4. Experiments & Linear Regression

Causal Data Science for Business Analytics

Christoph Ihl

Hamburg University of Technology

Monday, 4. May 2026



Randomized Experiments

Randomization is the only design that buys you independence from every covariate — observed and unobserved.

Randomization makes the treatment indicator independent of all covariates

Under a randomized assignment protocol, each unit's treatment value is drawn from a known distribution that does not depend on any of the unit's other characteristics:

$$P(D_i = d) = \frac{1}{k} \quad \text{for } k \text{ equally-sized treatment groups.}$$

Definition: Covariate Balance

We have covariate balance if the distribution of covariates $\mathbf{X}$ is the same across treatment groups:

$$P(\mathbf{X} \mid D = d) \overset{d}{=} P(\mathbf{X} \mid D = d') \quad \text{for all } d, d'.$$

Equivalently:

$$\mathbf{X} \perp\!\!\!\perp D.$$

Randomization implies covariate balance — across **all** covariates, including unobserved ones. This is the central reason RCTs dominate observational designs.

Perspective 1a: Backdoor adjustment collapses to a marginal under randomization

Let $\mathbf{X}$ be a sufficient adjustment set including all (un-)observed confounders. Backdoor adjustment gives:

$$P(Y \mid do(D = d)) = \sum_x P(Y \mid D = d,\, \mathbf{X} = \mathbf{x}) P(\mathbf{X} = \mathbf{x}).$$

Multiply by $\frac{P(D=d|\mathbf{X})}{P(D=d|\mathbf{X})}$ to recover the joint distribution in the numerator:

$$P(Y \mid do(D = d)) = \sum_x \frac{P(Y \mid D = d, \mathbf{X})\, P(D = d \mid \mathbf{X})\, P(\mathbf{X})}{P(D = d \mid \mathbf{X})} = \sum_x \frac{P(Y, D, \mathbf{X})}{P(D = d \mid \mathbf{X})}.$$

Apply randomization $\mathbf{X} \perp\!\!\!\perp D$, so $P(D \mid \mathbf{X}) = P(D)$:

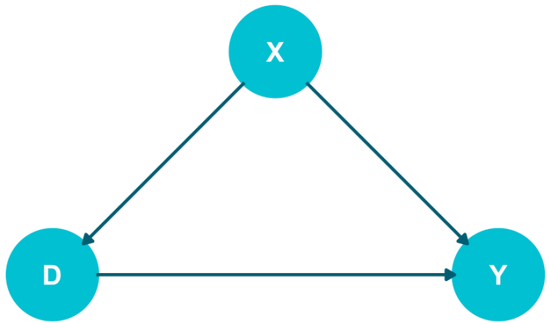
$$P(Y \mid do(D = d)) = \sum_x \frac{P(Y, D, \mathbf{X})}{P(D)} = \sum_x P(Y, \mathbf{X} \mid D = d) = P(Y \mid D = d). \quad \blacksquare$$

Intervention distribution **collapses to the conditional** — association *is* causation under randomization.

Perspective 1b: Randomization blocks every backdoor path — empty adjustment set suffices

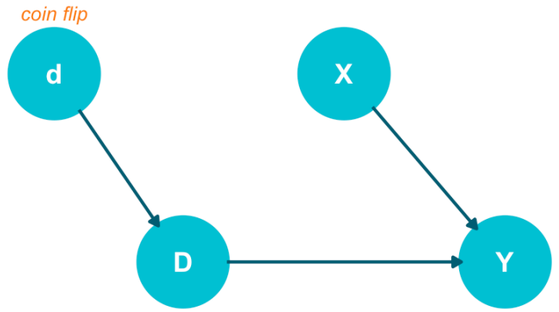
Graphically: randomization replaces every incoming arrow into D with a single edge from a coin-flip node d. **All backdoor paths are mutilated**, so the empty set $\emptyset$ satisfies the backdoor criterion.

Pre-intervention world



X confounds $D \to Y$: backdoor path $D \leftarrow X \to Y$ is open.

Post-intervention world (do-mutilation)



The arrow $X \to D$ is **deleted**: every backdoor path is blocked.

Perspective 2: Decomposing the associational ATE

The **causal ATE** is defined on potential outcomes: $ATE_{cs} = \mathbb{E}[Y_i(1) - Y_i(0)]$.

Let $p = P(D_i = 1)$. Decompose by conditioning on D:

$$ATE_{cs} = p\underbrace{\mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1]}_{\text{ATT}} + (1-p)\underbrace{\mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 0]}_{\text{ATU}}$$

Expanding the conditional expectations and naming the four group means:

$$\begin{aligned} ATE_{cs} &= p\,\mathbb{E}[Y_i(1) \mid D_i = 1] - p\,\mathbb{E}[Y_i(0) \mid D_i = 1] + (1-p)\,\mathbb{E}[Y_i(1) \mid D_i = 0] - (1-p)\,\mathbb{E}[Y_i(0) \mid D_i = 0] \\ &= p\,\mu_{11} - p\,\mu_{01} + (1-p)\,\mu_{10} - (1-p)\,\mu_{00}, \quad \text{where } \mu_{yd} = \mathbb{E}[Y_i(y) \mid D_i = d]. \end{aligned}$$

The **associational ATE** uses observed quantities only:

$$ATE_{as} = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \mu_{11} - \mu_{00}.$$

The second equality uses **consistency**: when $D_i = d$, the observed Y_i equals the potential outcome $Y_i(d)$, so $\mathbb{E}[Y_i \mid D_i = d] = \mathbb{E}[Y_i(d) \mid D_i = d]$.

