ECI: Week 5. Observational Studies

Jaewon ("Jay-one") Yoo

National Tsing Hua University, ISS5096

Contents

1 Identification in Observational Studies **2**

1.1 From experiments to observational studies . . . 2
1.2 The selection problem . . . 3
1.3 What is identification? . . . 3
1.4 Identification vs. estimation . . . 3
1.5 Confounding . . . 4

2 Selection on Observables **6**

2.1 The key assumptions . . . 6
2.2 Identification of the ATE . . . 6
2.3 Regression estimation . . . 6
2.4 Regression specifications and coefficients . . . 8
2.5 Variance estimation with bootstrap . . . 9
2.6 Nonlinear relationships . . . 10

3 Sensitivity Analysis **14**

3.1 Motivation . . . 14
3.2 Omitted variable bias in terms of partial R^2 . . . 14
3.3 Applied example: the Darfur study . . . 14
3.4 Connection to mediation analysis . . . 15

4 Partial Identification **19**

4.1 The credibility–informativeness tradeoff . . . 19
4.2 No-assumption bounds . . . 19
4.3 Narrowing bounds with domain knowledge . . . 20
4.4 Numerical example . . . 21

4.5 Why MTS only tightens the upper bound 21
4.6 Confidence regions for bounds 21

5 Summary **22**

---

1 Identification in Observational Studies

1.1 From experiments to observational studies

In the previous weeks, we studied experiments where the researcher controls the treatment assignment. In a **randomized experiment**, two properties hold: (1) **positivity**, where every unit has a nonzero probability of being assigned to treatment or control, $0 < \mathbb{P}[D_i = 1] < 1$; and (2) **unconfoundedness**, where treatment assignment does not depend on any potential outcomes, $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$. These two properties made identification of the average treatment effect straightforward.

Experiments are powerful, but not every causal question can be addressed with one. As researchers, we are often interested in studying treatments that arise naturally in the world, and we are simply observing them. Consider, for example, studying the health effects of marijuana use following its recent legalization in parts of the United States. What would an experiment look like here? It would mean randomly assigning individuals to consume a potentially harmful substance (with likely noncompliance), while forcing the control group to abstain. This raises serious ethical concerns. Many causal questions of this kind carry important policy or managerial implications, yet running a controlled experiment is either impractical or unethical.

In these situations, we have to work with data where treatment was not assigned by the researcher. This is an **observational study**. In observational studies, units select into treatment (or are selected into treatment) through processes that may be related to their potential outcomes. This creates the central challenge of causal inference: **confounding**.

The importance of observational studies in economics was recognized by the 2021 Nobel Prize in Economics, awarded to David Card, Joshua Angrist, and Guido Imbens for their contributions to establishing the *design-based approach*. As the Nobel Committee report (p. 2) put it:

> "Taken together, therefore, the Laureates' contributions have played a central role in establishing the so-called design-based approach in economics. This approach—aimed at emulating a randomized experiment to answer a causal question using observational data—has transformed applied work and improved researchers' ability to answer causal questions of great importance for economic and social policy using observational data."

1.2 The selection problem

What can we learn about the average treatment effect (ATE) from observational data? Consider the naive difference in means:

$$\begin{aligned}\mathbb{E}[Y_i|D_i=1]-\mathbb{E}[Y_i|D_i=0] &= \mathbb{E}[Y_i(1)|D_i=1]-\mathbb{E}[Y_i(0)|D_i=0] \\ &= \underbrace{\mathbb{E}[Y_i(1)-Y_i(0)|D_i=1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i=1]-\mathbb{E}[Y_i(0)|D_i=0]}_{\text{selection bias}}\end{aligned}$$

Without unconfoundedness, the naive difference in means equals the ATT *plus* selection bias. **Selection bias** measures how different the treated and control groups are in terms of their potential outcome under control. If the groups differ systematically, for instance, more motivated individuals are both more likely to seek job training *and* more likely to find employment regardless, then the naive comparison confounds the treatment effect with pre-existing differences.

As a concrete example, consider the effect of comment sections on support for online influencers. Suppose the naive estimate suggests that influencers do worse without comment sections than with them. This could reflect a negative ATT (disabling comments actually hurts), *or* it could reflect a positive ATT masked by large negative selection bias—influencers who disable comments may be systematically worse than those who keep them, even if they posted the same content. Without further assumptions, we simply cannot tell these apart.

With an unbounded outcome Y_i, we cannot even bound the ATT, because selection bias could in principle be anywhere from $-\infty$ to $+\infty$. We say the ATT (and the ATE) are **unidentified** without further assumptions.

1.3 What is identification?

Identification is the bridge between counterfactual quantities and observable data. There are two distributions at play:

- The **counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$; this is where causal quantities like the ATE live.
- The **observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$; this is what we can learn from data.

A quantity ψ is **identified** if it can be written as a function of the observational distribution $\mathbb{P}$. In other words: would we know this quantity if we had access to unlimited data? This is a question about the *logic* of the research design, not about sample size or estimation uncertainty.

Connecting counterfactuals to observables requires **assumptions**. The question "What is your identification strategy?" is really asking: what assumptions allow you to claim that what you estimated is a causal effect? Research designs (experiments, regression discontinuity, etc.) can help justify these assumptions, or you will need to justify them through argument.

1.4 Identification vs. estimation

It is important to keep identification and estimation separate:

- **Identification** tells us *what* to estimate: it connects the causal parameter to some function of $\mathbb{P}$.
- **Estimation** tells us *how* to estimate that function of $\mathbb{P}$ from a finite sample.

For example, once we identify the ATE as $\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$, the causal inference part is done. What remains is purely statistical: how do we estimate these conditional expectations from data? Without identification, the statistical properties of an estimator are irrelevant; you are estimating the wrong thing precisely.

1.5 Confounding

Confounding occurs when treatment and potential outcomes are not independent, that is, when there are common causes of both Y_i and D_i. This is the central concern in observational studies. Examples are pervasive in the social sciences:

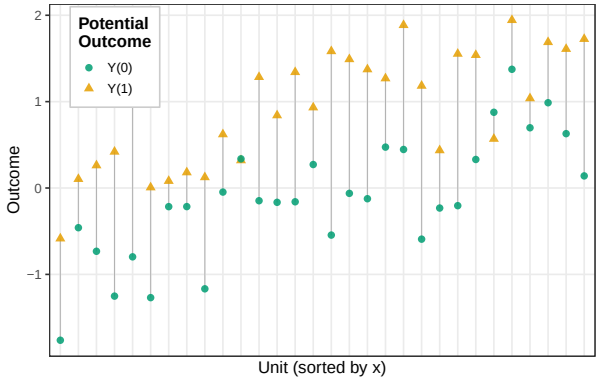
- Effect of job training on employment (confounder: motivation).
- Effect of college GPA on salary (confounder: intelligence).
- Effect of income on voting (confounder: age).
- Effect of corporate giants on economic development (confounder: previous economic development).

