Lecture 6
Regression for completely randomized experiment

Outline

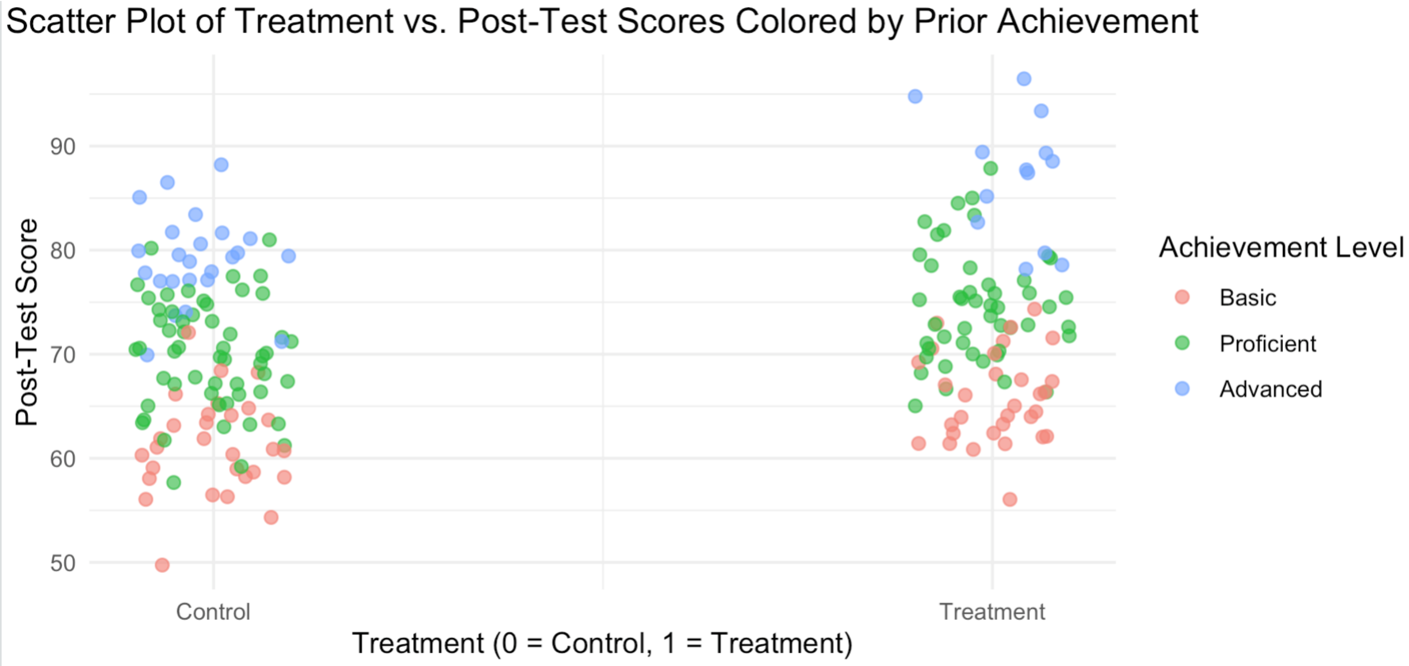
- Using regression with no covariates
- Using regression with covariates adjustments
- Using regression with covariates adjustments and interactions
- The LRC-CPPT cholesterol data example
- Suggested readings: Imbens and Rubin Chapter 7

Linear regression and causality

- Linear regression:

$$\mathbb{E}(Y_i|W_i,\boldsymbol{X}_i) = \alpha + \gamma W_i + \boldsymbol{\beta}^T \boldsymbol{X}_i$$

- Benefits of using linear regression:
  - Adjust for confounding variables
    - Not need for completely randomized experiments as pre-treatment covariates are not confounded
  - More accurate estimator if covariates explain part of the noise in the outcome



Linear regression and causality

- Linear regression:

$$\mathbb{E}(Y_i | W_i, \boldsymbol{X}_i) = \alpha + \gamma W_i + \boldsymbol{\beta}^T \boldsymbol{X}_i$$

- Question:
  - When can we interpret the coefficient(s) as causal effect?
  - How can we do correct inference if we take into account the randomization procedure of treatment assignments?

