potential outcome for individuals if they were assigned treatment $D = d$ and the mediator M were set (intervened) to the fixed value m. The controlled direct effect of the job-training workshop is defined, for any fixed mediator level m, by

$$\text{CDE}(d, d^*, m) \;=\; E\big[Y(d, m)\big] \;-\; E\big[Y(d^*, m)\big]$$

which can be written as

$$\text{CDE}(1, 0, m) \;=\; E\big[Y(1, m)\big] \;-\; E\big[Y(0, m)\big]$$

in this example.

Interpretation

$E[Y(d, m)]$: the average probability of re-employment if everyone is given treatment d and all have their self-efficacy set to m.

$E[Y(d^*, m)]$: the average probability of re-employment if everyone is given treatment d^* and all have their self-efficacy set to the same m.

$\text{CDE}(d, d^*, m)$ is the expected difference in the outcome if individuals were exposed to d instead of d^*, and if everyone's mediator were forced to the same level m. It captures the effect of D on Y that remains holding the mediator constant at a fixed value for everyone.

In our example, let $d = 1$ denote assignment to the job-training workshop and $d^* = 0$ denote assignment to the booklet. Let $m \in \{1, 2, 3, 4, 5\}$ be the fixed level of job search self-efficacy. Then:

$$\text{CDE}(1, 0, 4) = E[\,Y(1, 4)\,] \;-\; E[\,Y(0, 4)\,].$$

It can be interpreted as: if we could force everyone's self-efficacy to level 4, participating in the workshop (versus reading the booklet) would increase the probability of working at least 20 hours/week by $\text{CDE}(1, 0, 4)$. Thus, comparing $\text{CDE}(1, 0, m)$ across different m values shows whether the workshop's direct impact varies with the level of self-efficacy.

Thus, the CDE in this example is the average increase in re-employment probability if everyone were switched from receiving booklet ($d^* = 0$) to workshop ($d = 1$), while holding each person's job-search self-efficacy fixed at the common value m. It isolates the portion of the workshop's effect that does not operate by changing self-efficacy.

Contrast with the Natural Direct Effect (NDE)

The natural direct effect differs from the CDE in how the mediator is handled:

- **CDE:** Fixes the mediator to a value m for everyone, and compares

$$\text{CDE}(d, d^*, m) = E[\,Y(d, m)\,] \;-\; E[\,Y(d^*, m)\,].$$

- **NDE:** Fixes the mediator to each individual's own counterfactual value under control, $M_i(0)$, and compares

$$\text{NDE}(d, d^*) = E\big[\,Y\big(d,\, M(0)\big)\big] \;-\; E\big[\,Y\big(d^*,\, M(0)\big)\big].$$

The Control direct effect fixed mediator's value, then calculate the direct effect of D on Y, while the natural direct effect (NDE) averages the $D \to Y$ effects over the distribution of $M(0)$, it may mask how the direct effect actually varies with different mediator (here is self-efficacy) levels, and it generally requires stronger "cross-world" assumptions to be identified when there are treatment–mediator interactions.

Thus, we have, except under special conditions (e.g. no D - M interaction),

$$\text{CDE}(d, d^*, m) \;\neq\; \text{NDE}(d, d^*).$$

We Need the Controlled Direct Effect (CDE) because:

Heterogeneity under D–M interaction:

If the effect of D on Y varies with mediator level M (i.e. there is an interaction), then $\text{NDE}(d, d^*)$ averages direct effects across the distribution of $M_i(0)$. This average need not coincide with the direct effect at any particular mediator level, obscuring important heterogeneity.

Exposure-induced confounding:

When treatment D alters a variable L that also affects both M and Y, L becomes an intermediate confounder

By fixing M at a single level m, $\text{CDE}(d, d^*, m)$ measures the effect of D on Y at that mediator value, revealing how the direct effect varies with m. Since the mediator is intervened, exposure-induced confounders L no longer bias the comparison. The CDE can be identified under weaker assumptions (e.g. standard no-unmeasured-confounding for $D \to Y$ given C, and for $D \to M$).

Application to JOBS II:

Suppose the workshop's impact on employment interacts with self-efficacy level (e.g. greater direct benefit when M is low). The NDE would average over each person's $M_i(0)$ and mask this pattern. The CDE at $m = 2, 3, 4$ shows the workshop's direct effect when self-efficacy is held at those fixed levels, clarifying how direct benefit depends on M.

If the workshop also changes an intermediate belief or support network L that confounds the $M \to Y$ link, the NDE is not identifiable without cross-world assumptions, whereas the CDE remains well-defined by construction.

When CDE and NDE Coincide or Differ

Proof. By definition,

$$\text{NDE}(d, d^*) = E\big[Y\big(d,\, M(0)\big)\big] \;-\; E\big[Y\big(d^*,\, M(0)\big)\big].$$

Partitioning on the possible values m of $M(0)$ gives

$$\begin{aligned}\mathrm{NDE}(d, d^*) &= \sum_m \Big\{ E[Y(d,m)] - E[Y(d^*,m)] \Big\} \Pr\{M(0) = m\} \\ &= \sum_m \mathrm{CDE}(d, d^*, m) \Pr\{M(0) = m\}.\end{aligned}$$

Hence

$$\mathrm{NDE}(d, d^*) = E_{M(0)}\big[\mathrm{CDE}(d, d^*, M(0))\big].$$

Equality conditions.