Confounding leads to incomplete identification of the ATE, which in turn leads to biased estimators. The rest of this course covers different strategies for dealing with confounding.

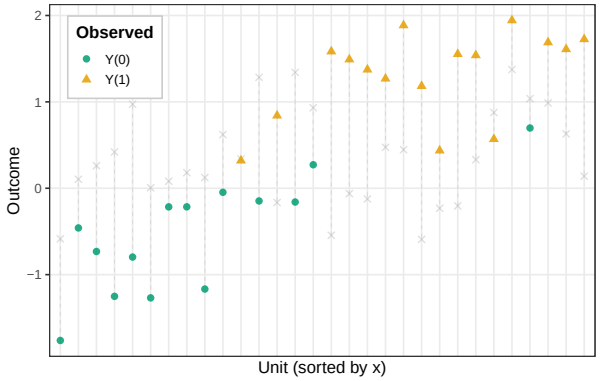
Visualizing confounding

The following sequence of figures, based on simulated data (true ATE = 1), illustrates why confounding makes causal inference difficult.

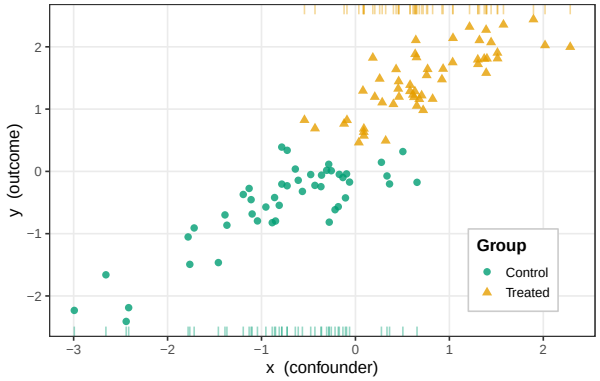
First, consider the ideal scenario: the *science table*, where we observe both potential outcomes for every unit. Each line segment represents the unit-level treatment effect $Y_i(1) - Y_i(0)$:



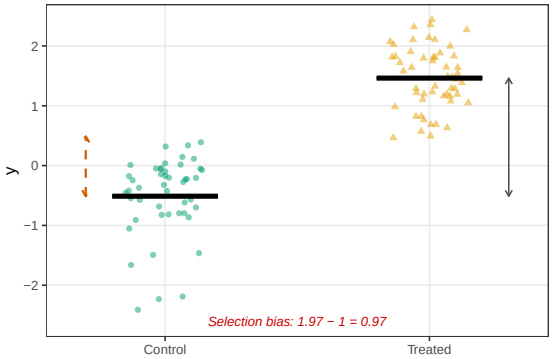
In reality, we face the *fundamental problem of causal inference*: we observe only one potential outcome per unit. The missing counterfactuals (grey marks) make individual treatment effects uncomputable:



When confounding is present, the covariate X_i affects *both* treatment assignment and the outcome. Units with higher X_i are more likely to be treated (notice the different distributions of treated and control units along the x-axis):



As a result, the simple mean comparison $\overline{Y}_1 - \overline{Y}_0$ does **not** equal the ATE. The treated group has higher X_i on average, so they would have had higher outcomes even without treatment:



2 Selection on Observables

2.1 The key assumptions

The most common set of assumptions for observational studies requires two conditions:

1. **No unmeasured confounding** (also called conditional unconfoundedness, weak ignorability, selection on observables, conditional exchangeability, or exogeneity):

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

   Conditional on some covariates $\mathbf{X}_i$, treatment assignment is effectively random. This is a claim that all confounders are captured by the observed covariates.

2. **Positivity** (or overlap):

$$0 < \mathbb{P}[D_i = 1 \mid \mathbf{X}_i] < 1$$

   For every value of the covariates, both treatment and control are possible. There must be both treated and untreated units at every level of $\mathbf{X}_i$.

The term "selection on observables" means that selection into treatment depends only on observable variables, i.e., there are no unobserved confounders. It is a strong assumption, and it **can be wrong**. We will discuss how to assess its plausibility later in this lecture.

2.2 Identification of the ATE

Under positivity and no unmeasured confounding, the population ATE is identified:

$$\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}_{\mathbf{X}}\big[\mathbb{E}[Y_i(1)|\mathbf{X}_i] - \mathbb{E}[Y_i(0)|\mathbf{X}_i]\big] \\
&= \mathbb{E}_{\mathbf{X}}\big[\mathbb{E}[Y_i|D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i|D_i = 0, \mathbf{X}_i]\big]
\end{aligned}$$

The key step is the third line: no unmeasured confounding allows us to replace the counterfactual conditional expectations $\mathbb{E}[Y_i(d)|\mathbf{X}_i]$ with observable conditional expectations $\mathbb{E}[Y_i|D_i = d, \mathbf{X}_i]$. After this substitution, the right-hand side involves only observable quantities; the causal inference part is done.

It is useful to define the **conditional expectation functions** (CEFs):

$$\mu_1(\mathbf{x}) = \mathbb{E}[Y_i(1)|\mathbf{X}_i = \mathbf{x}], \qquad \mu_0(\mathbf{x}) = \mathbb{E}[Y_i(0)|\mathbf{X}_i = \mathbf{x}]$$

These describe how the mean potential outcomes vary with the covariates. Under our assumptions, $\mu_1(\mathbf{x}) = \mathbb{E}[Y_i|D_i = 1, \mathbf{X}_i = \mathbf{x}]$ and $\mu_0(\mathbf{x}) = \mathbb{E}[Y_i|D_i = 0, \mathbf{X}_i = \mathbf{x}]$.

2.3 Regression estimation

Once we have identification, estimation proceeds by modeling the CEFs. The **regression/imputation estimator** of the ATE is:

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$$

The procedure is straightforward:

1. Obtain predicted values for all units under treatment ($D_i = 1$).
2. Obtain predicted values for all units under control ($D_i = 0$).
3. Take the average difference between these predicted values.

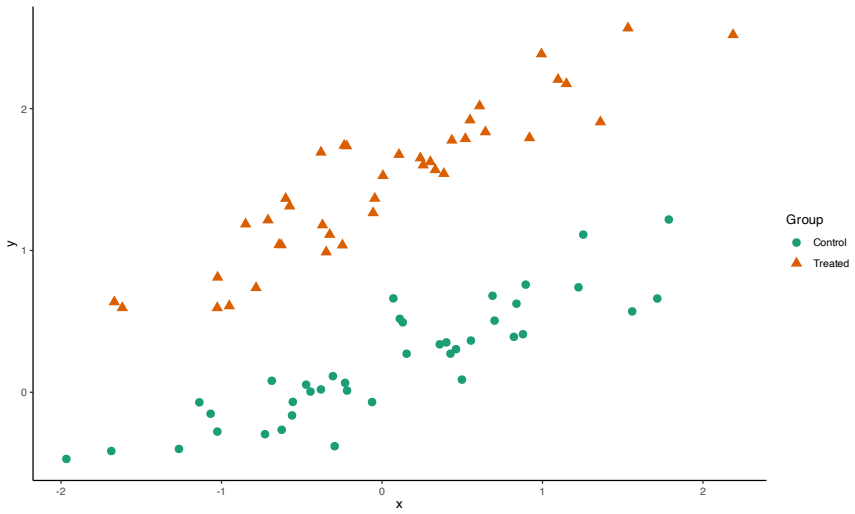
This is sometimes called the **imputation** or **plug-in** estimator, because we "plug in" the estimated CEFs and impute the missing potential outcomes for each unit.