- Some critiques
  - In completely randomized experiments, covariates are not confounders
  - Why do we want to assume a linear model if we don’t need to?
    - Model $\mathbb{E}(Y_i | W_i, \boldsymbol{X}_i) = \alpha + \gamma W_i + \boldsymbol{\beta}^T \boldsymbol{X}_i$ assumes same causal effect for all levels of $\boldsymbol{X}_i$

“Experiments should be analyzed as experiments, not as observational studies”

---- David A. Freedman, 2006

The LRC-CPPT cholesterol data

- An experiment to evaluate the effect of the drug cholestyramine on reducing cholesterol levels
- $N = 337$ patients are completely randomized
- **Pre-treatment covariates:** two cholesterol measurements before and after a suggestion of low-cholesterol diet, both measurements taken prior to the random assignment
    - cholp = 0.25 chol1 + 0.75 chol2

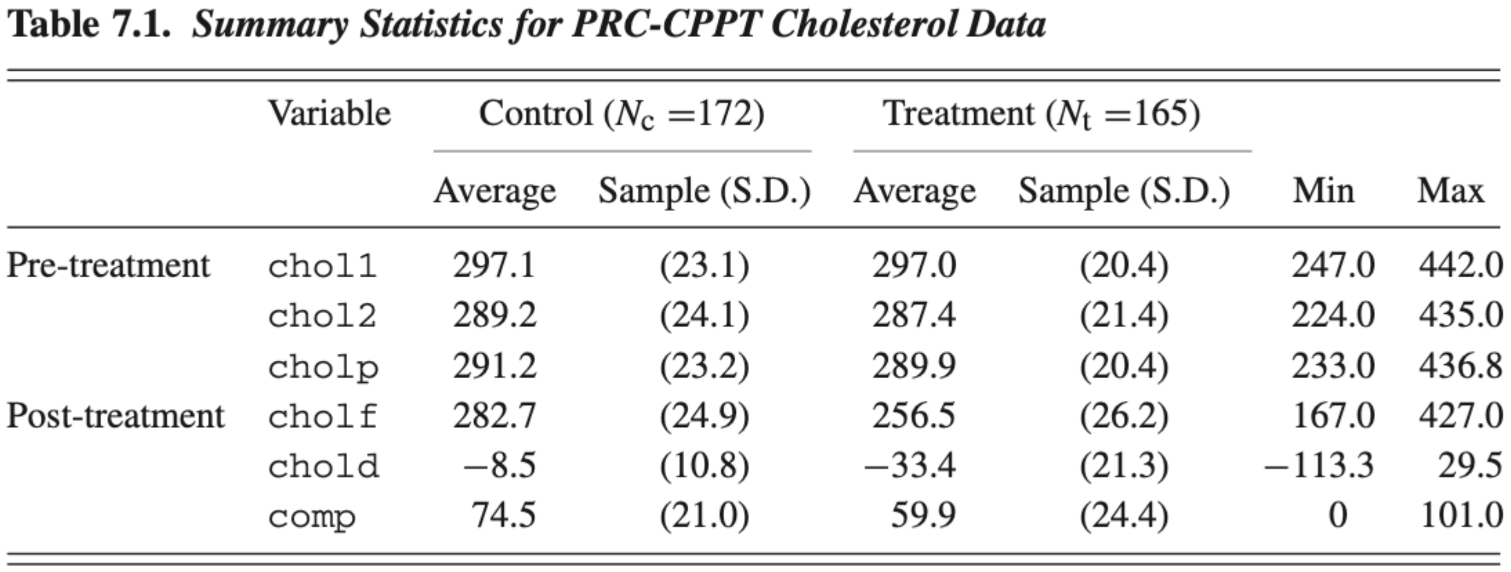


Table 7.1. ***Summary Statistics for PRC-CPPT Cholesterol Data***

	Variable	Control (N_c =172)		Treatment (N_t =165)			
		Average	Sample (S.D.)	Average	Sample (S.D.)	Min	Max
Pre-treatment	`chol1`	297.1	(23.1)	297.0	(20.4)	247.0	442.0
	`chol2`	289.2	(24.1)	287.4	(21.4)	224.0	435.0
	`cholp`	291.2	(23.2)	289.9	(20.4)	233.0	436.8
Post-treatment	`cholf`	282.7	(24.9)	256.5	(26.2)	167.0	427.0
	`chold`	−8.5	(10.8)	−33.4	(21.3)	−113.3	29.5
	`comp`	74.5	(21.0)	59.9	(24.4)	0	101.0

The LRC-CPPT cholesterol data

- An experiment to evaluate the effect of the drug cholestyramine on reducing cholesterol levels
- $N = 337$ patients are completely randomized
- **Post-treatment outcomes:**
    - cholf: post-treatment average cholesterol level
    - chold = cholf – cholp
    - comp: compliance rate, the percentage of individuals follow the treatment assignment

