Lecture 6
Regression for completely randomized experiment

Outline

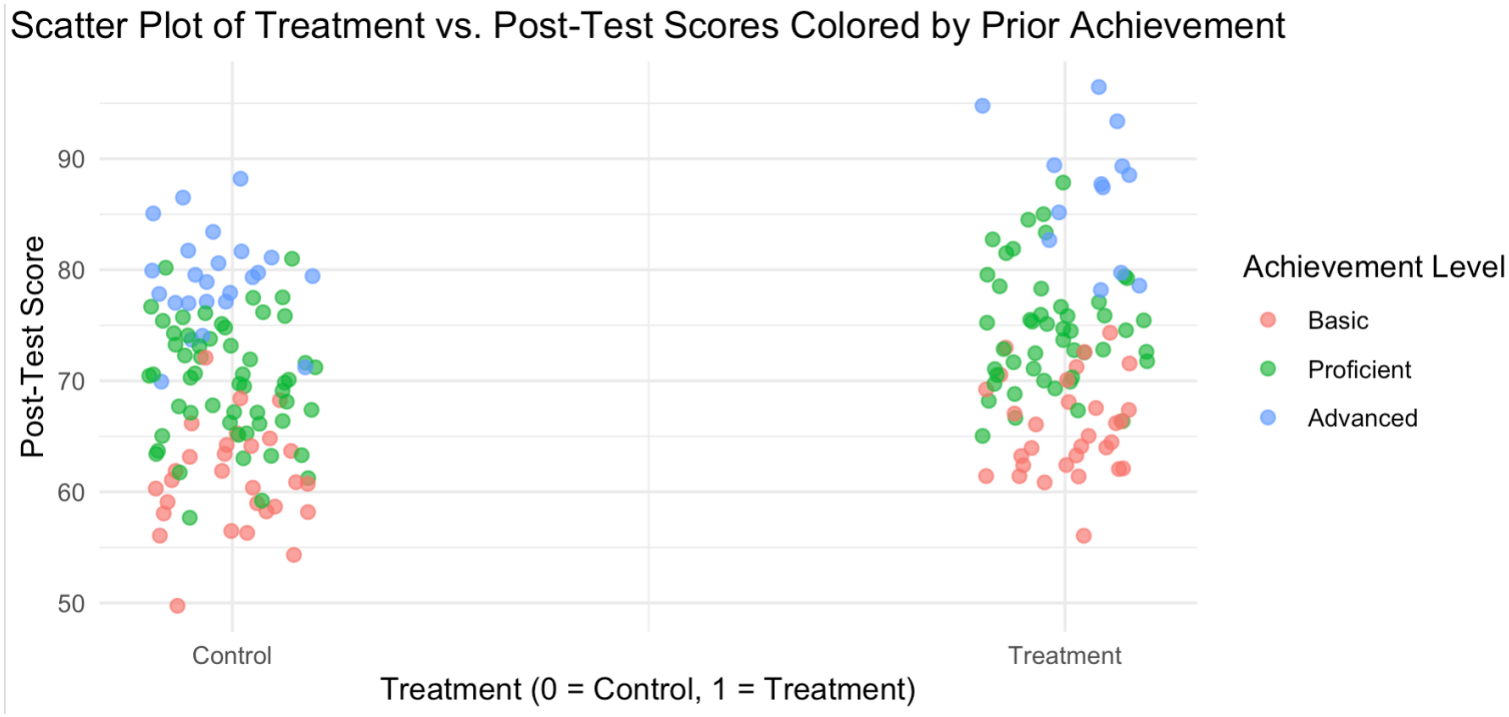
- Regression without covariates
- Regression with covariates adjustments: use baseline predictors to improve precision.
- Regression with covariates adjustments and interactions
- The LRC-CPPT cholesterol data example
- Reading: Imbens and Rubin, Chapter 7; Ding, Chapters 6 and 9

Potential outcomes and random assignment justify the treatment coefficient.

Recall: does random assignment make the realized groups exactly balanced?

Why adjust for prior achievement?

Teaching-method example: new instruction versus standard instruction.



Pair discussion: what variation could prior achievement explain?
Why is the treatment comparison causal?

The LRC-CPPT cholesterol experiment

337 patients: 165 assigned drug cholestyramine on reducing cholesterol and 172 assigned control.