No $D-M$ interaction: If $\mathrm{CDE}(d, d^*, m)$ is constant in m, say equal to c for all m, then

$$\mathrm{NDE}(d, d^*) = \sum_m c \Pr\{M(0) = m\} = c = \mathrm{CDE}(d, d^*, m).$$

In practice this holds when the direct effect of D on Y does not vary with the mediator level.

When they differ

Unless the above conditions holds, $\mathrm{CDE}(d, d^*, m)$ varies with m, so the weighted average $\mathrm{NDE}(d, d^*)$ will not equal $\mathrm{CDE}(d, d^*, m)$ at any single m. In particular, under exposure-mediator interaction or non-degenerate $M(0)$, the direct effect at a fixed m differs from the average direct effect over the distribution of $M(0)$. □

Corrected Identification of ATT

Original derivation The natural direct effect differs from the CDE in how the mediator is handled:

- **CDE:** Fixes the mediator to a value m for everyone, and compares

$$\text{CDE}(d, d^*, m) = E[\,Y(d, m)\,] \;-\; E[\,Y(d^*, m)\,].$$

- **NDE:** Fixes the mediator to each individual's own counterfactual value under control, $M_i(0)$, and compares

$$\text{NDE}(d, d^*) = E\big[\,Y\big(d,\, M(0)\big)\big] \;-\; E\big[\,Y\big(d^*,\, M(0)\big)\big].$$

The Control direct effect fixed mediator's value, then calculate the direct effect of D on Y, while the natural direct effect (NDE) averages the $D \to Y$ effects over the distribution of $M(0)$, it may mask how the direct effect actually varies with different mediator (here is self-efficacy) levels, and it generally requires stronger "cross-world" assumptions to be identified when there are treatment–mediator interactions.

Corrected Derivation For each observation, define

$$M(d) \;=\; \text{the value of the mediator it will naturally has if we set } D = d.$$

In particular,

$$M(0) \;=\; \text{the mediator level each individual would have under control } (D = 0)$$

Controlled Direct Effect (CDE) Fix a constant mediator value m for everyone. Then

$$\text{CDE}(d, d^*, m) \;=\; E\big[Y(d, m)\big] \;-\; E\big[Y(d^*, m)\big],$$

where $Y(d, m)$ is the potential outcome if D is set to d and M is held at m.

Natural Direct Effect (NDE) Hold each individual's mediator at the counterfactual value they would naturally have under baseline d^*. For $(d, d^*) = (1, 0)$,

$$\text{NDE}(1, 0) \;=\; E\big[Y\big(1,\, M(0)\big)\big] \;-\; E\big[Y\big(0,\, M(0)\big)\big].$$

Here:

- $Y\big(1, M(0)\big)$: outcome if $D = 1$ but M is forced to its value under $D = 0$.
- $Y\big(0, M(0)\big)$: outcome if $D = 0$ and M is its natural value under $D = 0$.

Thus NDE compares the expected change of outcome if changing the exposure level from d^* to d and set the mediator at the level it will naturally have under the exposure level d^*
The difference between NDE and CDE is that the NDE measures the direct effect of D on Y when the mediator takes on value it would naturally have been for each individual under a particular level of the exposure, here is d^*, and this value may differ across individuals. With the CDE, the mediator is set at the same value m for every individual regardless of the exposure.
If the effect of D on Y varies with mediator level M (i.e. there is an interaction), then $\text{NDE}(d, d^*)$ averages direct effects across the distribution of $M(0)$. This average need not coincide with the direct effect at any particular mediator level, obscuring important heterogeneity.

TA's Comment: I'm looking for the interpretation of M(d*) or M(0)

My Correction: Added the explanation of $M(d^*)$, and clearly state that the NDE measures the effect when the value of mediator is set to value it will naturally have under the exposure level d^*.

Question 2

Assumptions for CDE identification

To identify

$$\text{CDE}(d, d^*, m) = E[Y(d, m)] - E[Y(d^*, m)]$$

the following assumptions must hold:

CE.1 No unobserved D–Y confounding

$$Y(d, m) \;\perp\!\!\!\perp\; D \,\big|\, C$$

Interpretation: Given baseline covariates C, treatment assignment D must be statistically independent of the joint potential outcomes $Y(d, m)$. In other words, there are no unobserved factors that confound the exposure-outcome relationship.

In our example, let $D \in \{0, 1\}$ is the randomized indicator of assignment to the job-training workshop (1) versus the booklet (0). $Y \in \{0, 1\}$ is whether a participant was working at least 20 hours/week after the intervention. M is the 5-point self-efficacy index (the mediator). C is the set of baseline covariates (education, income, race, age, gender, financial strain). CE.1 requires that, once we condition on the pre-treatment variables C, there are no remaining unmeasured factors that influence both workshop assignment D and the potential re-employment outcome $Y(d, m)$.

CE.2 No unobserved M–Y confounding

$$Y(d, m) \;\perp\!\!\!\perp\; M \,\big|\, C,\, D,\, L$$

Interpretation: After adjusting for C, the treatment D, and any exposure-induced confounders L, the mediator M is independent of the potential outcomes $Y(d, m)$. In other words, this assumption requires that there must not be any unobserved factors that confound the mediator-outcome relationship.