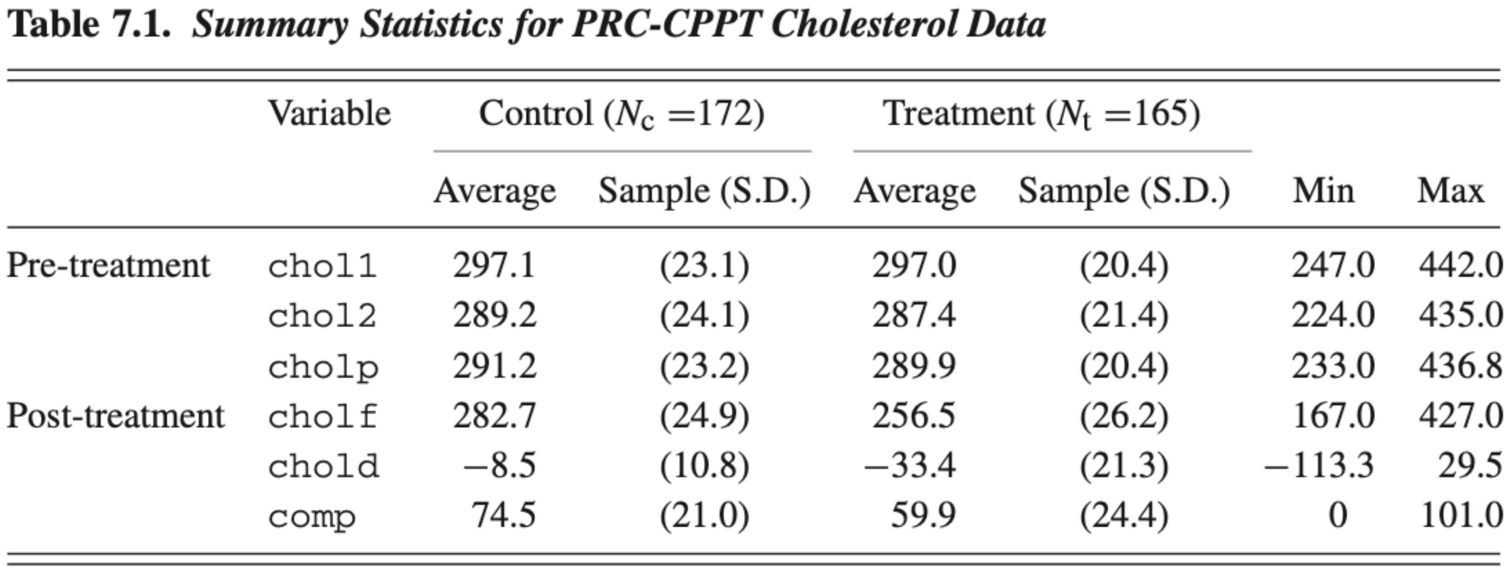


Table 7.1. ***Summary Statistics for PRC-CPPT Cholesterol Data***

	Variable	Control (N_c =172)		Treatment (N_t =165)			
		Average	Sample (S.D.)	Average	Sample (S.D.)	Min	Max
Pre-treatment	`chol1`	297.1	(23.1)	297.0	(20.4)	247.0	442.0
	`chol2`	289.2	(24.1)	287.4	(21.4)	224.0	435.0
	`cholp`	291.2	(23.2)	289.9	(20.4)	233.0	436.8
Post-treatment	`cholf`	282.7	(24.9)	256.5	(26.2)	167.0	427.0
	`chold`	−8.5	(10.8)	−33.4	(21.3)	−113.3	29.5
	`comp`	74.5	(21.0)	59.9	(24.4)	0	101.0

The LRC-CPPT cholesterol data

- Can we evaluate the drug effect by simply look at whether chold is positive or negative?
  - No! The before-after comparison is NOT necessarily causal
  - Even for the control group, chold is significantly negative

- The patient’s post-treatment cholesterol should be highly correlated with his/her pre-treatment cholesterol level
- How do we evaluate the causal effect after “adjusting for the pre-treatment cholesterol”?
  - Adjust for pre-treatment cholesterol by regression