Demonstration: linear imputation estimator

We illustrate this with simulated data. The data can be loaded directly from the following URL:

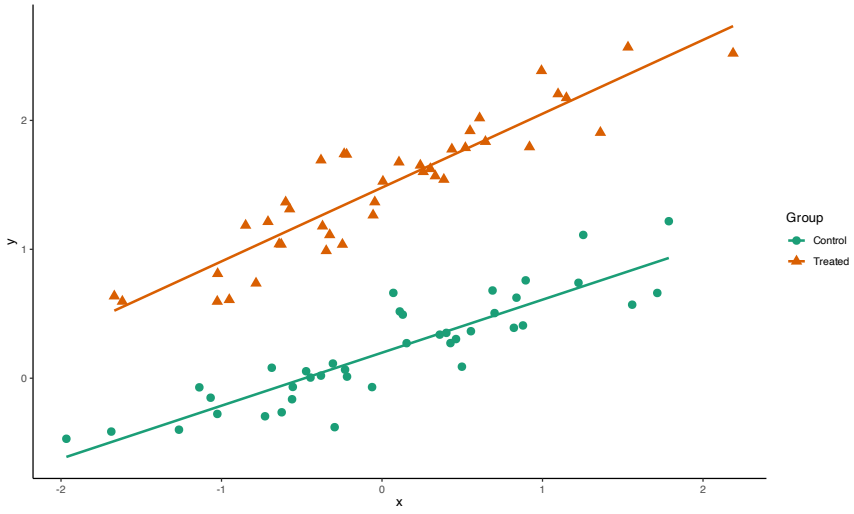
```
toy_data <- read_csv("https://bit.ly/3vOy2Ao")
```

The scatterplot below shows the raw data, with treated units in one color and control units in another. Notice that the treated and control groups occupy overlapping but different regions of the covariate space. This is characteristic of observational data where confounding is present.



We now fit separate linear regressions for the treated and control groups:

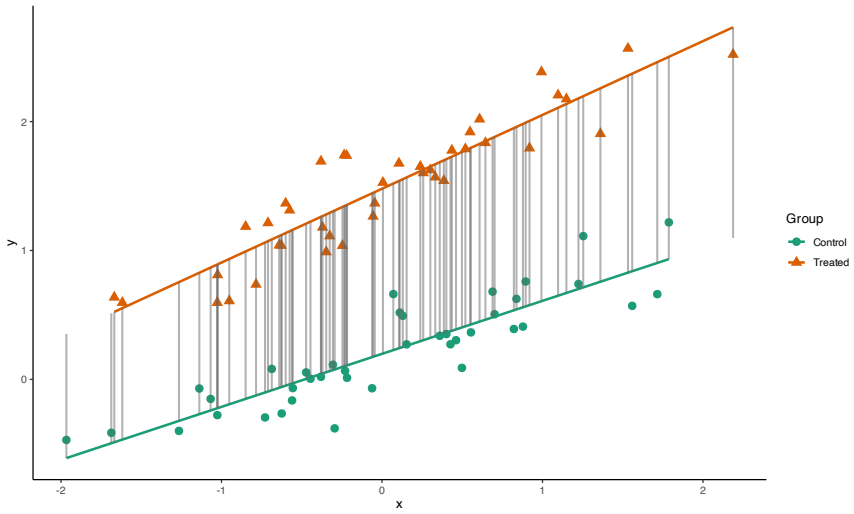
```
lm0 <- lm(y ~ x, data = toy_data, subset = d == 0)
lm1 <- lm(y ~ x, data = toy_data, subset = d == 1)
```



With the fitted models, we can impute the missing potential outcomes for each unit and compute the ATE estimate:

```
mu0.imps <- predict(lm0, toy_data)
mu1.imps <- predict(lm1, toy_data)
cat("Estimate of ATE:", mean(mu1.imps - mu0.imps))
# Estimate of ATE: 1.285176
```

The vertical segments in the figure below show the unit-level imputed treatment effects $\widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$ for each observation. The ATE estimate is the average of these segments.



2.4 Regression specifications and coefficients

Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to an OLS coefficient on D_i, but this depends on the specification:

Uninteracted OLS ($Y \sim D + X$): The CEFs for treated and control are forced to have the same slope; only the intercept differs. This implicitly assumes a **constant treatment effect** across all values of $\mathbf{X}$. Under this model, $\widehat{\tau}_{\text{reg}}$ equals the estimated coefficient on D_i.

Fully interacted OLS ($Y \sim D + \widetilde{\mathbf{X}} + D \cdot \widetilde{\mathbf{X}}$, with centered covariates $\widetilde{\mathbf{X}} = \mathbf{X} - \bar{\mathbf{X}}$): The interaction terms allow each group to have its own intercept and slopes. Consider the model $Y_i = \alpha + \tau D_i + \beta' \widetilde{\mathbf{X}}_i + \gamma'(D_i \widetilde{\mathbf{X}}_i) + \varepsilon_i$:

- When $D = 1$: $\mathbb{E}[Y|\widetilde{\mathbf{X}}] = (\alpha + \tau) + (\beta + \gamma)'\widetilde{\mathbf{X}}$
- When $D = 0$: $\mathbb{E}[Y|\widetilde{\mathbf{X}}] = \alpha + \beta'\widetilde{\mathbf{X}}$

Since the intercepts and slopes differ across groups, this is algebraically equivalent to running separate regressions for treated and control units (i.e., the imputation estimator). Under this specification, $\widehat{\tau}_{\text{reg}}$ also equals the estimated coefficient on D_i, but the variance will typically be smaller than the uninteracted model because we are modeling heterogeneity.

These two specifications make *very different assumptions* about the CEFs. The fully interacted model allows for heterogeneous treatment effects and is generally preferred.

Demonstration: equivalence of fully interacted OLS and imputation

We can verify this equivalence empirically. First, center the covariates and fit the fully interacted model:

```
toy_data$x_tilde <- toy_data$x - mean(toy_data$x)
mod_full <- lm(y ~ d + x_tilde + d * x_tilde, data = toy_data)

cat("Estimate of ATE (Imputation):", mean(mu1.imps - mu0.imps),
    "\nEstimated coefficient on Di from full int.",
    mod_full$coefficients["d"])
# Estimate of ATE (Imputation): 1.285176
# Estimated coefficient on Di from full int. 1.285176
```

The two estimates are identical. Under the fully interacted model with centered covariates, the coefficient on D_i *is* the imputation estimator. This would also hold for the uninteracted model, but the standard errors would be larger (less precision).