In our example, $D \in \{0, 1\}$ is randomized assignment to the job-training workshop (1) or booklet (0), M is the six-item self-efficacy index measured after treatment, $Y \in \{0, 1\}$ is re-employment status (20 hrs/week), C includes the baseline covariates: education, income, race, age, gender, and financial strain.

CE.2 requires that, once we condition on confouders C and exposure D, there are no unmeasured variables that jointly influence both a participant's self-efficacy score M and their potential re-employment outcome $Y(d, m)$.

CE. 3 Positivity

$$P(D = d \mid C = c) > 0, \quad P(M = m \mid C = c,\, D = d,\, L = \ell) > 0 \quad \forall\, d, c, m, \ell$$

Interpretation: This assumption requires that there must be a positive probability of all values for the exposure conditional on the baseline confounders. It also requires that there must be a positive probability of all values for the mediator conditional on the baseline confounders, exposure, and exposure-induced confounders. To be more specific:

- Treatment positivity $P(D = d \mid C = c) > 0$: For every combination of baseline covariates $C = c$, there must be some individuals who receive treatment, here receive the workshop ($D = 1$) and some who don't receive treatment here receive only the booklet ($D = 0$).
- Mediator positivity $P(M = m \mid C = c,\, D = d,\, L = \ell) > 0$: For each fixed covariate $C = c$, treatment$D = d$, and any of exposure-induced confounders $L = \ell$, each possible level m of job-search self-efficacy must occur with positive probability. Here, if M is recorded on a 5-point scale $\{1, 2, 3, 4, 5\}$, then in every stratum defined by (C, D, L) there must exist at least one participant whose self-efficacy score equals m.

CE.4 Consistency

$$Y = Y(D, M)$$

Interpretation: The observed outcome equals the potential outcome under the observed treatment and mediator values. In other words, consistency requires that the potential outcome function $Y(d, m)$ exactly reproduces the observed outcome whenever (d, m) equals the actually received (D_i, M_i).

To be specific:

- No multiple versions of treatment or mediator: There must be a single, well-defined version of the "workshop" intervention and a single well-defined measurement of the self-efficacy index. Different implementations (e.g. workshop in different locations or variations in the booklet) would violate consistency.
- No interference: One participant's outcome depends only on their own (D_i, M_i), not on others' treatment or mediator values.
- Implications for identification: Under consistency, we can replace unobservable potential outcomes $Y(d, m)$ with the observed Y for those individuals whose (D, M) happen to equal (d, m). Which means that we are measuring what we want to measure.

Assessment in the JOBS II Experiment

- **CE.1 (No unobserved D–Y confounding) holds** Since treatment (D) was randomized within strata of C, so for every level of education, income, age, gender, race and financial strain,

$$Y(d, m) \perp\!\!\!\perp D \mid C.$$

- **CE.2 (No unobserved M–Y confounding) not holds in general.** Self-efficacy (M) was not randomized. There may exist unmeasured factors L (e.g. motivation, innate ability, social support) that are affected by the workshop and also influence employment. Without randomizing or fully observing L,

$$Y(d, m) \perp\!\!\!\perp M \mid C, D, L$$

  cannot be guaranteed.

- **CE.3 (Positivity) mostly holds.** By design, the randomization guarantees $P(D = 1 \mid C = c) = P(D = 0 \mid C = c) > 0$ for all c.

$$P(D = 1 \mid C = c) = P(D = 0 \mid C = c) = 0.5 > 0 \quad \forall\, c,$$

  so every covariate pattern c has a positive probability of receiving each treatment.

  The self-efficacy mediator M takes values in $\{1, 2, 3, 4, 5\}$. In the observed data:

$$P(M = m \mid C = c,\, D = d, L = \ell) > 0 \quad \forall\, m \in \{1, \ldots, 5\},\ \forall\, m, c, d, \ell.$$

  That is, for each combination of baseline covariates $C = c$, treatment $D = d$ and exposure-induced confounder L, every self-efficacy level m appears at least once might be true, but some extreme combinations (c, d, ℓ) might have very few or no observations for certain m. Therefore, as long as the self-efficacy index M varies continuously (no perfect floor/ceiling) within each treatment–covariate-exposure-induced confounders cell, mediator positivity holds.

- **CE.4 (Consistency) mostly holds.** This assumption holds if the workshop protocol was standardized (single "version" of D) and the self-efficacy index was measured uniformly via the same six-item questionnaire (single "version" of M). Also, there should be no interference between participants: one person's employment outcome does not depend on others' assignments, and we could say the observed outcome equals the potential outcome under the observed treatment and mediator values. Hence, for nearly all practical purposes, $Y = Y(D, M)$ holds, though measurement error in D, M might violate the consistency assumption in its intended sense.

Question 3

We load `jobs2.dta` from the JOBS II experiment and define

- `treat`: indicator for workshop assignment (1 = treatment, 0 = control),
- `job_seek`: continuous measure of job-search self-efficacy,
- `work1`: indicator for employment at first follow-up (1 = employed, 0 = not employed),
- baseline covariates {`econ_hard`, `sex`, `age`, `nonwhite`, `educ`, `income`},

We fit the age, econhard as continuous variables, sex and nonwhite as binary variables, and educ and income as ordinal variables with each category recoded.

We compute the 80th percentile of `job_seek`, then fit two outcome models (one is linear and one is logit model), with a treatment–mediator interaction and extract the Controlled Direct Effect (CDE) at that percentile via regression-imputation.