Baseline: chol1 and chol2, measured before random assignment.
cholp = 0.25 chol1 + 0.75 chol2. Outcome: cholf.

Mean cholesterol	**Control**	**Treatment**
Baseline: cholp	291.2	289.9
Post-treatment: cholf	282.7	256.5
Change: chold = cholf − cholp	−8.5	−33.4

Both groups improved. Which comparison estimates the treatment effect?

Imbens & Rubin, Chapter 7. Summary values are rounded.

Potential-outcome means and PATE

- Assume that there is a super-population and the potential outcomes $(Y_i(0), Y_i(1), X_i)$ are i.i.d. samples
- Define the two population means and their difference:

$$\alpha = \mathrm{E}[Y_i(0)], \qquad \tau = \mathrm{E}[Y_i(1) - Y_i(0)]$$

- For $w = 0, 1$, define each unit's deviation from its population mean:

$$\varepsilon_i(w) := Y_i(w) - \mathrm{E}[Y_i(w)] = Y_i(w) - (\alpha + \tau w)$$

$$\mathrm{E}[\varepsilon_i(w)] = \mathrm{E}[Y_i(w)] - (\alpha + \tau w) = 0$$

Mean zero follows by definition.

$$Y_i(w) = \alpha + \tau w + \varepsilon_i(w)$$

- For binary w, this expresses two means. Individual effects may differ.

Quick check: express $Y_i(1) - Y_i(0)$ using τ, $\varepsilon_i(1)$, and $\varepsilon_i(0)$.
What condition makes every unit's effect equal to τ?

Unconfoundedness and the assignment mechanism

Lecture 3: W, X, and $Y(w)$ collect information for all N study units.

$$\Pr(\mathrm{W} = \mathrm{w} \mid \mathrm{X}, \mathrm{Y}(0), \mathrm{Y}(1)) = \Pr(\mathrm{W} = \mathrm{w} \mid \mathrm{X})$$

Once baseline X is fixed, assignment probabilities do not depend on potential outcomes.

Can assignment probabilities depend on X under unconfoundedness?
Does unconfoundedness require the W_i's to be independent?

Complete randomization is a special case. We connect it to regression next.

Randomization justifies the observed regression

Complete randomization chooses N_t treated units independently of the potential outcomes and baseline covariates.

Unconditional unconfoundedness (Lecture 5) and consistency:

$$\big(Y_i(0), Y_i(1)\big) \perp\!\!\!\perp W_i, \qquad Y_i^{\text{obs}} = Y_i(W_i) = W_i Y_i(1) + (1 - W_i) Y_i(0)$$

$$\mathrm{E}[Y_i^{obs} \mid W_i = w] = \mathrm{E}[Y_i(w) \mid W_i = w]$$

$$= \mathrm{E}[Y_i(w)] = \alpha + \tau w$$

Thus the observed regression has treatment coefficient τ = PATE.

Before conditioning on assignment, are Y_i^{obs} i.i.d. under complete randomization?
Conditional on the entire $(W_1 \cdots, W_n)$, are Y_i^{obs} independent across i? Are they identically distributed?

Linear regression without covariates

$$Y_i^{obs} = a + bW_i + \varepsilon_i$$

Least squares fits one mean in each treatment group:

$$\hat{a} = \overline{Y}_c^{obs}, \qquad \hat{a} + \hat{b} = \overline{Y}_t^{obs}$$

$$\hat{b} = \overline{Y}_t^{obs} - \overline{Y}_c^{obs} = \hat{\tau}$$

If we reverse the 0/1 coding of treatment, what happens to the coefficient and the fitted outcomes?

What supplies the regression uncertainty?

- regression model

$$Y_i^{\text{obs}} = \alpha + \tau W_i + \varepsilon_i$$

  where observed errors are $\varepsilon_i = \varepsilon_i(W_i)$

- PATE inference adds a sampling model for the study units:

$$(Y_i(0), Y_i(1), X_i) \text{ i. i. d.} \sim P$$

- Follow the linear regression convention, we perform inference conditional on $(W_1, \cdots, W_N)$
  - we treat assignment vectors as fixed
  - Then $\varepsilon_i = \varepsilon_i(W_i)$ are independent across i as $(\varepsilon_i(0), \varepsilon_i(1))$ are independent across i

Does randomization imply equal error variances in the two groups?