2.5 Variance estimation with bootstrap

How do we obtain standard errors for $\widehat{\tau}_{\text{reg}}$? Analytic expressions exist but can be complicated, especially for the imputation estimator. The **bootstrap** provides a simulation-based alternative to analytic standard error formulas:

1. Randomly resample n rows of the data with replacement.
2. Refit the regressions on the bootstrapped data.
3. Calculate $\widehat{\tau}_{\text{reg}}$ in each bootstrap sample.

4. Repeat many times and use the empirical variance of the bootstrap estimates.

The bootstrap approximates the distribution of the estimator by resampling, allowing us to compute standard errors and confidence intervals without deriving the analytic formula.

Demonstration: bootstrap variance estimation

```
set.seed(02138); sims <- 500; tau_hat_draws <- rep(NA, sims)
for (i in 1:sims) {
    # 1. Randomly resample n rows with replacement
    sample_boot <- dplyr::slice_sample(toy_data,
                        n = nrow(toy_data), replace = TRUE)

    # 2. Refit the regressions on the bootstrapped data
    model <- lm(y ~ d + x_tilde + d*x_tilde, data = toy_data)
    dat1 <- sample_boot; dat1$d <- 1
    dat0 <- sample_boot; dat0$d <- 0
    mu1_hat <- predict(model, newdata = dat1)
    mu0_hat <- predict(model, newdata = dat0)

    # 3. Calculate tau_hat in each bootstrap
    tau_hat_draws[i] <- mean(mu1_hat - mu0_hat)
}

# 4. Use empirical variance of the bootstraps
var(tau_hat_draws)
# [1] 0.000254049
```

The bootstrap variance of approximately 0.00025 gives a standard error of $\sqrt{0.00025} \approx 0.016$, which we can use for confidence intervals.

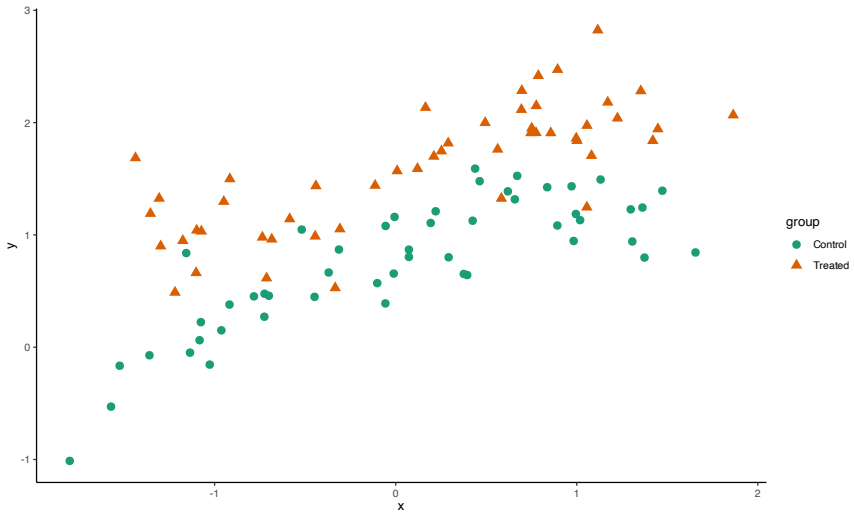
2.6 Nonlinear relationships

When the relationship between Y_i and $\mathbf{X}_i$ is nonlinear, linear models may be misspecified. In such cases, more flexible methods like **generalized additive models** (GAMs) can be used for $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$. The imputation estimator $\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_i \widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$ works identically regardless of whether the underlying models are linear or nonlinear; only the model for the CEFs changes.

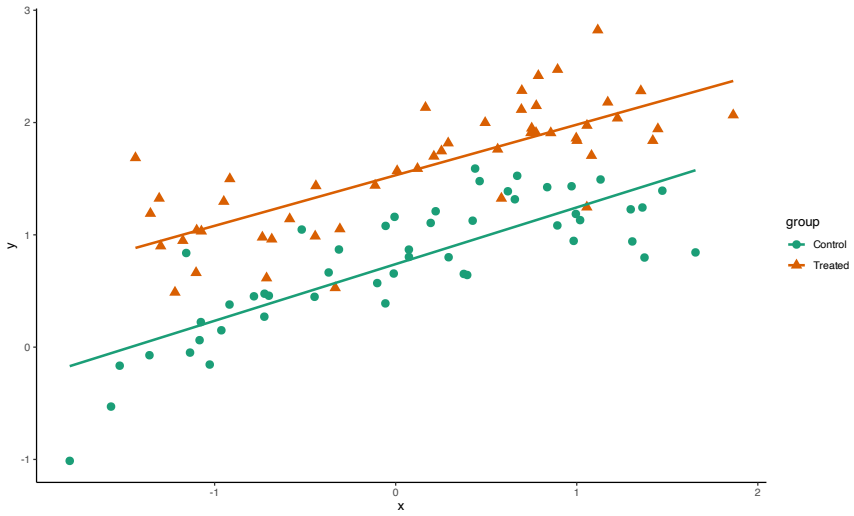
Demonstration: nonlinear CEFs with GAMs

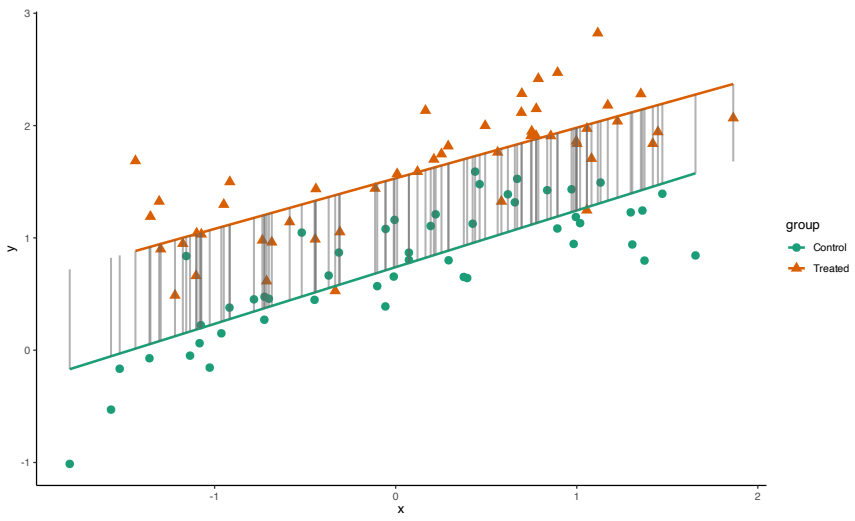
Consider a second simulated dataset where the CEFs are nonlinear:

```
toy_data_02 <- read_csv("https://bit.ly/4cOg0Pl")
```