R code

```
> m0 <- quantile(jobs_clean$job_seek, .8, na.rm=TRUE)
>
> ##  linear-model CDE
> mod_lin <- lm(
+   work1 ~ treat * job_seek +
+           econ_hard + sex + age + nonwhite +
+           educ + income,
+   data = jobs_clean
+ )
>
> cde_lin <- impcde(
+   data    = jobs_clean,
+   model_y = mod_lin,
+   D       = "treat",
+   M       = "job_seek",
+   d       = 1,
+   dstar   = 0,
+   m       = m0
+ )
>
> print(cde_lin)
[1] 0.01368073
>
>
> ## logit-model CDE
> mod_log <- glm(
+   work1 ~ treat * job_seek +
+           econ_hard + sex + age + nonwhite +
+           educ + income,
+   data   = jobs_clean,
+   family = binomial(link = "logit")
+ )
>
> cde_log <- impcde(
```

```
+   data    = jobs_clean,
+   model_y = mod_log,
+   D       = "treat",
+   M       = "job_seek",
+   d       = 1,
+   dstar   = 0,
+   m       = m0
+ )
>
> print(cde_log)
[1] 0.01181084
```

Results:

$$\widehat{\text{CDE}}_{\text{linear}} = 0.01368, \quad \widehat{\text{CDE}}_{\text{logit}} = 0.01181.$$

Interpretation

These CDE estimates quantify the effect of attending the workshop on employment probability, holding each individual's job-search self-efficacy constant at the 80th percentile. Under the linear model, the estimated CDE is

$$\widehat{\text{CDE}}_{\text{linear}} \;=\; 0.01368,$$

meaning that if everyone's self-efficacy were fixed at that 80% value, being assigned to the workshop increases the probability of re-employment at follow-up by about 1.368 percentage points, comparing to not attending the workshop.

Under the logit model, the corresponding CDE on the probability scale is

$$\widehat{\text{CDE}}_{\text{logit}} \;=\; 0.01181,$$

which shows if everyone's self-efficacy were fixed at that 80% value, being assigned to the workshop increases the probability of re-employment at follow-up by about 1.181 percentage points, comparing to not attending the workshop.

The close agreement of these two estimates suggests that our result is not sensitive to the choice of link function.

Because these are controlled direct effects, they describe the impact of the workshop on employment when self-efficacy is held at a high (80th percentile) level. The fact that these CDEs are relatively small means that for individuals whose self-efficacy is already high, the workshop confers only a small additional benefit via paths other than boosting self-efficacy.

Question 4

We used the following code to generate the CI and p-value:

```
cde_log_boot <- impcde(
  data    = jobs_clean,
  model_y = mod_log,
  D       = "treat",
  M       = "job_seek",
  d       = 1,
  dstar   = 0,
  boot    = TRUE,
  boot_reps = 1000,
```

```
  boot_conf_level = 0.90,
  boot_seed = 42,
  m       = m0
)

tab_cde <- with(cde_log_boot,
  data.frame(
    Estimate    = CDE,
    `Lower (90%)` = ci_CDE[1],
    `Upper (90%)` = ci_CDE[2],
    `p-value`     = pvalue_CDE
  )
)

print(tab_cde}
```

The result:

Table 1: Controlled Direct Effect of Treatment on Employment at the 80th Percentile of Self-Efficacy (Logit Model, 90% CI)

	Estimate	Lower (90%)	Upper (90%)	p-value
CDE $(d = 1, d^* = 0, m = m_{0.8})$	0.01181	-0.0645	0.0860	0.808

Interpretation.

These results quantify the workshop's impact on re-employment probability when all participants' job-search self-efficacy is fixed at a high level (the 80th percentile, $m_{0.8}$).

- **Point estimate.** The estimated CDE is 0.01181, if everyone's self-efficacy were fixed at that 80% value, being assigned to the workshop increases the probability of re-employment at follow-up by about 1.181 percentage points, compared to only receiving the booklets.
- **Confidence interval.** The 90% bootstrap CI ranges from –6.45 percentage points to +8.60 percentage points. This interval is wide and includes zero, meaning that the true controlled direct effect could be slightly negative, exactly zero, or positive, and we therefore cannot rule out the possibility of no CDE of the workshop once self-efficacy is held at a high level.
- **Statistical test.** The two-sided p-value is 0.808, which is above 0.1. We therefore fail to reject the null hypothesis of zero controlled direct effect at the 80th percentile of self-efficacy at $a = 0.1$ level, which means that there is no statistically significant evidence that being assigned to the workshop has impact on re-employment if self-efficacy is fixed at this high level.

Corrected Identification of ATT

Precision of the Estimate The precision of the CDE estimate can be assessed by its associated variability via the standard error or the width of the confidence interval. In our bootstrap procedure, we obtained a 90% CI of $[-0.0645,\ 0.0860]$, whose total length is 0.1505. Because this interval is relatively wide compared to the point estimate (0.01181), it indicates a high degree of uncertainty around the estimate. In other words, the confidence interval's breadth shows that the true CDE could reasonably lie anywhere between about -6.45% and $+8.60\%$. Consequently, even though the point estimate is positive, the large variability means we lack sufficient precision to draw a reliable conclusion about the workshop's controlled direct effect at the 80th percentile of self-efficacy.

TA's Comment: Should also discuss whether the result is precise

My Correction: Added a paragraph discussing the width of the CI, indicating that the point estimate is imprecise, since the 90% CI is wide.