Perspective 2: Decomposing the associational ATE

Working through the algebra (see session notes for derivation), the associational ATE decomposes as:

$$
\begin{aligned}
ATE_{as} &= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)]}_{ATE_{cs}} \\
&+ \underbrace{\mathbb{E}[Y_i(0) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0]}_{\text{Confounding Bias}} \\
&+ \underbrace{(1 - \mathbb{E}[D_i])\big[\underbrace{\mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 1]}_{\text{ATT}} - \underbrace{(\mathbb{E}[Y_i(1) \mid D_i = 0] - \mathbb{E}[Y_i(0) \mid D_i = 0])}_{\text{ATU}}\big]}_{\text{Heterogeneity Bias}}
\end{aligned}
$$

This is the **gold equation** of the session — we will use it again in Exercise 1 (Christmas / socks). Two distinct biases separate the observed mean difference from the true causal ATE:

- Confounding bias — treated and untreated differ in their *baseline* potential outcome $Y(0)$.
- Heterogeneity bias — *effect* itself differs across treated and untreated subpopulations (ATT $\neq$ ATU).

Perspective 2: Exchangeability zeroes both bias terms

Exchangeability under randomization: $\{Y(1), Y(0)\} \perp\!\!\!\perp D$, which implies:

- Before treatment: $\mathbb{E}[Y_i(0) \mid D_i = 0] = \mathbb{E}[Y_i(0) \mid D_i = 1] = \mathbb{E}[Y_i(0)]$.
- After treatment: $\mathbb{E}[Y_i(1) \mid D_i = 1] = \mathbb{E}[Y_i(1) \mid D_i = 0] = \mathbb{E}[Y_i(1)]$.

$$
\begin{aligned}
ATE_{as} &= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)]}_{ATE_{cs}} \\
&+ \underbrace{\mathbb{E}[Y_i(0) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0]}_{\text{Confounding Bias} \,=\, 0} \\
&+ \underbrace{(1 - \mathbb{E}[D_i])\big[\mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 1] - (\mathbb{E}[Y_i(1) \mid D_i = 0] - \mathbb{E}[Y_i(0) \mid D_i = 0])\big]}_{\text{Heterogeneity Bias} \,=\, 0}
\end{aligned}
$$

$$
\boxed{ATE_{as} \;=\; ATE_{cs} \quad \text{under randomization.}}
$$

Blocked randomization: enforce balance within strata

Simple randomization gives **balance in expectation** — but in any one realisation, treated and control groups can drift apart on important pre-treatment covariates (especially in small samples).

Block (stratified) randomization: partition units into strata $S_1, \ldots, S_K$ on a pre-treatment covariate, then randomize **within each stratum**. Balance on S is now guaranteed *exactly*, not just in expectation.

Within each block k, compute the local mean difference $\hat{\tau}_k = \bar{Y}_{1,k} - \bar{Y}_{0,k}$. Combine using stratum weights $\pi_k = n_k/n$:

$$\hat{\tau}_{\text{strat}} = \sum_{k=1}^{K} \pi_k \, \hat{\tau}_k.$$

Variance is reduced when within-stratum outcome variance is smaller than the unconditional outcome variance — which is exactly the case when S is predictive of Y.

When to block: prior-period outcome (Y_{pre}), demographic strata, geographic regions, or any pre-treatment variable known to be strongly predictive of Y. Avoid blocking on too many strata (sparse cells, lost variance reduction).

Worked example: blocking on prior-period spend in a marketing experiment

Simulated experiment with $n = 1{,}000$ users, pre-period spend X predicts outcome spend Y, and we compare simple vs. blocked random assignment.

R Python

▶ Show code

```
SE_simple SE_blocked
    0.834      0.434
```

The blocked design typically yields a substantially smaller standard error when X is predictive of Y — variance reduction by design.

Effect Identification

From a population identity to an estimator: mean comparison and linear regression are the same thing.

Identification = expressing the unobservable ATE in observable quantities

Identification is the step of expressing the inherently unobservable ATE in terms of observable population quantities — by appeal to randomization or further assumptions that remove confounding and heterogeneity bias.

Under randomization of D, the **mean comparison** identifies the ATE:

$$\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)] \;=\; \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0].$$

The corresponding **sample estimator** is the difference of two sample means:

$$\hat{\tau}_{ATE} \;=\; \frac{1}{\sum_i D_i} \sum_i D_i Y_i \;-\; \frac{1}{\sum_i (1 - D_i)} \sum_i (1 - D_i) Y_i.$$

Job Corps RCT: 9,240 disadvantaged 16–24 year olds, weekly earnings in year 4

Job Corps is a US Department of Labor education program targeting disadvantaged 16–24 year olds. The RCT randomized program access between November 1994 and February 1996; outcome is weekly earnings four years post-assignment.

R Python

▶ Show R code

```
[1] 197.9258
```

▶ Show R code

```
[1] 213.981
```

▶ Show R code

```
[1] 16.05513
```

Sources: Schochet, Burghardt & Glazerman (2001); Schochet, Burghardt & McConnell (2008). Data: 1994–96 US, distributed via the causalweight R package.

CATE is identified the same way under conditional independence

A randomized experiment also delivers **conditional independence** for any pre-treatment covariate X:

$$\{Y(1), Y(0)\} \perp\!\!\!\perp D \mid X.$$

Hence the **Conditional Average Treatment Effect (CATE)** is identified:

$$\tau_{CATE}(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] = \mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x].$$

X must be pre-treatment

The conditioning set X must be measured before treatment assignment – formally, we condition on X, not $X(D)$. Conditioning on a post-treatment variable can both block part of the causal effect and open collider paths through unobservables (back to Session 02's collider lesson – and previewed again in the *Pre- vs. post-treatment covariates* slide).

Estimators of CATEs (causal forests, meta-learners) are the topic of **Session 09: Treatment Effect Heterogeneity**.