Fitting linear models here would clearly be a poor choice, as the relationship between Y and X is curved. A linear model imposes straight-line CEFs that miss the curvature, and the resulting imputation segments are visibly off:

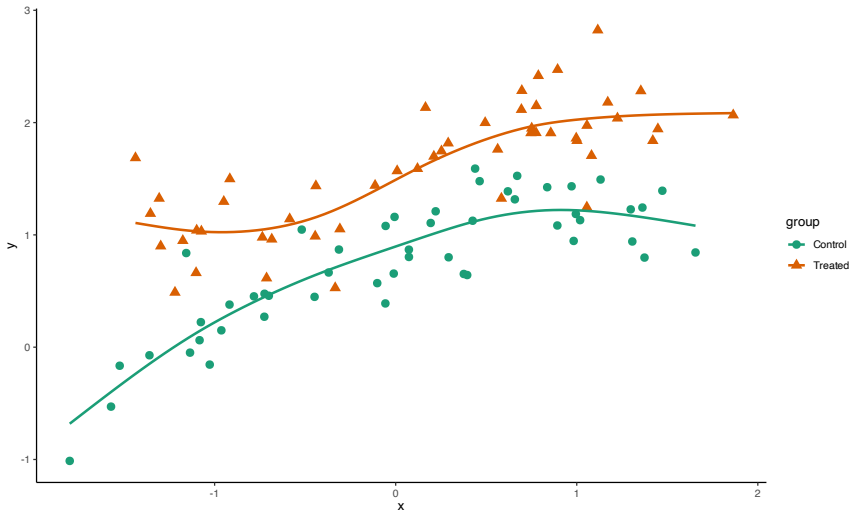




Instead, we can use GAMs from the `mgcv` package, which fit flexible smooth functions:

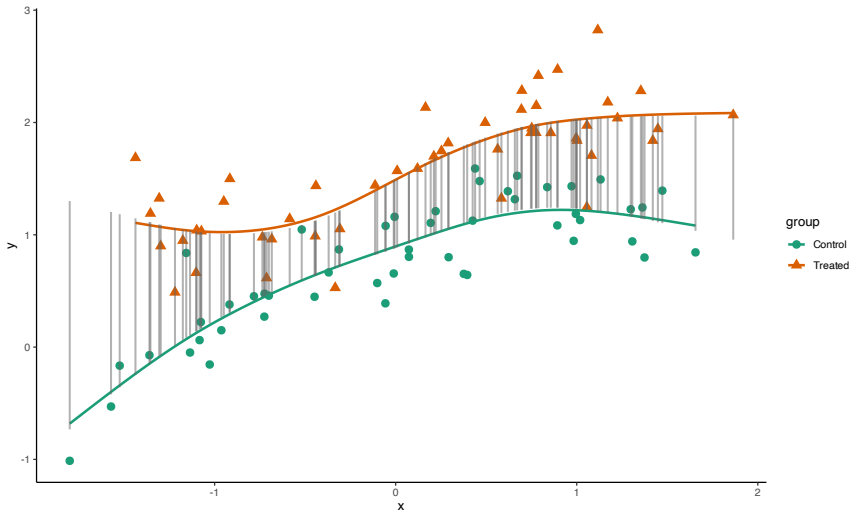
```
library(mgcv)
gam1 <- gam(y ~ s(x), data = toy_data_02, subset = group == "Treated")
gam0 <- gam(y ~ s(x), data = toy_data_02, subset = d == 0)
summary(gam1)

# Parametric coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  1.59527    0.04571    34.9   <2e-16 ***
#
# Approximate significance of smooth terms:
#       edf Ref.df     F p-value
# s(x) 3.73  4.642 19.85  <2e-16 ***
#
# R-sq.(adj) =  0.651   Deviance explained = 67.7%
# GCV = 0.11539  Scale est. = 0.10447  n = 50
```

The GAM fits capture the nonlinearity in the CEFs:



The imputation estimator then uses these flexible fits:

```
cat("Estimate of ATE (GAM):", mean(predict(gam1) - predict(gam0)))
# Estimate of ATE (GAM): 0.8379884
```



The key takeaway: the imputation estimator framework is the same regardless of the modeling choice. What changes is *how* we estimate $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$: we can use OLS, GAMs, random forests, or any other method. The important thing is that the model reasonably approximates the true CEFs.

3 Sensitivity Analysis

3.1 Motivation

The no unmeasured confounding assumption is a claim about *unmeasured* data, and is therefore inherently untestable. We can never verify from the data alone that all confounders have been accounted for. However, we can ask: **how strong would an unobserved confounder need to be to overturn our findings?** This is the idea behind sensitivity analysis.

3.2 Omitted variable bias in terms of partial R^2

Consider the standard regression model and suppose the true model contains an omitted variable U_i:

$$Y_i = \alpha + \tau D_i + \mathbf{X}_i'\beta + \gamma U_i + \varepsilon_i$$

The standard omitted variable bias (OVB) formula gives us $\widehat{\tau} = \tau + \gamma \times \delta$, where δ captures the relationship between U and D after partialing out $\mathbf{X}$.

Cinelli and Hazlett (JRSS-B, 2020) reformulate OVB in terms of **partial** R^2 values, which are easier to reason about:

$$|\text{bias}| = \sqrt{\frac{R^2_{Y\sim U|D,\mathbf{X}} \cdot R^2_{D\sim U|\mathbf{X}}}{1 - R^2_{D\sim U|\mathbf{X}}} \cdot \frac{\mathbb{V}(Y^{\perp\!\!\!\perp \mathbf{X},D})}{\mathbb{V}(D^{\perp\!\!\!\perp \mathbf{X}})}}$$

The intuition is that the bias is proportional to two quantities: how much U predicts the outcome Y (after accounting for D and $\mathbf{X}$), and how much U predicts the treatment D (after accounting for $\mathbf{X}$). A confounder that strongly predicts both treatment and outcome produces large bias; one that predicts only one or neither produces little bias.

The partial R^2 is the incremental predictive value of a variable:

$$R^2_{Y\sim U|D,\mathbf{X}} = \frac{R^2_{Y\sim D+\mathbf{X}+U} - R^2_{Y\sim D+\mathbf{X}}}{1 - R^2_{Y\sim D+\mathbf{X}}} = \frac{\text{additional variance explained by } U}{\text{variance unexplained by } D, \mathbf{X}}$$

Sensitivity analysis then varies the two unknown parameters $R^2_{Y\sim U|D,\mathbf{X}}$ and $R^2_{D\sim U|\mathbf{X}}$ to see how the bias and adjusted estimate change.

Note: this framework applies equally to fully interacted models (cf. Week 4).