Question 5

Setup and Notation

Let

- D be the proportion of a country's population descended from historically plow-using communities (the exposure).
- $M = \log(\text{per-capita income})$ be the mediator.
- Y be the proportion of seats held by women in parliament (the outcome).
- C be a vector of baseline covariates (domesticated animals, terrain suitability, tropical climate).
- $\mathcal{M}(d \mid C)$ denote a draw from the distribution of M under $D = d$, given C.

Formal Definition of IDE

$$\text{IDE}(d, d^*) \;=\; E\Big[Y\big(d,\, \mathcal{M}(d^* \mid C)\big) \;-\; Y\big(d^*,\, \mathcal{M}(d^* \mid C)\big)\Big].$$

Interpretation

It is the expected change in the outcome if individual exposed to d rather than d^*, while setting the mediator M at a value randomly drawn from its distribution under the exposure d^* given baseline confounders C. It captures an effect of the exposure D on the outcome Y that does not operate through its impact on the population distribution of the mediator M

In our plow-use example, D is the proportion of a country's population descended from plow-using ancestors. Hence $\text{IDE}(d, d^*)$ represents the expected change in the share of parliamentary seats held by women if the ancestral plow-use proportion were increased from d^* to d, while holding log per-capita income fixed at a value randomly drawn from its distribution when the plow-use proportion was d^*, conditional on baseline covariates C. This isolates the interventional direct effect not operating through income.

Interventional Indirect Effect (IIE)

$$\text{IIE}(d, d^*) \;=\; E\Big[Y\big(d,\, \mathcal{M}(d \mid C)\big) \;-\; Y\big(d,\, \mathcal{M}(d^* \mid C)\big)\Big].$$

This is the expected change in Y when individuals had been exposed to d then experienced a level of the mediator randomly drawn from its distribution under exposure d rather than from its distribution under exposure d^*. It captures an effect of the exposure D on the outcome Y that arises from a shift in the population distribution of the mediator caused by a change in the exposure D.

In our plow-use application, let D be the proportion of a country's population descended from historical plow-using communities, M the log of per-capita income, and Y the share of parliamentary seats held by women. Then $IIE(d, d^*)$ represents the expected change in women's representation in parliament due to the change in the log-per capita income population distribution induced by raising the ancestral plow-use proportion from d^* to d, while holding the proportion of plow-use at the level d. In other words, it captures the effect of plow-use proportion on women's representation in parliament that operates through the shift of distribution of log-per capita income.

Question 6

To nonparametrically identify the interventional direct effect and indirect effect, we need the following assumptions:

IE.1 No unmeasured exposure–outcome confounding

$$Y(d, m) \perp D \mid C$$

Interpretation: This assumption requires that the exposure D must be statistically independent of the joint potential outcomes $Y(d, m)$, conditional on the baseline confounders C, this assumption requires that there must not be any unobserved factors that confound the exposure-outcome relationship.

In our example, Here D is the proportion of a country's population with ancestral plow agriculture, Y is the proportion of parliamentary seats held by women in 2000, and C includes baseline geography (terrain suitability, tropical climate) and availability of large domesticated animals. IE.1 requires that once we adjust for those pre-industrial covariates C, there are no other unmeasured factors that jointly influence both a country's ancestral plow-use share and its modern female representation.

IE.2: No unmeasured mediator–outcome confounding

$$Y(d, m) \perp M \mid C, D, L$$

After adjusting for C, the treatment D, and any exposure-induced confounders L, the mediator M is independent of the potential outcomes $Y(d, m)$. In other words, this assumption requires that there must not be any unobserved factors that confound the mediator-outcome relationship.

In our example, M is the log of per-capita income in 2000, Y is the share of parliamentary seats held by women, D is the ancestral proportion of plow agriculture, C are baseline geographic and animal-domestication covariates, and L is the 2000 Polity score (authoritarian vs. democratic governance). IE.2 requires that, conditional on C, the exposure D, and the governance measure L, there are no unobserved factors that jointly influence both a country's income level and its women's representation in parliament.

IE.3 No unmeasured exposure–mediator confounding

$$M(d) \perp D \mid C$$

This assumption requires that the exposure D must be statistically independent of the potential values of the mediator $M(d)$, conditional on the baseline confounders C. Substantively, this assumption requires that there must not be any unobserved factors that confound the exposure-mediator relationship.

In our example, here D is the ancestral proportion of plow agriculture, $M(d)$ is the country's log per-capita income that would obtain under plow-use level d, and C includes geographic suitability, animal domestication and tropical climate indicators. IE.3 requires that, once we condition on these baseline factors C, there are no unmeasured variables that jointly influence both a country's ancestral plow-use share and its potential per capita income levels.

IE.4 Positivity

$$P\big(D = d \mid C = c\big) \;>\; 0 \quad \text{and} \quad P\big(M = m \mid C = c,\, D = d,\, L = \ell\big) \;>\; 0$$

This assumption requires that there must be a positive probability of all values for the exposure conditional on the baseline confounders. It also requires that there must be a positive probability of all values for the mediator conditional on the baseline confounders, exposure, and exposure-induced confounders.