One estimate, two standard errors

Equal error variances (homoscedastic error): use a pooled within-group variance.

$$s_p^2 = \frac{(N_t - 1)s_t^2 + (N_c - 1)s_c^2}{N - 2}, \qquad \mathrm{SE}_{hom}^2 = s_p^2\left(\frac{1}{N_t} + \frac{1}{N_c}\right)$$

Allow unequal variances (heteroscedastic): use the robust HC2 standard error.

$$\mathrm{SE}_{HC2}^2 = \frac{s_t^2}{N_t} + \frac{s_c^2}{N_c} = \hat{V}$$

HC2 equals Neyman's variance estimator for this unadjusted regression.

$$95\%\ \mathrm{CI}:\ \hat{\tau} \pm 1.96 \times \mathrm{SE}$$

Covariates in the potential-outcome model

Let X be a pre-treatment predictor, such as baseline cholesterol.

Conditional unconfoundedness and overlap:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i, \qquad 0 < \Pr(W_i = 1 \mid X_i = x) < 1$$

- At the same X_i, assignment gives no information about potential outcomes.
- Complete randomization guarantees this for baseline X_i.
- Two additional conditional-mean assumptions needed for the additive regression:

$$\mathrm{E}[Y_i(0) \mid X_i = x] = \alpha + \beta x, \qquad \mathrm{E}[Y_i(1) - Y_i(0) \mid X_i = x] = \tau$$

  - The second assumption requires PATE constant across levels of X_i
- Consistency and conditional unconfoundedness give:

$$\mathrm{E}[Y_i^{obs} \mid W_i = w, X_i = x] = \mathrm{E}[Y_i(w) \mid X_i = x] = \alpha + \tau w + \beta x$$

Does including X in a regression itself guarantee unconfoundedness?

How covariate adjustment changes the estimate

$$\hat{\tau}_{\text{adj}} = \overline{Y}_t^{obs} - \overline{Y}_c^{obs} - \hat{\beta}(\overline{X}_t - \overline{X}_c)$$

- Adjust the observed difference for baseline imbalance and its fitted slope.
- Even if the model is incorrect, $\hat{\tau}_{\text{adj}}$ is still consistent for PATE.

Teaching-method simulation

Dashed line: true SATE



- The efficiency gain might be lost if the linear model is incorrect.

If the treated group has higher baseline achievement and β > 0,
which way does the adjustment move the estimate?

Statistical inference after covariate adjustment

- Assume homoscedastic error assumption:

$$\mathbb{V}(\varepsilon_i(0)) = \mathbb{V}(\varepsilon_i(1)) = \sigma^2 = \mathbb{V}\left(Y_i^{\text{obs}} | W_i, \boldsymbol{X}_i\right)$$

  We can follow standard linear regression inference and estimate variance of $\hat{\tau}^{\text{ols}}$ as

$$\hat{\mathbb{V}}_{\text{sp}}^{\text{homo}} = \frac{1}{N\,(N-1-\dim(X_i))} \cdot \frac{\sum_{i=1}^{N}\left(Y_i^{\text{obs}} - \hat{\alpha}^{\text{ols}} - \hat{\tau}^{\text{ols}} - X_i\hat{\beta}^{\text{ols}}\right)^2}{\overline{W}\cdot(1-\overline{W})}$$

- Use the robust variance HC2 estimator without assuming homoscedasticity

$$\hat{\mathbb{V}}_{\text{sp}}^{\text{hetero}} = \frac{1}{N\,(N-1-\dim(X_i))} \cdot \frac{\sum_{i=1}^{N}\left(W_i-\overline{W}\right)^2 \cdot \left(Y_i^{\text{obs}} - \hat{\alpha}^{\text{ols}} - \hat{\tau}^{\text{ols}} - X_i\hat{\beta}^{\text{ols}}\right)^2}{\left(\overline{W}\cdot(1-\overline{W})\right)^2}$$

Interactions allow the effect to vary with X

Is this assumption $\tau \equiv \tau(\boldsymbol{x}) = \mathbb{E}(\tau_i | \ \boldsymbol{X}_i = \boldsymbol{x})$ reasonable?