From mean comparison to OLS — derivation

Start from the observed-outcome equation for each unit:

$$Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i) = Y_i(0) + (Y_i(1) - Y_i(0))\, D_i.$$

Take the conditional expectation given D_i in the population:

$$\mathbb{E}[Y_i \mid D_i] = \mathbb{E}[Y_i(0) \mid D_i] + \big(\mathbb{E}[Y_i(1) \mid D_i] - \mathbb{E}[Y_i(0) \mid D_i]\big)\, D_i.$$

Apply exchangeability — under randomization, $\{Y(1), Y(0)\} \perp\!\!\!\perp D$, so the conditional expectations do not depend on D:

$$\mathbb{E}[Y_i(0) \mid D_i] = \mathbb{E}[Y_i \mid D_i = 0], \qquad \mathbb{E}[Y_i(1) \mid D_i] = \mathbb{E}[Y_i \mid D_i = 1].$$

Substituting:

$$\mathbb{E}[Y_i \mid D_i] \;=\; \underbrace{\mathbb{E}[Y_i \mid D_i = 0]}_{\beta_0} \;+\; \underbrace{\big(\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]\big)}_{\beta_1} \cdot D_i.$$

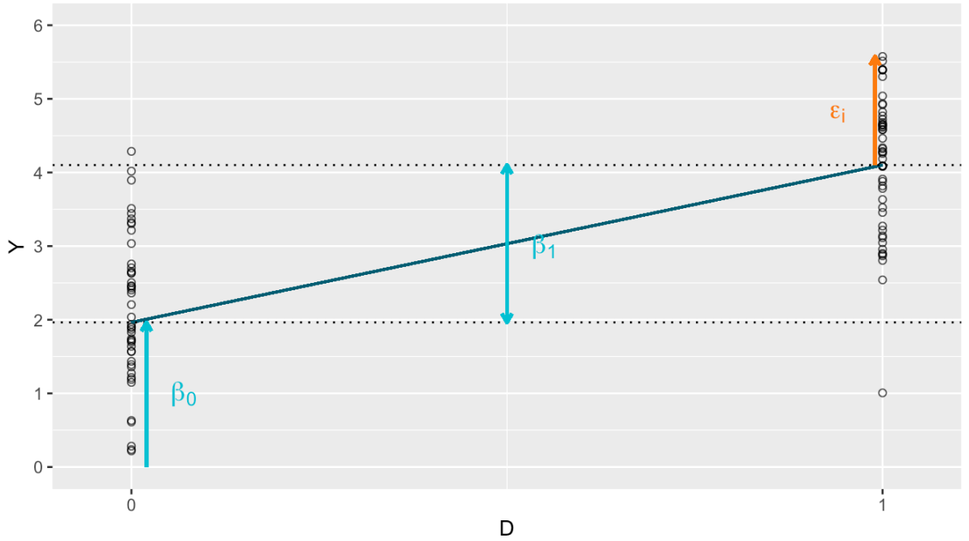
This is the population linear-regression model with $\beta_1 = ATE$.

Residuals visualised: β_0, β_1, ϵ_i on a 2-group scatter

The residual is the gap between the observed outcome and the conditional mean: $\epsilon_i = Y_i - \underbrace{(\beta_0 + \beta_1 D_i)}_{\mathbb{E}[Y_i|D_i]}$.

R Python

▶ Show code



OLS = minimising the sum of squared residuals

The sample estimator is the OLS minimiser:

$$\hat{\beta}_0, \hat{\beta}_1 \;=\; \arg\min_{\beta_0^*,\, \beta_1^*} \sum_{i=1}^{n} \underbrace{(Y_i - \beta_0^* - \beta_1^* D_i)^2}_{\hat{\epsilon}_i^2}.$$

Solving the first-order conditions yields the closed-form estimators:

$$\hat{\beta}_1 \;=\; \frac{\widehat{\mathrm{Cov}}(Y_i, D_i)}{\widehat{\mathrm{Var}}(D_i)},$$

$$\widehat{\mathrm{Cov}}(Y_i, D_i) = \tfrac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \tfrac{1}{n} \sum_j Y_j \right) \left(D_i - \tfrac{1}{n} \sum_j D_j \right),$$

$$\widehat{\mathrm{Var}}(D_i) = \tfrac{1}{n-1} \sum_{i=1}^{n} \left(D_i - \tfrac{1}{n} \sum_j D_j \right)^2, \qquad \hat{\beta}_0 \;=\; \bar{Y} - \hat{\beta}_1 \bar{D}.$$

Statistical Inference

A point estimate is not a finding — uncertainty quantification is what turns a number into a claim.

Under random assignment, OLS estimators are unbiased, consistent, asymptotically normal

Under standard regularity conditions (Wooldridge 2010), the OLS estimators of β_0, β_1 have three desirable properties:

Unbiased

On average across hypothetical samples, equal to the true parameter: $\mathbb{E}[\hat{\beta}_1] = \beta_1, \quad \mathbb{E}[\hat{\beta}_0] = \beta_0$.

Consistent

Converges in probability to the truth as $n \to \infty$: $\hat{\beta}_1 \xrightarrow{p} \beta_1, \quad \hat{\beta}_0 \xrightarrow{p} \beta_0$.

Asymptotically normally distributed

Approximately normal in large samples — the basis for t-tests and confidence intervals:

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\beta_1}), \qquad \sqrt{n}(\hat{\beta}_0 - \beta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\beta_0}).$$

Proofs: Wooldridge (2010), *Econometric Analysis of Cross Section and Panel Data*, Ch. 4. Skipped in lecture.

In an RCT, who actually randomizes?

The coin flip ***— not nature drawing from an infinite population.***

Neyman vs. Fisher inference: large-sample asymptotics vs. exact randomization tests

Both fix the units in front of you and treat the **assignment vector $\mathbf{D}$** as the only randomness — they disagree on what to *do* with it.

Neyman (1923) — large-sample

Asks *"how big is the average effect, and how precisely?"* — tests the weak null $\mathbb{E}[Y(1) - Y(0)] = 0$. Treats the potential outcomes as fixed; randomization is the source of variability.