In our example, since both ancestral plow-use D and log per-capita income M are continuous, IE. 4 should be written as:

$$f_{D|C}(d \mid c) > 0 \quad \text{and} \quad f_{M|C,D,L}(m \mid c, d, \ell) > 0$$

In the example, this assumption requires that for every fixed value of the baseline covariates $C = c$, the conditional density of D at any point d is positive, and for every fixed (c, d, ℓ), the conditional density of M at any point m is also positive. This assumption requires that for every combination of baseline covariates $C = c$, all levels of ancestral plow-use d has positive probability of occurring, and likewise, for every triple (c, d, ℓ), each possible value of log per-capita income m must occur with positive probability.

IE.5 Consistency

$$Y \;=\; Y\big(D,\, M\big) \quad \text{and} \quad M \;=\; M(D).$$

This assumption requires that the observed outcome must be consistent with the joint potential outcomes. It also requires that the observed value of the mediator is consistent with its potential values.

In our example, consistency means that the outcome we actually observe in each country—its proportion of women's seats in parliament is exactly the potential outcome under that country's realized values of ancestral plow-use D and log per-capita income M. Likewise, the observed log income is the potential income that would arise under the country's realized plow-use D. In other words, it requires that the variable "ancestral plow-use proportion" D and the variable "log per-capita income" M must each be well-defined quantities, so that specifying $D = d$ and $M = m$ uniquely pins down the potential outcomes $Y(d, m)$. In particular, there cannot be qualitatively different "kinds" of plow-use or income—only one version of each level. Also, for each country, if that country's true ancestral plow-use is $D = d$ and its true log-income is $M = m$, then the observed proportion of women's parliamentary seats must equal the potential outcome $Y(d, m)$, and the observed mediator must equal the potential mediator $M(d)$.

Assumptions Satisfied by Design in Alesina et al. (2013)

IE.1 No unmeasured exposure–outcome confounding

Not met by design. Ancestral plow-use D is not randomized among countries, although we control for a rich set of pre–industrial covariates C (terrain suitability, tropical climate indicator, availability of draft animals), ancestral plow-use D was not randomly assigned across societies. There may remain unobserved historical or cultural processes, such as colonial institutions, legal traditions, missionary activity, or ethnic fractionalization, that simultaneously influenced both the early adoption of plow agriculture and the evolution of gender norms, including women's political representation.

IE.2 No unmeasured mediator–outcome confounding

Not met by design. Although we adjust for C (geography, climate, animal domestication), the exposure D (ancestral plow-use share), and the exposure-induced confounder L (Polity score in 2000), which together block many back-door paths from M (log per-capita income) to Y (women's parliamentary share). However, there may remain unobserved contemporaneous factors, such as gender-targeted policy reforms, that simultaneously influence both per-capita income and female political representation. For instance, a global commodity boom could raise incomes across many countries and also alter women's labor market opportunities and political mobilization, creating a spurious association between M and Y not captured by C, D, or L. Because these modern economic or policy shocks are not randomized or fully observed, IE.2 cannot be assumed to hold strictly by design.

IE 3. No unmeasured exposure–mediator confounding

Not met by design. Although we condition on C (terrain suitability, tropical climate, animal domestication), ancestral plow-use D was not experimentally randomized across countries. Many unobserved historical or institutional factors—such as the timing and nature of state formation and pre–industrial trade could have simultaneously influenced both a society's adoption of the plow and its long-run economic development. Because these historical determinants are neither randomized nor fully captured by our baseline covariates, we cannot assume that D is independent of the potential mediator values $M(d)$ conditional on C.

IE.4 Positivity

Not met by design. The exposure D is the continuous proportion of ancestral plow-use in $[0, 1]$. In any finite sample of countries and for a given stratum $C = c$, we only observe a discrete set of D-values. Hence, one cannot nonparametrically guarantee positive density at every real d. Likewise, the mediator M (log per-capita income) is continuous. Conditional on $(C = c, D = d, L = \ell)$, we again only see a finite number of income values, violating the strict requirement that each possible m has positive probability. Thus, because both D and M are truly continuous, nonparametric identification fails. In practice, one must impose a parametric model (e.g. specify a regression or density model for $D \mid C$ or for $M \mid C, D, L$) to fill in the unobserved support and enable estimation of the interventional effects.

IE.5 Consistency

Mostly met by design but with caution. In principle, each country's observed log-income M and female-seat in parliament Y are consistent with the potential outcomes under that country's realized ancestral plow-use proportion D. But there is little ambiguity about "versions" of the treatment or mediator: we must have a single well-defined measure of ancestral plow-use and a single well-defined economic statistic (per-capita income), if the historical data on plow adoption or modern income suffer from measurement error, or if the concept of "plow agriculture" varied in practice (e.g. different implements or cultivation methods), then the mapping $D \mapsto M$ or $(D, M) \mapsto Y$ may not be one-to-one, and consistency could be violated.