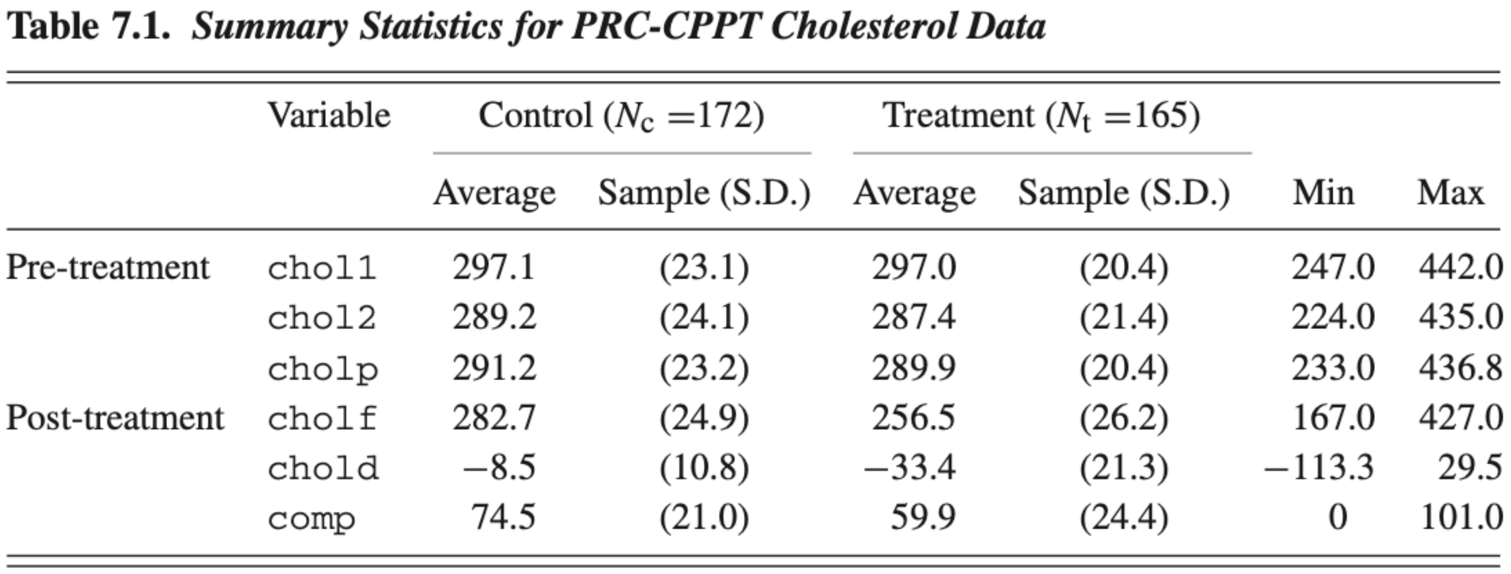


Table 7.1. ***Summary Statistics for PRC-CPPT Cholesterol Data***

	Variable	Control (N_c =172)		Treatment (N_t =165)			
		Average	Sample (S.D.)	Average	Sample (S.D.)	Min	Max
Pre-treatment	chol1	297.1	(23.1)	297.0	(20.4)	247.0	442.0
	chol2	289.2	(24.1)	287.4	(21.4)	224.0	435.0
	cholp	291.2	(23.2)	289.9	(20.4)	233.0	436.8
Post-treatment	cholf	282.7	(24.9)	256.5	(26.2)	167.0	427.0
	chold	−8.5	(10.8)	−33.4	(21.3)	−113.3	29.5
	comp	74.5	(21.0)	59.9	(24.4)	0	101.0

Linear regression with no covariates

- Neyman's approach

$$\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$$

$$\overline{Y}_{\text{c}}^{\text{obs}} = \frac{1}{N_{\text{c}}} \sum_{i:W_i=0} Y_i^{\text{obs}} \quad \text{and} \quad \overline{Y}_{\text{t}}^{\text{obs}} = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} Y_i^{\text{obs}}$$

- An alternative way to get $\hat{\tau}^{\text{dif}}$ is by linear regression

$$(\hat{a}, \hat{b}) = \arg\min_{(a,b)} \sum_{i=1}^{N} (Y_i - a - bW_i)^2$$

- It is easy to show that $\hat{b} = \hat{\tau}^{\text{dif}}$
  - $\arg\min_{(a,b)} \sum_{i=1}^{n} (Y_i - a - bW_i)^2 = \operatorname{argmin}_{(a,b)} \left[\sum_{i:W_i=0} (Y_i - a)^2 + \sum_{i:W_i=1} (Y_i - a - b)^2 \right]$
  - $\hat{a} = \overline{Y}_c^{\text{obs}}$, $\hat{a} + \hat{b} = \overline{Y}_t^{\text{obs}}$
- However, if we use R function lm(), the variance estimator is $\frac{1}{N} \frac{(N_c-1)s_c^2 + (N_t-1)s_t^2}{N-2}$
  - Different from $\frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}$ in Neyman's approach