3.3 Applied example: the Darfur study

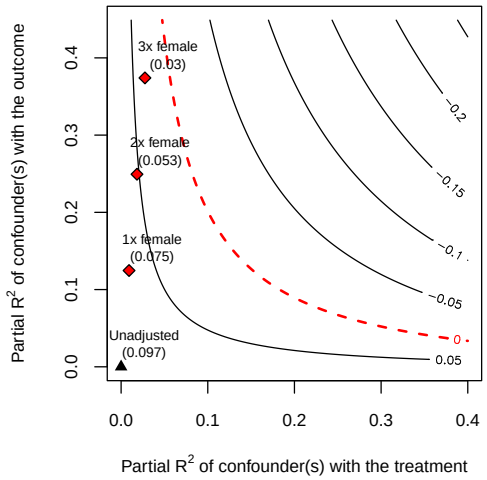
The `sensemakr` R package implements this framework. We demonstrate it using data from the Darfur conflict (Sudan, 2003), a survey of conflict refugees where D indicates whether the individual was directly harmed by violence, and Y measures support for peacemaking. The finding is counterintuitive: those directly harmed are *more* supportive of peace. But could this be driven by an unobserved confounder?

```
library(sensemakr)
```

```r
data("darfur")

# Run the regression model
model <- lm(peacefactor ~ directlyharmed + age + farmer_dar +
              herder_dar + pastvoted + hhsize_darfur + female +
              village, data = darfur)

# Conduct the sensitivity analysis
sensitivity <- sensemakr(model = model,
                         treatment = "directlyharmed",
                         benchmark_covariates = "female",
                         kd = 1:3)
```

Here, `female` is the strongest known confounder in the data and serves as our **benchmark**. The `kd` parameter specifies "knock-on distance," i.e., how many times stronger the hypothetical unobserved confounder is relative to this benchmark. Setting `kd = 1:3` asks: what would happen if the confounder were 1×, 2×, or 3× as strong as `female`?

```r
plot(sensitivity)
```



Reading the contour plot: The x-axis is $R^2_{D\sim U|\mathbf{X}}$ (how strongly U predicts treatment) and the y-axis is $R^2_{Y\sim U|D,\mathbf{X}}$ (how strongly U predicts the outcome). Each contour line shows the adjusted point estimate for a hypothetical confounder with that combination of strengths. The diamond marks the benchmark covariate `female`.

The key finding: even a confounder **three times as strong as `female`** would not drive the estimated effect to zero. Our estimate is robust to plausible levels of unobserved confounding.

3.4 Connection to mediation analysis

This subsection is supplementary reading for students who plan to use mediation analysis in their research. Causal mediation will be covered in more detail later in the course;

the material here provides early context on why the sensitivity analysis tools above are directly relevant to mediation. If this does not apply to your research, you may skip ahead to Section 4. For a deeper treatment, see Imai, Keele, Tingley, and Yamamoto (2011).

Many research projects aim to go beyond "does D cause Y?" and ask "*through what channel* does D affect Y?" This is the domain of **causal mediation analysis**, which decomposes the total causal effect into a *direct effect* ($D \to Y$) and an *indirect effect* operating through a mediator M ($D \to M \to Y$).

Why standard assumptions are not enough

A key insight from Imai, Keele, Tingley, and Yamamoto (2011) is that the standard assumptions used to identify total causal effects—the no unmeasured confounding assumption we discussed above—are **insufficient** for identifying causal mechanisms. Even in a randomized experiment where treatment is randomly assigned:

- The ATE is identified (as usual).
- But the *average causal mediation effect* (ACME) and the *average direct effect* (ADE) are **not** identified.

This is because the mediator M is never randomized—even when D is. The mediator is an intermediate outcome that arises naturally, so the $M \to Y$ relationship is fundamentally observational regardless of the research design.

Sequential ignorability

Identifying direct and indirect effects requires an additional assumption. Imai, Keele, and Yamamoto (2010) formalize this as **sequential ignorability**:

1. $\{Y_i(d', m),\, M_i(d)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$: treatment is as good as random given pre-treatment covariates. This is essentially the no unmeasured confounding assumption from earlier in this lecture.
2. $Y_i(d', m) \perp\!\!\!\perp M_i(t) \mid D_i = d, \mathbf{X}_i$: the mediator is as good as random, given the treatment status and pre-treatment covariates.

The assumption is called "sequential" because two ignorability conditions are imposed in sequence. The first is standard. The second is the problematic one: it requires that there are **no unobserved confounders**—including post-treatment ones—between the mediator M and the outcome Y. The conditioning set must include only pre-treatment covariates $\mathbf{X}_i$; we cannot condition on post-treatment variables (those affected by D) without introducing new biases. This makes the second condition a strong and inherently untestable claim.

What goes wrong without sequential ignorability

Imai et al. (2011) provide a striking numerical example. Suppose the average treatment effect on the mediator is positive (+0.2), and the average effect of the mediator on the outcome is also

positive (+0.2). One might naturally conclude that the indirect effect through M must also be positive. But this is wrong: in their example, the true ACME is *negative* (−0.2). The reason is causal heterogeneity: units for whom the treatment raises the mediator happen to be exactly those for whom the mediator lowers the outcome. Without sequential ignorability, "multiplying the paths" as in traditional structural equation models can give the wrong sign entirely.

Sensitivity analysis for mediation

Because sequential ignorability is strong and untestable, Imai et al. (2011) argue that **sensitivity analysis should be a routine part of any mediation analysis**. Under the linear structural equation framework, sequential ignorability implies that the error terms in the mediator model and the outcome model are uncorrelated. Imai, Keele, and Tingley (2010) propose using the correlation ρ between these residuals as a sensitivity parameter:

- When $\rho = 0$: sequential ignorability holds, and the ACME is identified.
- As ρ deviates from 0: the ACME estimate changes, reflecting the bias from unobserved confounding between M and Y.

By plotting the estimated ACME against ρ, we can determine the value of ρ at which the ACME would be zero or change sign. If this "tipping point" is large, the finding is robust; if small, the mediation result is fragile.

Imai, Keele, and Yamamoto (2010) also offer an R^2-based formulation of the sensitivity analysis, where the sensitivity parameters are the proportions of residual variance in the mediator and outcome models explained by the unobserved confounder—closely analogous to the partial R^2 framework of Cinelli and Hazlett (2020) covered earlier in this section.

The bottom line

The logic of sensitivity analysis for mediation is the same as for observational studies more broadly: we cannot prove an untestable assumption, but we can ask *how badly it would need to be violated* to overturn our conclusions. The difference is that mediation demands a *stronger* assumption (two stages of ignorability rather than one), making sensitivity analysis even more critical. All of these methods—estimation and sensitivity analysis—are implemented in the `mediation` R package (Tingley, Yamamoto, Hirose, Keele, and Imai, 2014). We will cover mediation analysis in detail later in the course.