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}(0,\, V_N).$$

Output: point estimate $\hat{\tau}$ + asymptotic SE + CI. Variance V_N is conservatively estimated under treatment-effect heterogeneity (Neyman's variance bound).

Pros: closed-form CIs, scales to large n, dominates textbooks. Cons: only valid asymptotically; conservative in small samples.

Fisher (1925) — exact

Asks *"could this difference be a coincidence under no effect for anyone?"* — tests the sharp null $H_0 : Y_i(1) = Y_i(0)$ for *every* unit i.

Under H_0, observed outcomes are fixed; the test-statistic distribution comes from re-shuffling treatment labels across all $\binom{n}{n_1}$ permutations.

Output: an exact finite-sample p-value

$$p = \frac{1}{|\mathcal{P}|} \sum_{\pi \in \mathcal{P}} \mathbf{1}\{|\hat{\tau}_\pi| \geq |\hat{\tau}_{\text{obs}}|\}.$$

Pros: no asymptotic approximation; works in tiny samples. Cons: tests sharp null (not average null); compute-intensive.

From here on we ride with Neyman — but Fisher's permutation logic returns next session as the engine behind A/B-test p-values.

Inference goals: standard errors, hypothesis tests, confidence intervals

A point estimate $\hat{\beta}_1$ has zero information about its own precision. We need three primitives:

- **Standard errors** — $se(\hat{\beta}_1)$ measures the sample-to-sample variability of the estimator.
- **Hypothesis tests** — assess whether the data are inconsistent with a specific null value (typically $\beta_1 = 0$).
- **Confidence intervals** — report a range of plausible values for β_1 at a chosen confidence level.

These three are not three separate calculations: they are three views on the same sampling-distribution object. Once you have the SE, you have everything else by simple arithmetic.

Heteroskedasticity-robust standard error of $\hat{\beta}_1$

The asymptotic variance of $\hat{\beta}_1$ in the population (heteroskedasticity-robust / sandwich form):

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\mathbb{E}\left[\epsilon^2 (D_i - \mathbb{E}[D_i])^2\right]}{n \cdot (\mathrm{Var}(D_i))^2}.$$

Plug-in sample analog using residuals $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i$:

$$\widehat{\mathrm{Var}}(\hat{\beta}_1) = \frac{\sum_i \hat{\epsilon}_i^2 (D_i - \bar{D})^2}{\left[\sum_i (D_i - \bar{D})^2\right]^2}, \qquad se(\hat{\beta}_1) = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)}.$$

This is the **HC0** sandwich estimator. R's `sandwich::vcovHC(fit, type = "HC1")` and Python's `cov_type="HC1"` apply a small-sample correction factor of $n/(n-k)$.

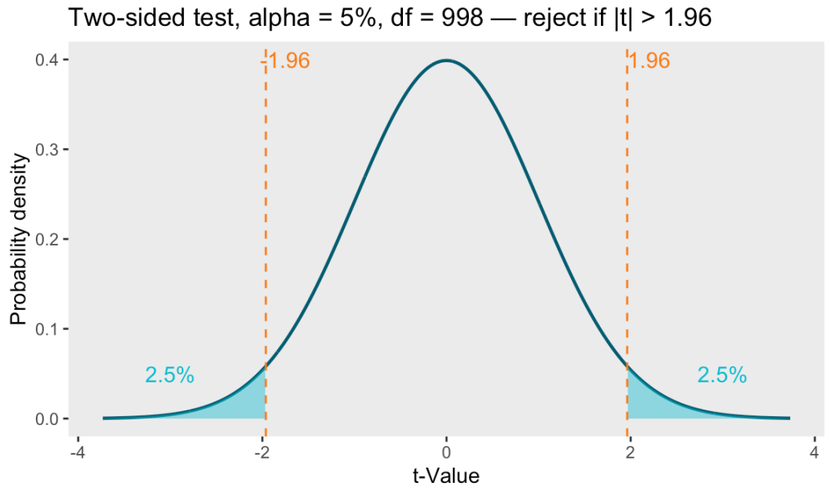
HC vs. classical SE

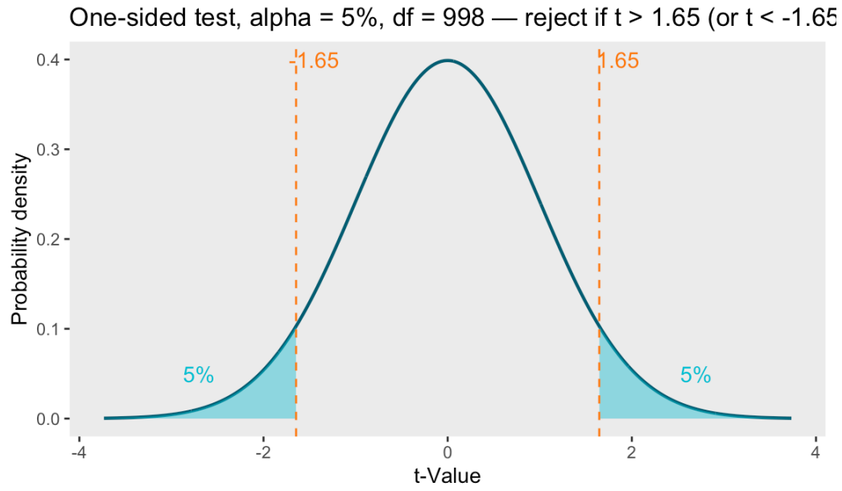
The expression above is the heteroskedasticity-robust (HC0/HC1) sandwich form. The classical "homoskedastic" SE simplifies to $\sigma_\epsilon^2/(n \cdot \mathrm{Var}(D))$ and assumes $\mathrm{Var}(\epsilon \mid D) = \sigma_\epsilon^2$ – a constraint we rarely want to impose. Default to HC.