Causal interpretation of the linear model

- Linear model on the potential outcomes
$$Y_i(w) = \alpha + \tau_i w + \varepsilon_i^* = \alpha + \tau w + \varepsilon_i(w)$$
where $\mathbb{E}(\varepsilon_i^*) = 0$ and $\varepsilon_i(w) = \varepsilon_i^* + (\tau_i - \tau)w$
  - Not really an assumption if w only has two values
$$Y_i(0) = \alpha + \varepsilon_i^*, Y_i(1) = Y_i(0) + \tau_i$$

- Assume that there is a super-population and the potential outcomes are i.i.d. samples
- The observed outcomes $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$ are not i.i.d. samples under complete randomization

- Define PATE: $\tau = \mathbb{E}(\tau_i) = \mathbb{E}\big(Y_i(1) - Y_i(0)\big)$
- $\alpha = \mathbb{E}\big(Y_i(0)\big)$ and $\mathbb{E}\big(\varepsilon_i(w)\big) = 0$

Linear regression with no covariates

- Causal model on the potential outcomes

$$Y_i(w) = \alpha + \tau_i w + \varepsilon_i^* = \alpha + \tau w + \varepsilon_i(w)$$

where $\mathbb{E}(\varepsilon_i^*) = 0$ and $\varepsilon_i(w) = \varepsilon_i^* + (\tau_i - \tau)w$

  - If the treatment is binary ($w = 0,1$), then the above model essentially has no assumption on $Y_i(0)$ and $Y_i(1)$
  - If the treatment is continuous, the model assumes a linear but heterogenous causal effect on each individual

  - How to estimate τ from observed data?
  - When does the above model imply the linear regression model on observed data?

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

Linear regression with no covariates

$$Y_i(w) = \alpha + \tau_i w + \varepsilon_i^* = \alpha + \tau w + \varepsilon_i(w)$$

We assume the following identification conditions

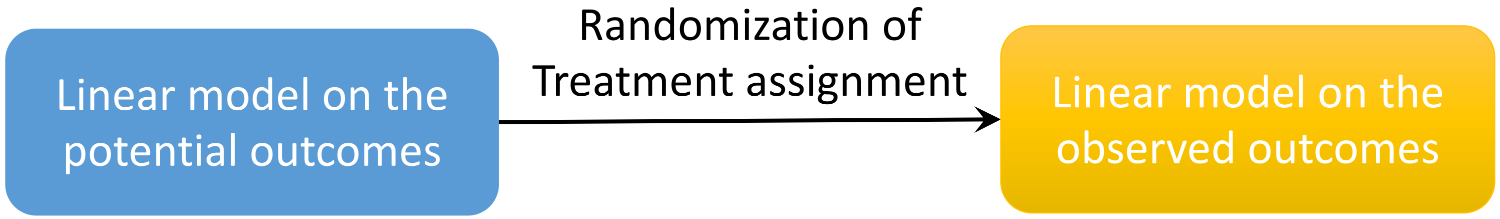
- Randomization of the treatment:

$$(\boldsymbol{Y}(0), \boldsymbol{Y}(1)) \perp \boldsymbol{W}$$

  - Satisfied in completely randomized experiments
  - Then, $\mathbb{E}\big(Y_i(w)\big) = \mathbb{E}(Y_i(w) | W_i = w) = \mathbb{E}\big(Y_i^{\text{obs}} | W_i = w\big) = \alpha + \tau w$
  - So this implies a regression model

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

where $\varepsilon_i = \varepsilon_i(W_i) = \varepsilon_i^* + (\tau_i - \tau) W_i$



Linear model on the potential outcomes

Randomization of Treatment assignment

Linear model on the observed outcomes

  - What is the correct statistical inference?

Linear regression with no covariates

- regression model

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

  where $\varepsilon_i = \varepsilon_i(W_i) = \varepsilon_i^* + (\tau_i - \tau)W_i$

- Follow the linear regression convention, we perform statistical inference conditional on $(W_1, \cdots, W_N)$
  - we treat assignment vectors as fixed