Question 7

We used the following code to get the results:

```
rm(list=ls())
library(dplyr)
library(tidyr)
library(foreign)
source("https://raw.githubusercontent.com/causalMedAnalysis/causalMedR/refs/heads/main/utils.R")
source("https://raw.githubusercontent.com/causalMedAnalysis/causalMedR/refs/heads/main/rwrlite.R")

plow <- read.dta("/Users/yubin/Desktop/Thesis/Causal Mediation Analysis/hw3/plowUse.dta") %>%
  drop_na(
    plow,
    ln_income,
    women_politics,
    agricultural_suitability,
    tropical_climate,
    large_animals,
    rugged,
    polity2_2000
  ) %>%
  mutate(
    std_women_politics = (women_politics - mean(women_politics)) /
                         sd(women_politics)
  )
```

```
D <- "plow"
C <- c("agricultural_suitability",
       "tropical_climate",
       "large_animals",
       "rugged")

# Specify the three model formulas
Lform <- polity2_2000 ~ plow +
          (agricultural_suitability +
           tropical_climate +
           large_animals +
           rugged)

Mform <- ln_income ~ plow +
          (agricultural_suitability +
           tropical_climate +
           large_animals +
           rugged)

Yform <- std_women_politics ~ (ln_income * plow) +
           (agricultural_suitability +
            tropical_climate +
            large_animals +
            rugged +
            polity2_2000)

# Estimate via regression-with-residuals
rwres <- rwrlite(
  data            = plow,
  D               = D,
  C               = C,
  L_formula_list  = list(Lform),
  M_formula       = Mform,
  Y_formula       = Yform
)

results <- tibble::tibble(
  effect   = c("Overall", "IDE", "IIE", "CDE(m=0)"),
  estimate = c(rwres$OE, rwres$IDE, rwres$IIE, rwres$CDE)
)

print(results)
```

Models Being Fit

- **Governance model (L-model):**

$$\text{Polity score}_{2000} \sim \text{plow} + \{\text{agricultural\_suitability, tropical\_climate, large\_animals, rugged}\}.$$

- **Mediator model (M-model):**

$$\ln(\text{income}) \sim \text{plow} + \{\text{agricultural\_suitability, tropical\_climate, large\_animals, rugged}\}.$$

- **Outcome model (Y-model):**

$$\text{std\_women\_politics} \sim (\ln(\text{income}) \times \text{plow}) + \{\text{agricultural\_suitability, tropical\_climate, large\_animals, rugged}\} + \text{polity2\_200}$$

We get the following results:

Table 2: Interventional Effects of Ancestral Plow Use on Standardized Female Parliamentary Representation

Effect	Estimate (SD units)
Overall Effect	-0.0839
Interventional Direct Effect	-0.3393
Interventional Indirect Effect	$+0.2555$
Controlled Direct Effect at $m = 0$	$+0.7268$

Interpretation

These estimates are expressed in standard-deviation units of the outcome (std_women_politics), which was obtained by centering and scaling the observed proportion of parliamentary seats held by women.

Interventional Direct Effect.

In our regression-with-residuals analysis we estimated

$$\widehat{\text{IDE}} \;=\; -0.3393 \quad \text{(in SD units of } Y\text{)}.$$

Therefore, if we set each country's log per-capita income to a value randomly drawn from its distribution under no–plow use ($d^* = 0$) conditional on the baseline covariates C, increasing the ancestral plow-use proportion by one unit (from 0 to 1) is associated with a 0.3393 Standard Deviation decrease in the standardized share of parliamentary seats held by women. This effect isolates the component of plow use's impact on parliamentary seats held by women that does not operate through changes in income.

Interventional Indirect Effect.

In our application, with both variables standardized, we find

$$\widehat{\text{IIE}} \;=\; +0.2555 \text{ SD},$$

which means that if we set every country's ancestral plow-use proportion to $d = 1$, but for one scenario we draw each country's log per-capita income M from the distribution it would have under plow-use proportion $d = 1$, and for the other scenario we draw M from the distribution it would have under plow-use proportion $d^* = 0$. Then IIE is the average difference in women's parliamentary-seat share between these two scenarios. Numerically, one-unit increase of the plow-use share from $d = 0$ to $d = 1$ shifts the distribution of log-income in such a way that women's representation increases by about 0.2555 standard-deviation units solely via the income pathway.

Note that the exposure D is measured as a proportion in the interval $[0, 1]$. Hence a "one-unit increase" corresponds to moving from no-plow ($D = 0$) to full-plow adoption ($D = 1$). In practice, it may be more interpretable to rescale these effects to a 10 percentage-point increase in plow share (i.e. $\Delta D = 0.1$), which would multiply all estimates by 0.1. Moreover, since the outcome was standardized prior to analysis, all estimated effects are expressed in units of the outcome's standard deviation.

Conclusion

Historical plow agriculture appears to have two opposing pathways on modern female political representation: a direct negative effect (consistent with persistent gender-norm transmission) and a positive indirect effect via its influence on raising income levels. The negative direct pathway slightly dominates, producing a small net negative overall association.

Question 8

We used the following code to get the bootstrap CI and pvalue.

```
rwres <- rwrlite(
  data            = plow,
  D               = D,
  C               = C,
  L_formula_list  = list(Lform),
  M_formula       = Mform,
  Y_formula       = Yform,
  boot            = TRUE,
  boot_reps       = 1000,
  boot_conf_level = 0.95,
  boot_seed       = 60637,
  boot_parallel   = TRUE
)

library(tibble)
res_tbl <- tibble(
  effect   = c("Overall", "IDE", "IIE", "CDE(0)"),
  estimate = c(rwres$OE, rwres$IDE, rwres$IIE, rwres$CDE),
  lower95  = c(rwres$ci_OE[1], rwres$ci_IDE[1],
               rwres$ci_IIE[1], rwres$ci_CDE[1]),
  upper95  = c(rwres$ci_OE[2], rwres$ci_IDE[2],
               rwres$ci_IIE[2], rwres$ci_CDE[2]),
  p_value  = c(rwres$pvalue_OE,
               rwres$pvalue_IDE,
               rwres$pvalue_IIE,
               rwres$pvalue_CDE)
) %>%
  mutate(across(-effect, ~ round(.x, 3)))

print(res_tbl)
```

The result:

Table 3: 95% Percentile Bootstrap CIs and p-values for Interventional Effects

Effect	Estimate	Lower 95%	Upper 95%	p-value
Overall	−0.084	−0.489	0.310	0.668
Interventional DE (IDE)	−0.339	−0.732	0.043	0.078
Interventional IE (IIE)	+0.255	+0.035	+0.528	0.024
Controlled DE at $m = 0$	+0.727	−1.414	+2.856	0.500

Interpretation

Interventional Direct Effect (IDE).