Reading the t-distribution: 5% two-sided vs. 5% one-sided critical values

The standardised statistic with **known** sd is normal; with the **estimated** se it is t-distributed with $df = n - 2$:

$$z_1 = \frac{\hat{\beta}_1 - \beta_1}{sd(\hat{\beta}_1)} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{vs.} \quad t_1 = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{n-2}.$$





Hypothesis testing — null/alternative + significance level

Step 1. State the null H_0 and alternative H_1:

- **Two-sided (undirected):** H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$.
- **One-sided (directed, positive):** H_0: $\beta_1 \leq 0$ vs. H_1: $\beta_1 > 0$.
- **One-sided (directed, negative):** H_0: $\beta_1 \geq 0$ vs. H_1: $\beta_1 < 0$.

Step 2. Set the significance level α:

- The maximally accepted **type I error** probability — the chance of rejecting H_0 when it is in fact true.
- $\alpha = 0.05$ means at most a 5% rate of false rejections in repeated samples.
- Conventional alternatives: $\alpha = 0.01$ (1%), $\alpha = 0.10$ (10%).

Hypothesis testing — critical value, t-statistic, p-value

Step 3. Compute the **critical value** c:

- Quantile of the t-distribution at α, given H_0.
- Two-sided, $\alpha = 0.05$, $df = 1000 - 2$: $c = t_{\alpha/2=0.025,\, 998} \approx 1.96$.
- One-sided, $\alpha = 0.05$, $df = 998$: $c = t_{\alpha=0.05,\, 998} \approx 1.65$.

Step 4. Compute the **test statistic** and **p-value**:

- Test statistic: $t_1 = \dfrac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$.
- p-value: probability of a test statistic as extreme as t_1 under H_0:
  - Two-sided: $p = P(|A| \geq |t_1|)$ for $A \sim t_{n-2}$.
  - One-sided: $p = P(A \geq t_1)$ or $p = P(A \leq t_1)$.

Decision rules:

- Two-sided: reject H_0 if $|t_1| > c$, equivalently $p < \alpha$.
- One-sided: reject H_0 if $t_1 > c$ (or $t_1 < -c$), equivalently $p < \alpha$.

Confidence intervals: $\hat{\beta}_1 \pm c \cdot se(\hat{\beta}_1)$

A $(1-\alpha)$ **confidence interval** is the range of values β_1^* for which the corresponding null $H_0 : \beta_1 = \beta_1^*$ would *not* be rejected at level α.

Frequentist interpretation: across hypothetical repeated samples, $(1-\alpha)$ of the constructed CIs would cover the true β_1.

Two-sided CI:

$$CI = \hat{\beta}_1 \pm c \cdot se(\hat{\beta}_1).$$

Worked example: $\hat{\beta}_1 = 0.30$, $se = 0.05$, $c = 1.96 \rightarrow CI = 0.30 \pm 1.96 \cdot 0.05 = [0.202,\ 0.398]$.

One-sided CI:

$$CI = \hat{\beta}_1 - c \cdot se(\hat{\beta}_1) \quad \text{(upper open)}.$$

Worked example with $c = 1.65$: $CI = [0.218,\ \infty)$.

Common misinterpretations of α and CIs — and where the intuitive reading actually lives

The frequentist machinery is precise but not intuitive. Students (and reviewers, and journalists) routinely misread it.

What people *want* it to mean

- "$p = 0.03$ means there's a 3% chance the null is true."
- "There's a 95% chance the true β_1 lies in this CI."
- "$\alpha = 0.05$ is the probability I'm wrong."
- "A non-significant result means H_0 is true."

These are all probability statements about hypotheses or parameters given data, $P(H_0 \mid \text{data})$ or $P(\beta_1 \in CI \mid \text{data})$.

What it *actually* means

- p: probability of data this extreme given H_0 true, $P(\text{data} \mid H_0)$.
- 95% CI: across hypothetical re-samples, 95% of the *constructed intervals* cover β_1. Any single realised interval either covers or doesn't.
- α: long-run type I error rate of the procedure.
- Non-significant $\neq$ "no effect" – only "no detectable effect at this n and α."

All are statements about the procedure, not the parameter.

The intuitive readings *are* legitimate — under Bayesian inference. With a prior on β_1, the posterior $P(\beta_1 \mid \text{data})$ supports direct statements like *"there's a 95% chance* $\beta_1 \in [0.20, 0.40]$*"*. That's the **credible interval** — coming up in **Session 5: online A/B testing.**

R-squared in the single-treatment regression equals the squared correlation

General: percentage of variance in Y explained by all variables in the model:

$$R^2 = 1 - \frac{SSR}{SST}, \qquad SSR = \sum_i (Y_i - \hat{Y}_i)^2, \qquad SST = \sum_i (Y_i - \bar{Y})^2.$$

Specific case — only a single binary treatment D in the model:

$$R^2 = r^2 = \left(\frac{\sum_i (Y_i - \bar{Y})(D_i - \bar{D})}{\sqrt{\sum_i (Y_i - \bar{Y})^2} \cdot \sqrt{\sum_i (D_i - \bar{D})^2}} \right)^2.$$

In a causal-inference context, R^2 is largely **diagnostic** — high R^2 does not mean the regression is causally identified, and low R^2 does not preclude precise ATE estimation. We rarely report it as a primary result.

Job Corps via OLS with HC-robust SEs

Estimate the ATE of Job Corps assignment on year-4 weekly earnings using OLS with heteroskedasticity-robust standard errors.