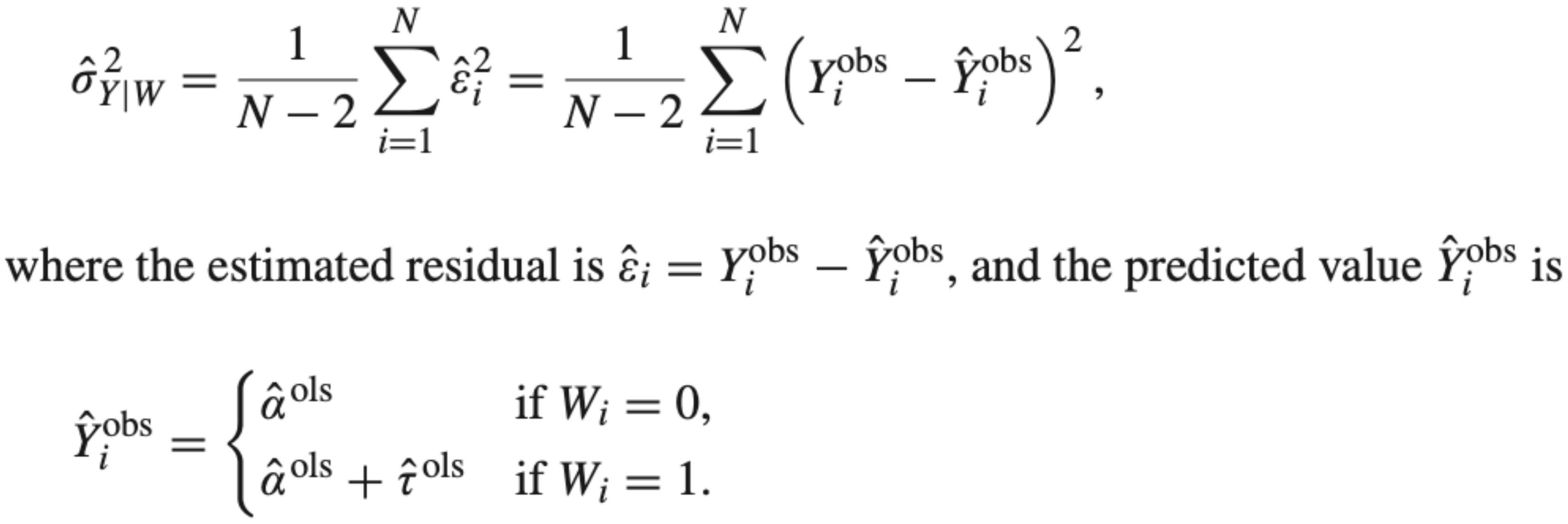
- Random sampling of the units
  - $(\varepsilon_i(0), \varepsilon_i(1))$ are independent across i
  - This implies that ε_i in the linear regression model are independent as W_i are treated as fixed
  - But they may not follow the same distribution

Homoscedastic error assumption

Homoscedastic error assumption: $\mathbb{V}(\varepsilon_i(0)) = \mathbb{V}(\varepsilon_i(1)) = \sigma^2$

- Then$\mathbb{V}\left(Y_i^{\text{obs}} | W_i\right) = \varepsilon_i = \varepsilon_i(W_i)$ always has variance σ^2

- Under homoscedasticity, OLS estimates of the variance is



$$\hat{\sigma}^2_{Y|W} = \frac{1}{N-2}\sum_{i=1}^{N} \hat{\varepsilon}_i^2 = \frac{1}{N-2}\sum_{i=1}^{N}\left(Y_i^{\text{obs}} - \hat{Y}_i^{\text{obs}}\right)^2,$$

where the estimated residual is $\hat{\varepsilon}_i = Y_i^{\text{obs}} - \hat{Y}_i^{\text{obs}}$, and the predicted value $\hat{Y}_i^{\text{obs}}$ is

$$\hat{Y}_i^{\text{obs}} = \begin{cases} \hat{\alpha}^{\text{ols}} & \text{if } W_i = 0, \\ \hat{\alpha}^{\text{ols}} + \hat{\tau}^{\text{ols}} & \text{if } W_i = 1. \end{cases}$$

  - Same as the standard linear regression approach

Heteroscedastic errors

- If we don’t want to assume $\mathbb{V}(\varepsilon_i(0)) = \mathbb{V}(\varepsilon_i(1))$, then the homoscedastic error assumption fails
  - ε_i has the same distribution for $W_i = 0$, and the same distribution for $W_i = 1$
  - We should use same variance within the treated and control group
  - That leads to the variance estimator in Neyman’s approach
- This is also called the Sandwich estimator that is robust to the violation of the homoscedastic noise assumption in linear regression
  - In R, it corresponds to Sandwich estimator with HC2 adjustment

Linear regression wit no covariates

To summarize the logic

- We build a (linear) model on the potential outcomes
- This model implies a linear regression model on the observed outcome if $(\boldsymbol{Y}(0), \boldsymbol{Y}(1)) \perp \boldsymbol{W}$
- The coefficient on W_i in the linear regression model is the average causal effect (PATE)
- The linear regression model treat $\boldsymbol{W}$ as fixed so it works for any randomization assignment mechanism that satisfy $(\boldsymbol{Y}(0), \boldsymbol{Y}(1)) \perp \boldsymbol{W}$
- Noise in the linear regression model are independent as long as potential outcomes are independent across units