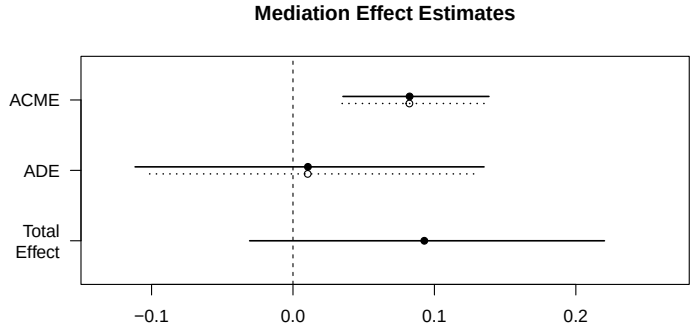
Demonstration: mediation and sensitivity analysis with `mediation`

We illustrate using data from the media framing experiment in Brader, Valentino, and Suhay (2008), included in the `mediation` package. The treatment D is exposure to a negative immigration news story featuring a Hispanic immigrant; the mediator M is the subject's anxiety level (`emo`); the outcome Y is whether the subject sent a message to their member of Congress opposing immigration (`cong_mesg`).

```
library(mediation)
```

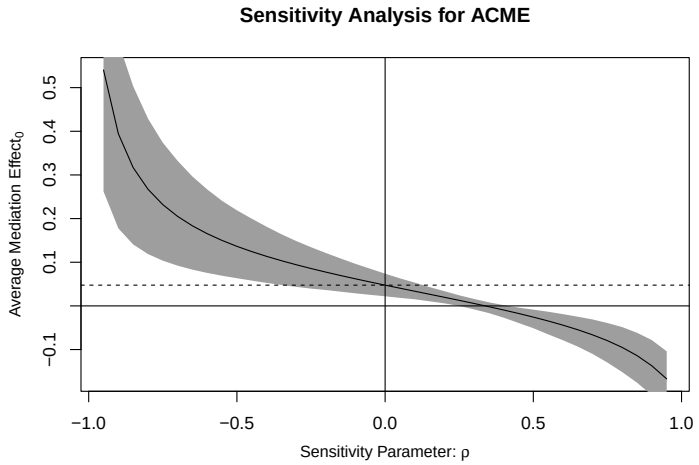
```r
data("framing", package = "mediation")

# Step 1: Mediator model
med.fit <- lm(emo ~ treat + age + educ + gender + income,
              data = framing)

# Step 2: Outcome model
out.fit <- glm(cong_mesg ~ emo + treat + age + educ + gender + income,
               data = framing, family = binomial("probit"))

# Step 3: Estimate mediation effects
set.seed(02138)
med.out <- mediate(med.fit, out.fit, treat = "treat",
                   mediator = "emo", robustci = TRUE, sims = 1000)
summary(med.out)
# ACME (average)      0.0824  [0.035, 0.138]  p < 0.001 ***
# ADE (average)       0.0105  [-0.107, 0.132]  p = 0.876
# Total Effect        0.0929  [-0.031, 0.220]  p = 0.148
```

The ACME (average causal mediation effect) is 0.082 and statistically significant: the negative immigration story increases opposition to immigration *through* heightened anxiety. The direct effect (ADE) is small and insignificant, suggesting most of the total effect operates through the anxiety channel.



But these estimates rely on sequential ignorability. How robust are they? We use the sensitivity analysis:

```r
sens.out <- medsens(med.out, rho.by = 0.05,
                    effect.type = "indirect", sims = 1000)
summary(sens.out)
# Rho at which ACME = 0: 0.35
# R^2_M~ * R^2_Y~ at which ACME = 0: 0.0672
```



Reading the sensitivity plot: The left panel shows the estimated ACME as a function of ρ (the correlation between the mediator and outcome model residuals). At $\rho = 0$, sequential ignorability holds and the ACME matches our original estimate. The ACME crosses zero at $\rho = 0.35$, meaning an unobserved confounder that induces a residual correlation of 0.35 between anxiety and congressional messaging would be enough to explain away the mediation effect entirely. The right panel expresses the same analysis in terms of the $\widetilde{R}^2$ of the mediator and outcome models that the unobserved confounder would need to explain.

Whether $\rho = 0.35$ is "large" or "small" is a substantive judgment. If one believes that unmeasured predispositions (e.g., fear disposition, ideology) could plausibly correlate with both anxiety and political action at this level, the mediation finding may be fragile. This is exactly the kind of reasoning that sensitivity analysis is designed to support.

4 Partial Identification

4.1 The credibility–informativeness tradeoff

Manski's **law of decreasing credibility** states that the credibility of inferences decreases with the strength of the assumptions. Point identification requires strong assumptions (like no unmeasured confounding); weaker assumptions may not pin down a single value for the treatment effect, but they can still **bound** it. This is the idea behind **partial identification**: instead of learning the exact value of τ, we learn that τ must lie within some interval.

4.2 No-assumption bounds

If the outcome Y is bounded, $Y \in [y_L, y_U]$, then logically $\tau \in [y_L - y_U,\ y_U - y_L]$. Can we improve on this using data?

Rewrite the ATE using the law of total expectation, with $p = \mathbb{P}(D_i = 1)$:

$$\tau = \underbrace{\mathbb{E}[Y_i|D_i=1]}_{\text{observed}} \cdot p + \underbrace{\mathbb{E}[Y_i(1)|D_i=0]}_{\text{unobserved}} \cdot (1-p) - \underbrace{\mathbb{E}[Y_i(0)|D_i=1]}_{\text{unobserved}} \cdot p - \underbrace{\mathbb{E}[Y_i|D_i=0]}_{\text{observed}} \cdot (1-p)$$

Each potential outcome mean decomposes into an observed part (from the group that actually received that treatment status) and an unobserved part (the counterfactual for the other group). The observed parts are pinned down by data; the unobserved parts can only be bounded using the logical limits y_L and y_U:

$$\tau \geq \mathbb{E}[Y_i|D_i = 1] \cdot p + y_L(1-p) - y_U \cdot p - \mathbb{E}[Y_i|D_i = 0](1-p)$$
$$\tau \leq \mathbb{E}[Y_i|D_i = 1] \cdot p + y_U(1-p) - y_L \cdot p - \mathbb{E}[Y_i|D_i = 0](1-p)$$

These bounds have width $|y_U - y_L|$, i.e., half of the logical bounds. The data have already narrowed the range, but these bounds will always contain zero. Weak assumptions lead to weak inferences.

4.3 Narrowing bounds with domain knowledge

Domain knowledge can be formalized as assumptions that narrow the bounds further. We consider two types of monotonicity assumptions:

MTR (Monotone Treatment Response): Treatment does not hurt anyone.

$$Y_i(1) \geq Y_i(0) \quad \text{for all } i \qquad \Longrightarrow \qquad \tau \geq 0$$

For example, job training will not make someone *worse* at finding a job. This is an individual-level assumption about the direction of the treatment effect. It pins down the **lower bound** of τ: the effect cannot be negative.