R Python

```r
library(readr)
library(here)
library(sandwich)
library(modelsummary)

jc <- read_csv(here("materials/data/jc.csv"),
               show_col_types = FALSE)
ols <- lm(earny4 ~ assignment, data = jc)

modelsummary(ols,
  vcov = sandwich::vcovHC,
  estimate = "{estimate} (se = {std.error}, t = {statistic}){stars}",
  statistic = "p = {p.value}, CI = [{conf.low}, {conf.high}]",
  stars = TRUE,
  gof_map = c("r.squared", "nobs"),
  output = "markdown")
```

	1.
(Intercept)	197.926 (se = 3.073, t = 64.416)***
	p = <0.001, CI = [191.903, 203.949]
assignment	16.055 (se = 4.074, t = 3.941)***
	p = <0.001, CI = [8.069, 24.041]
R2	0.002
Num.Obs.	9240

Significance: + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.

Bootstrap: resample to estimate $se(\hat{\beta}_1)$

The sample stands in for the population: resampling with replacement mimics fresh draws, and the resulting variability proxies the true sampling variability.

The procedure (Huber 2023):

1. Draw B bootstrap samples of size n **with replacement** from the original sample.
2. Estimate $\hat{\beta}_1^{(b)}$ in each.
3. The bootstrap SE is the standard deviation across the B replicates:

$$se(\hat{\beta}_1) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}_1^{(b)} - \bar{\hat{\beta}}_1)^2}.$$

R Python

```r
library(readr); library(here); library(boot)
jc <- read_csv(here("materials/data/jc.csv"),
               show_col_types = FALSE)
bs <- function(data, idx) {
  coef(lm(earny4 ~ assignment, data = data[idx, ]))[2]
}
set.seed(1)
res <- boot(data = jc, statistic = bs, R = 1999)
c(est = res$t0, boot_se = sd(res$t))
```

```
est.assignment        boot_se
      16.05513        3.95481
```

Bootstrap SE ≈ HC-robust SE in well-behaved samples.

When bootstrap > HC-robust: small n, heavy tails, clustered/dependent data, or non-smooth statistics (medians, quantiles, ratios). Otherwise HC1 is faster and equally honest.

Extensions

OLS extends naturally to multivalued and continuous treatments — and to covariate-adjusted variance reduction.

Multivalued treatments: dummy-encode each level

Treatment D can take discrete values, ordered or unordered:

- Ordered: $D \in \{0, 1, 2, \ldots\}$ (e.g., 0 / 1 / 2 weeks of training).
- Unordered: $D \in \{A, B, C, \ldots\}$ (e.g., no training, IT, management).

The independence assumption generalises to all treatment levels:

$$\{Y_i(0), Y_i(1), Y_i(2), \ldots, Y_i(J)\} \perp\!\!\!\perp D_i.$$

Encode each non-baseline level as a binary dummy $D_{ij} = \mathbf{1}\{D_i = j\}$ and run OLS:

$$\mathbb{E}[Y_i \mid D_i] = \underbrace{\beta_0}_{\mathbb{E}[Y|D=0]} + \underbrace{\beta_1}_{\Delta \text{ vs } 0 \text{ at } 1} D_{i1} + \underbrace{\beta_2}_{\Delta \text{ vs } 0 \text{ at } 2} D_{i2} + \ldots + \underbrace{\beta_J}_{\Delta \text{ vs } 0 \text{ at } J} D_{iJ}.$$

Each β_j is the ATE of level j vs. the baseline level 0.

Multivalued treatments example

804 university students randomized into **three mutually exclusive arms: control** (default survey: professional-plans questions first, then personal-info questions), **information** (shown a graph of median monthly Swiss wages before being asked their own expectation), **order** (default sequence reversed — *personal questions first, professional plans second* — pure framing, no extra information). Outcome wexpect2 = own wage expectation 3 years after graduation, in 500 CHF brackets (Fernandes, Huber, and Vaccaro 2020).

R Python

```r
library(readr); library(here)
library(sandwich); library(modelsummary)
wexpect <- read_csv(here("materials/data/wexpect.csv"),
                    show_col_types = FALSE)
ols <- lm(wexpect2 ~ treatmentinformation + treatmentorder,
          data = wexpect)
modelsummary(ols,
  vcov = sandwich::vcovHC,
  estimate = "{estimate} (se = {std.error}, t = {statistic}){stars}",
  statistic = "p = {p.value}, CI = [{conf.low}, {conf.high}]",
  stars = TRUE,
  gof_map = c("r.squared", "nobs"),
  output = "markdown")
```

	1.
(Intercept)	9.408 (se = 0.159, t = 59.268)***
	p = <0.001, CI = [9.096, 9.719]
treatmentinformation	0.345 (se = 0.243, t = 1.421)
	p = 0.156, CI = [-0.132, 0.822]
treatmentorder	-0.173 (se = 0.234, t = -0.741)
	p = 0.459, CI = [-0.633, 0.286]
R2	0.006
Num.Obs.	804

Significance: + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.

Continuous treatments: ATE = marginal effect $\partial \mathbb{E}[Y \mid D]/\partial D$

Treatment D takes (continuum-) many values respecting cardinality: marketing spend in EUR, years of education, dose in mg.

Independence assumption for every value of the continuous treatment:

$$Y_i(d) \perp\!\!\!\perp D_i \qquad \text{for all } d.$$

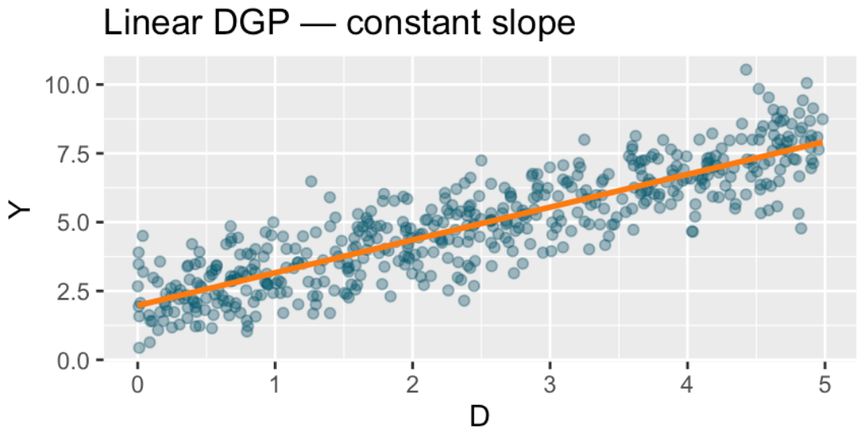
Two strategies:

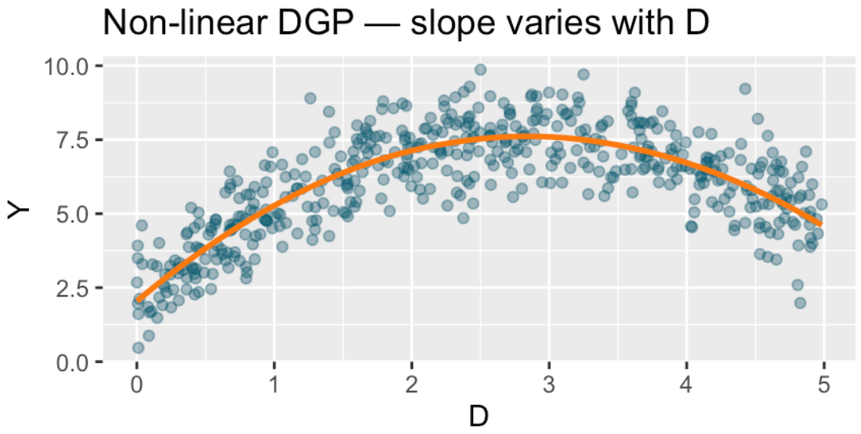
1. **Discretise** by binning into narrow brackets — $D_{i0} = \mathbf{1}\{D_i \leq 100\}$, $D_{i1} = \mathbf{1}\{100 < D_i \leq 200\}$, ...; reduce to multivalued case.
2. **Include D_i continuously** in the regression: $\mathbb{E}[Y_i \mid D_i] = \beta_0 + \beta_1 D_i$. Then the **ATE = marginal effect:**

$$ATE = \frac{\partial \mathbb{E}[Y_i \mid D_i]}{\partial D_i} = \beta_1.$$

Continuous treatments — non-linearity via quadratic specification

Linearity assumption: $\partial \mathbb{E}[Y_i(D = t')]/\partial D_i = \partial \mathbb{E}[Y_i(D = t)]/\partial D_i$ for all $t \neq t'$ in the support of D_i — i.e., the ATE is constant in D.





Quadratic specification: $\mathbb{E}[Y_i \mid D_i] = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 \Rightarrow ATE(D_i) = \partial \mathbb{E}[Y_i \mid D_i]/\partial D_i = \beta_1 + 2\beta_2 D_i.$

For maximum flexibility: higher-order polynomials, splines, or kernel regression.

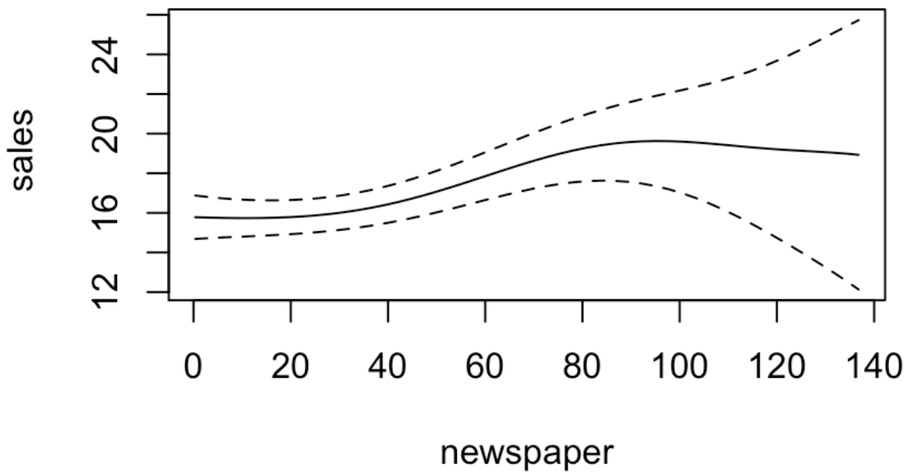
Continuous treatments example: newspaper ad spend → sales (kernel regression)

Kernel regression estimates $\widehat{\mathbb{E}}[Y \mid D = d]$ as a *locally weighted average* of nearby Y_i — observations close to d get high weight, distant ones get low weight; the **bandwidth** controls "close".

200 brand-week observations, marketing-spend dataset (datarium::marketing). Outcome: sales; treatment: newspaper-ad spend (continuous, USD).

R Python

```r
library(readr); library(here); library(np)
options(np.messages = FALSE)
mkt <- read_csv(here("materials/data/marketing.csv"),
                show_col_types = FALSE)
bw <- npregbw(sales ~ newspaper, data = mkt)
plot(bw, plot.errors.method = "asymptotic")
```



Including covariates reduces variance — under randomization, NOT bias correction

Under randomization, $D \perp X$, so adding pre-treatment covariates does **not** change the population value of $\hat{\beta}_1$ — it cannot fix bias that does not exist.

What it does change is **precision**: the residual variance shrinks, and so does $se(\hat{\beta}_1)$.

Augment the regression with covariates:

$$Y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 D_i + \hat{\beta}_{X_1} X_{i1} + \ldots + \hat{\beta}_{X_K} X_{iK}}_{\widehat{\mathbb{E}}[Y_i | D_i, X_i]} + \hat{\epsilon}_i.$$

The covariates explain part of the variation in Y that was previously residual:

$$R^2 \uparrow, \qquad \frac{\widehat{\mathrm{Var}}(\hat{\epsilon}_i)}{\widehat{\mathrm{Var}}(Y_i)} \downarrow, \qquad se(\hat{\beta}_1) \downarrow.$$

Under non-random assignment, covariates serve a different purpose

When D is not randomized, including covariates is for ***bias correction*** (de-confounding), not just variance reduction – Session 06 spends a full lecture on that. Don't confuse the two roles.