- For statistical inference
  - The OLS estimator estimator is always unbiased
  - We can apply standard linear regression inference results if we assume $\mathbb{V}(\varepsilon_i(0)) = \mathbb{V}(\varepsilon_i(1))$
  - If $\mathbb{V}(\varepsilon_i(0)) \neq \mathbb{V}(\varepsilon_i(1))$, we need to use the robust variance estimator

Linear regression with covariates adjustment

- What are model assumptions on the potential outcomes that lead to

$$Y_i^{\text{obs}} = \alpha + \tau W_i + \boldsymbol{\beta}^T \boldsymbol{X}_i + \varepsilon_i$$

a linear model on the observed outcome

- Assumption 1: $\mathbb{E}(Y_i(0)|\, \boldsymbol{X}_i) = \alpha + \boldsymbol{\beta}^T \boldsymbol{X}_i$
- Assumption 2: CATE $\tau(\boldsymbol{x}) = \mathbb{E}(\tau_i|\, \boldsymbol{X}_i = \boldsymbol{x}) \equiv \tau = \text{PATE}$ constant across levels of $\boldsymbol{X}_i$
  - We can allow for heterogeneous causal effect but need $\mathbb{E}(\tau_i - \tau \mid \boldsymbol{X}_i) = 0$ (individual causal effects are independent from the pre-treatment covariates)

- Then $\mathbb{E}(Y_i(w)|\, \boldsymbol{X}_i) = \mathbb{E}(Y_i(0) + \tau_i w|\, \boldsymbol{X}_i) = \alpha + \tau w + \boldsymbol{\beta}^T \boldsymbol{X}_i$

- Unconfoundedness property:

$$\big(\boldsymbol{Y}(0), \boldsymbol{Y}(1)\big) \perp \boldsymbol{W} \mid \boldsymbol{X}$$

- $\mathbb{E}\big(Y_i^{\text{obs}}|W_i = w, \boldsymbol{X}_i = \boldsymbol{x}\big) = \mathbb{E}(Y_i(w)|\boldsymbol{X}_i = \boldsymbol{x}) = \alpha + \tau w + \boldsymbol{\beta}^T \boldsymbol{X}_i$
  - Statistical inference is conditional on both $\boldsymbol{X}_i$ and W_i

OLS with covariates adjustment

$$(\hat{\alpha}^{\text{ols}}, \hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}) = \arg\min_{\alpha,\tau,\beta} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i \beta \right)^2$$

- The estimator $\hat{\tau}^{\text{ols}}$ is unbiased for the causal estimand τ
- Even if the model is incorrect (either the violation of $\mathbb{E}(Y_i(0)|\, \boldsymbol{X}_i) = \alpha + \boldsymbol{\beta}^T \boldsymbol{X}_i$ or $\tau \equiv \mathbb{E}(\tau_i|\ \boldsymbol{X}_i = \boldsymbol{x})$), $\hat{\tau}^{\text{ols}}$ still converges to the PATE $\mathbb{E}(\tau_i)$ under complete randomization

Efficiency gain from regression

- If the model is correct, we have

$$\mathbb{V}\left(\hat{\tau}^{\text{ols}}\right) \approx \frac{\mathbb{E}\{\mathbb{V}(Y_i(1)|\, \boldsymbol{X}_i)\}}{N_t} + \frac{\mathbb{E}\{\mathbb{V}(Y_i(0)|\, \boldsymbol{X}_i)\}}{N_c} \leq \frac{\sigma_c^2}{N_c} + \frac{\sigma_t^2}{N_t}$$

  - If $\boldsymbol{X}_i$ is predictive of the (potential) outcomes, we have a more accurate estimator
- If the linear model is incorrect, the efficiency might be lost (Freedman 2008, *Adv. Appl. Math.)*

Estimate of the variance of $\hat{\tau}^{\mathrm{ols}}$ with covariates adjustment

- Assume homoscedastic error assumption:

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

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

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

- The robust variance estimator (Sandwich estimator) without assuming homoscedasticity

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

Linear regression with covariates adjustment and interactions

What if the assumption $\tau \equiv \tau(\boldsymbol{x}) = \mathbb{E}(\tau_i | \boldsymbol{X}_i = \boldsymbol{x})$ constant across levels of $\boldsymbol{X}_i$ is incorrect?