Our estimate,

$$\widehat{\text{IDE}} = -0.339 \text{ (SD units)},$$

The 95% percentile bootstrap CI is $(-0.732,\ 0.043)$, and the two-sided p-value is 0.078.

Holding log per-capita income at its counterfactual distribution under no plow use ($d^* = 0$), one unit increase, which means increasing the ancestral plow-use share from 0 to 1 is associated with a -0.339 point change in the standardized women's share. The 95% CI $(-0.732,\ 0.043)$ just includes zero and yields $p = 0.078$, so we fail to reject the null at $\alpha = 0.05$ but note a suggestive negative direct effect. Thus at the $\alpha = 0.05$ level we fail to reject

the null of zero direct effect and should conclude that there is no significant IDE, yet the point estimate suggests a potentially negative impact of plow-use that does not operate through income.

Interventional Indirect Effect (IIE).

Our estimate,

$$\widehat{\text{IIE}} \;=\; +0.255 \text{ (SD units)},$$

with a 95% percentile-bootstrap confidence interval of $(0.035,\ 0.528)$ and a two-sided p-value of 0.024.

Holding the ancestral plow-use share at value $d = 1$, we compare two hypothetical populations whose log-income distributions are drawn under $d = 1$ versus under no plow use ($d^* = 0$). The IIE quantifies the change in the standardized women's parliamentary share attributable solely to that shift in the income distribution. Here, one-unit increase, or increasing ancestral plow-use from 0 to 1 induces a shift in log per-capita income that, on average, raises women's representation by 0.255 standard-deviation units. Because the 95% CI excludes zero and $p = 0.024 < 0.05$, we reject the null of no interventional indirect effect at the 5% level, concluding that a substantial component of plow-use's impact on female representation operates through its effect on income.

Question 9

Conclusions

Our regression-with-residuals analysis of the Alesina et al. (2013) data yields the following findings:

Overall effect: A one–unit increase in ancestral plow-use share (from 0 to 1) is associated with a -0.084-SD change in standardized female parliamentary representation, but this net effect is not statistically siginificant ($p = 0.668$).

Interventional direct effect (IDE): Holding log per-capita income at the level random draw from its distribution under no plow use $d^* = 0$, the IDE is -0.339 SD (95% CI $-0.732,\ 0.043$), $p = 0.078$. The p-value is above 0.05, so we fail to reject the hypothesis of none IDE, however, the point estimate suggests a potentially meaningful negative impact of plow use on women's representation that does not operate through income.

Interventional indirect effect (IIE): Changing plow-use from 0 to 1 shifts the income distribution in a way that raises women's representation by $+0.255$ SD (95% CI $0.035,\ 0.528$), $p = 0.024$. This pathway via higher income is statistically significant at the 5% level.

Taken together, these results imply two opposing mechanisms of historical plow use on modern female political representation. On the one hand, plow-use increased economic development (income), which in turn supported greater women's representation (positive and significant IIE). On the other hand, plow-use appears to have left a negative cultural or institutional legacy that depresses women's representation independent of income (negative and not significant IDE). The overall effect is small and statistically null. Thus, we may conclude that the plow-use does increase income, then promote women's representation, however, there remains a cultural or institutional factors offsetting much of those gains, making a negative but not significant IDE, and small and statistically null overall effect.

Thus, we fail to conclude that ancestral plow use has any overall effect on contemporary women's political representation, while we can conclude there is a significant positive effect through income of plow-use to women's representations, which can be interpreted as prow-use makes agriculture more efficient, then wealth accumulation increases the gender equality, represented by higher seats shared by women in parliament. But it is offset by a negative direct effect, which can be interpreted as plow cultivation makes men had a comparative advantage in farming, leading to a division of labor along gender lines in societies that practiced plow agriculture, with long-term consequences for gender inequality, but this direct effect is not operating through increasing income, and is non significant, leaving the overall impact indistinguishable from zero.

Threats to Validity

Several potential limitations could undermine our causal conclusions:

Unmeasured confounding: Although we adjust for baseline geography and animal-domestication covariates (C) and governance (L), there may remain omitted historical or cultural factors influencing both plow-use, income, and women's representation.

Model misspecification: Our linear and interaction terms may not capture nonlinearities or higher-order dependencies in the D–M–Y system.

Measurement error: The mediator ($\ln\_income$) and outcome (women's seat share) may be measured with error, biasing both direct and indirect estimates.

Consistency: We assume each country's potential outcomes under a given plow-use level and income level are well-defined (no multiple versions of the treatment or mediator). Historical data heterogeneity may violate these assumptions.