- Effect heterogeneity across gender, age, pre-existing conditions …
- How do we allow such heterogeneity in linear regression?

Let the conditional average effect depend on baseline X:

$$\tau(x) = \mathrm{E}[Y_i(1) - Y_i(0) \mid X_i = x] = \tau + \gamma(x - \mu_X)$$

With $\mu_X = E(X)$, averaging $\tau(X)$ gives PATE $= \tau$.

- Still assume $\mathrm{E}[Y_i(0) \mid X_i = x] = \alpha + \beta x$
- Complete randomization implies that:

$$\mathrm{E}[Y_i^{obs} \mid W_i = w, X_i = x] = \alpha + \beta x + \tau w + \gamma(x - \mu_X)w$$

Treatment slope: β + γ. Control slope: β.

Centering and the treatment coefficient

$$X_i^c = X_i - \overline{X}, \qquad Y_i^{obs} = \alpha + \tau W_i + \beta X_i^c + \gamma W_i X_i^c + \varepsilon_i$$

$$\hat{\tau}(X_i) = \hat{\tau} + \hat{\gamma}(X_i - \overline{X})$$

$$\frac{1}{N}\sum_{i=1}^{N} \hat{\tau}(X_i) = \hat{\tau}$$

- What is the benefit of adding interactions
  - More flexible model assumptions
  - Further increase efficiency
  - In completely randomized experiments, with the interaction terms, we can always guarantee no efficiency loss even when the linear model is wrong (Peng's book section 6.2.2)

1. What does the W coefficient mean without centering?
2. Does centering change the fitted outcomes?

Fitting the models in R

Y: outcome. W: assignment (0/1). X: a numeric baseline covariate.

```
d$Xc <- d$X - mean(d$X)
m0 <- lm(Y ~ W, data = d)
m1 <- lm(Y ~ W + Xc, data = d)
m2 <- lm(Y ~ W * Xc, data = d)
```

```
fit <- m2
est <- coef(fit)["W"]
se_hom <- sqrt(vcov(fit)["W", "W"])
se_HC2 <- sqrt(sandwich::vcovHC(fit, type = "HC2")["W", "W"])
```

Several covariates: W * (X1c + X2c + …). Center indicator columns too.

Precision and model assumptions

Predictive baseline covariates can reduce unexplained outcome variation.

Under complete randomization, adjusted estimates remain consistent even if the linear mean model is imperfect.

Fully interacted adjustment has an asymptotic precision guarantee relative to the unadjusted mean difference, under regularity conditions.

Would you expect every added covariate to reduce the estimated SE in this particular dataset? Why?

Assignment and compliance in the cholesterol trial

Random assignment: drug (W = 1) or placebo (W = 0).
Patients may take less than the assigned dose.

Compliance (comp) is the percentage of the assigned dose actually taken.
It measures adherence to drug in one group and to placebo in the other.

Two outcomes for the same randomized comparison

Cholesterol: does assignment lower post-treatment cholesterol?
Compliance: does assignment change adherence?

Compliance is measured after assignment and can be affected by it.

Quick check: for the cholesterol analysis, can we adjust for observed compliance as we do for baseline cholesterol?

Results for the cholesterol experiment

Drug versus placebo assignment. Entries are estimates (standard errors).

Compliance estimates: about 15 percentage points lower under drug assignment.

Covariates	**Post-treatment cholesterol**	**Compliance (percentage points)**
None	−26.22 (3.93)	−14.64 (3.51)
cholp	−25.01 (2.60)	−14.68 (3.51)
chol1, chol2	−25.02 (2.59)	−14.95 (3.50)
chol1, chol2 and treatment interactions	−25.04 (2.56)	−14.94 (3.49)

Pair discussion: why might adjustment help cholesterol precision much more than compliance precision?

Imbens & Rubin, Chapter 7. Reported results retained from the original example.

Regression and the randomized comparison

Randomization connects potential-outcome means to observed regression.

Baseline adjustment can improve precision. Interactions allow conditional average effects to vary.

Fisher inference can use the regression estimate as its statistic.
Refit it under each assignment using the same sharp-null procedure.

Final check: explain the roles of randomization, centering,
and robust standard errors in one sentence each.

Reading: Imbens & Rubin, Chapter 7; Ding, Chapters 6 and 9