MTS (Monotone Treatment Selection): Positive selection into treatment.

$$\mathbb{E}[Y(d)|D_i = 1] \geq \mathbb{E}[Y(d)|D_i = 0] \quad \text{for } d \in \{0, 1\}$$

People with higher baseline outcomes are more likely to seek treatment. This formalizes the *direction* of selection bias as an assumption. For instance, more motivated individuals are both more likely to enroll in job training *and* more likely to have good outcomes regardless.

MTS constrains the **unobserved counterfactual means**. To see how, note that the upper bound of τ requires finding the largest $\mathbb{E}[Y(1)]$ and smallest $\mathbb{E}[Y(0)]$ consistent with the data and assumptions. Each potential outcome mean has an observed and an unobserved component:

$$\mathbb{E}[Y(1)] = \underbrace{\mathbb{E}[Y_i|D_i = 1]}_{\text{observed}} \cdot p + \underbrace{\mathbb{E}[Y_i(1)|D_i = 0]}_{\text{unobserved}} \cdot (1-p)$$

$$\mathbb{E}[Y(0)] = \underbrace{\mathbb{E}[Y_i(0)|D_i = 1]}_{\text{unobserved}} \cdot p + \underbrace{\mathbb{E}[Y_i|D_i = 0]}_{\text{observed}} \cdot (1-p)$$

MTS says treated units have weakly higher potential outcomes than control units. This means the unobserved $\mathbb{E}[Y_i(1)|D_i = 0]$ cannot exceed the observed $\mathbb{E}[Y_i|D_i = 1]$ (control group cannot beat treated group under treatment), and the unobserved $\mathbb{E}[Y_i(0)|D_i = 1]$ cannot be smaller than the observed $\mathbb{E}[Y_i|D_i = 0]$ (treated group's control potential outcome is at least as high as control group's). The following table summarizes how MTS replaces logical bounds with tighter, **data-driven bounds**:

To find the upper bound...	No assumptions	With MTS
Largest $\mathbb{E}[Y(1)\|D = 0]$ possible	$\leq y_U$	$\leq \mathbb{E}[Y_i\|D_i = 1]$
Smallest $\mathbb{E}[Y(0)\|D = 1]$ possible	$\geq y_L$	$\geq \mathbb{E}[Y_i\|D_i = 0]$

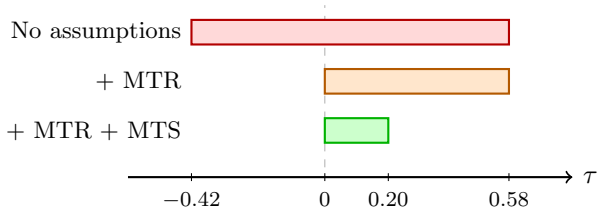
This pins down the **upper bound** of τ.

Crucially, each assumption contributes a different piece of information: MTR pins down the lower bound (from domain knowledge), while MTS pins down the upper bound (from domain knowledge *combined with* data).

4.4 Numerical example

Consider a setting with $Y \in [0, 1]$, $p = \mathbb{P}(D_i = 1) = 0.4$, $\mathbb{E}[Y_i|D_i = 1] = 0.7$, and $\mathbb{E}[Y_i|D_i = 0] = 0.5$.

Assumptions	Bounds for τ	Width
No assumptions (data only)	$[-0.42,\ \ 0.58]$	1.00
+ MTR	$[\ \ 0,\ \ 0.58]$	0.58
+ MTR + MTS	$[\ \ 0,\ \ 0.20]$	0.20



Notice that the final bound $[0, 0.20]$ is tight and informative. The lower bound of 0 comes from MTR (treatment does not hurt), while the upper bound of 0.20 equals the naive difference in means. MTS says selection is positive, so the observed gap is the *most* the true effect could be. Neither assumption alone gives this; domain knowledge and data each contribute a different side.

4.5 Why MTS only tightens the upper bound

A natural question is why MTS does not also tighten the lower bound. The answer lies in the direction of the constraints.

To find the **upper bound**, we seek the largest possible τ, which requires pushing $\mathbb{E}[Y(1)]$ up and $\mathbb{E}[Y(0)]$ down. This means setting the unobserved $\mathbb{E}[Y_i(1)|D_i = 0]$ as large as possible (ceiling) and $\mathbb{E}[Y_i(0)|D_i = 1]$ as small as possible (floor). MTS provides exactly these constraints: it caps $\mathbb{E}[Y_i(1)|D_i = 0]$ at $\mathbb{E}[Y_i|D_i = 1]$ and floors $\mathbb{E}[Y_i(0)|D_i = 1]$ at $\mathbb{E}[Y_i|D_i = 0]$. Both constraints **bind**.

To find the **lower bound**, the directions reverse: we want $\mathbb{E}[Y_i(1)|D_i = 0]$ as *small* as possible and $\mathbb{E}[Y_i(0)|D_i = 1]$ as *large* as possible. But MTS constrains these quantities from the *opposite side*: it gives a ceiling when we need a floor, and a floor when we need a ceiling. So MTS constraints simply **do not bind** for the lower bound. MTR provides the lower bound instead.

4.6 Confidence regions for bounds

The bounds $[\delta_L, \delta_U]$ are population quantities that must be estimated. Since $\widehat{\delta}_L$ and $\widehat{\delta}_U$ are subject to sampling uncertainty, we need confidence intervals. A confidence interval for the **true**

parameter τ takes the form:

$$\left[\widehat{\delta}_L - z_{1-\alpha}\,\widehat{\text{se}}(\widehat{\delta}_L), \quad \widehat{\delta}_U + z_{1-\alpha}\,\widehat{\text{se}}(\widehat{\delta}_U)\right]$$

This uses the one-sided critical value $z_{1-\alpha}$ (not $z_{1-\alpha/2}$) because each bound is a one-sided constraint: the lower bound only needs to guard against τ falling below it, and the upper bound only needs to guard against τ exceeding it. The coverage properties are:

- If $\tau = \delta_L$ or $\tau = \delta_U$ (boundary): coverage $\to 1 - \alpha$.
- If $\delta_L < \tau < \delta_U$ (interior): coverage $\to 1$.

When the true parameter is in the interior of the identification region, both bounds have "slack," and the confidence interval covers with probability approaching 1. The hardest case is when τ sits exactly at one of the bounds.

5 Summary

- Conditional unconfoundedness is an assumption about unmeasured data and is inherently untestable. However, it can be indirectly assessed through sensitivity analysis.
- Sensitivity analysis asks: how strong would an unobserved confounder need to be to overturn the finding? The partial R^2 framework (Cinelli and Hazlett, 2020) provides a principled way to answer this question.
- Without point identification, we can still learn from **bounds** under weaker assumptions. Domain knowledge and data each contribute: assumptions pin down one bound, data tightens the other.
- Up next: directed acyclic graphs (DAGs), a graphical framework for reasoning about which covariates to condition on.