Lin (2013) covariate-adjusted estimator: interact D with mean-centred X

Freedman (2008): naively adding covariates additively, $Y \sim D + X$, can introduce small-sample bias when treatment-effect heterogeneity is present and X is unbalanced.

Why: $Y \sim D + X$ assumes one effect β_1 for everyone. With true heterogeneity $\tau(X)$, OLS instead recovers a $\widehat{\mathrm{Var}}(D \mid X)$-weighted average of conditional ATEs — and finite-sample imbalance in X tilts those weights off the marginal density that defines the ATE.

Lin (2013) proposes the canonical fix — interact D with **mean-centred** covariates:

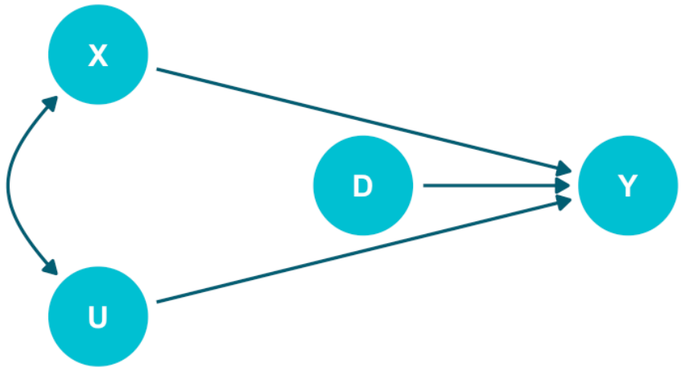
$$Y_i \;=\; \beta_0 + \beta_1 D_i + \boldsymbol{\gamma}^\top (X_i - \bar{X}) + \boldsymbol{\delta}^\top D_i \cdot (X_i - \bar{X}) + \epsilon_i.$$

The coefficient $\hat{\beta}_1$ is the Lin estimator of the ATE.

We'll use the Lin estimator in the **coding lab** and compare its SE to the unadjusted mean comparison and to the additive `lm(Y ~ D + X)`.

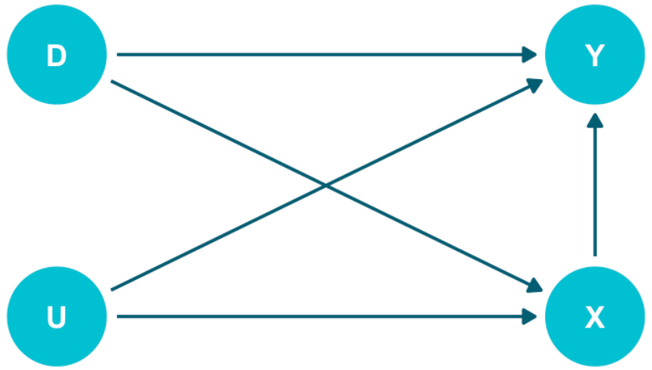
Pre-treatment vs. post-treatment covariates: only the former are safe

Pre-treatment X — safe



X is exogenous to D. Adding X to the regression reduces variance without changing the expectation of $\hat{\beta}_1$.

Post-treatment X — unsafe



Conditioning on X here opens a collider path through U — bridge to S03's collider lesson.

Two harms from conditioning on post-treatment X: (1) **mediator-blocking** — closes part of $D \to X \to Y$; (2) **collider-opening** — opens $D \to X \leftarrow U \to Y$, breaking randomization.

Including-covariates example: coffee-leaflet RCT

522 university students, randomly assigned to receive (or not) an environmental-impact leaflet about coffee. Outcome: awareness about waste production (awarewaste); covariates: mother's education (mumedu) and student sex.

R Python

```r
library(readr); library(here)
library(sandwich); library(modelsummary)
cl <- read_csv(here("materials/data/coffeeleaflet.csv"),
               show_col_types = FALSE)
ols <- lm(awarewaste ~ treatment + mumedu + sex, data = cl)
modelsummary(ols,
  vcov = sandwich::vcovHC,
  estimate = "{estimate} (se = {std.error}, t = {statistic}){stars}",
  statistic = "p = {p.value}, CI = [{conf.low}, {conf.high}]",
  stars = TRUE,
  gof_map = c("r.squared", "nobs"),
  output = "markdown")
```

	1.
(Intercept)	1.187 (se = 0.249, t = 4.770)***
	p = <0.001, CI = [0.698, 1.676]
treatment	0.332 (se = 0.096, t = 3.449)***
	p = <0.001, CI = [0.143, 0.521]
mumedu	0.272 (se = 0.090, t = 3.007)**
	p = 0.003, CI = [0.094, 0.450]
sex	0.137 (se = 0.100, t = 1.370)
	p = 0.171, CI = [-0.060, 0.333]
R2	0.046
Num.Obs.	466

Significance: + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.

The coefficient on treatment is the ATE of the leaflet on waste-awareness, with mumedu and sex absorbing residual variation in Y — variance reduction, not bias correction (since assignment was randomized).

References

Fernandes, Ana, Martin Huber, and Giannina Vaccaro. 2020. "Gender Differences in Wage Expectations." https://arxiv.org/abs/2003.11496.

Huber, Martin. 2023. *Causal Analysis: Impact Evaluation and Causal Machine Learning with Applications in R*. Cambridge, MA: MIT Press.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

Thank you for your attention!

TUHH
Institute of
Entrepreneurship
STARTUPENGINEER.io



www.startupengineer.io/team/ihl/

christoph-ihl

christophihl