- Assume CATE $\tau(\boldsymbol{x}) = \mathbb{E}(\tau_i | \boldsymbol{X}_i = \boldsymbol{x}) = \tau + \boldsymbol{\gamma}^T(\boldsymbol{x} - \overline{\boldsymbol{X}})$
  - τ is still the population average treatment effect
- Still assume $\mathbb{E}(Y_i(0) | \boldsymbol{X}_i) = \alpha + \boldsymbol{\beta}^T \boldsymbol{X}_i$

- Then $\mathbb{E}(Y_i(w) | \boldsymbol{X}_i) = \mathbb{E}(Y_i(0) + \tau_i w | \boldsymbol{X}_i) = \alpha + \tau w + \boldsymbol{\beta}^T \boldsymbol{X}_i + \boldsymbol{\gamma}^T(\boldsymbol{X}_i - \overline{\boldsymbol{X}})w$

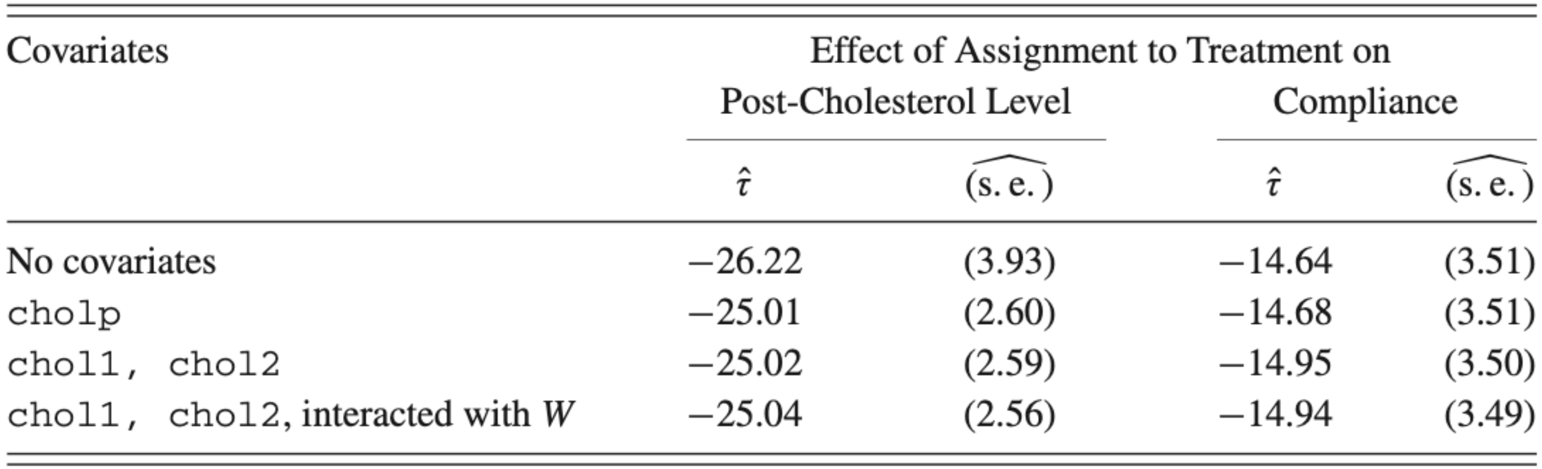
- When does the above model imply the linear regression model with interactions on observed data?

$$Y_i^{\text{obs}} = \alpha + \tau W_i + \boldsymbol{\beta}^T \boldsymbol{X}_i + \boldsymbol{\gamma}^T(\boldsymbol{X}_i - \overline{\boldsymbol{X}})W_i + \varepsilon_i$$

  - Unconfoundedness property → check by yourself
  - 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)

Results on the LRC-CPPT cholesterol data

- We estimate the PATE for both the post-treatment cholesterol level cholf and compliance
    - A considerable reduction of the variance of $\hat{\tau}^{\text{ols}}$ for cholf when we add the pre-treatment cholesterol levels in the regression
    - Our goal is always estimating PATE even after “covariates adjustment”
    - In randomized experiments satisfying $(\boldsymbol{Y}(0), \boldsymbol{Y}(1)) \perp \boldsymbol{W}$, adjusting for covariates or not, our estimate of PATE is always valid, we only change the efficiency of our estimate



Covariates	Effect of Assignment to Treatment on			
	Post-Cholesterol Level		Compliance	
	$\hat{\tau}$	$\widehat{(\text{s. e.})}$	$\hat{\tau}$	$\widehat{(\text{s. e.})}$
No covariates	−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`, interacted with *W*	−25.04	(2.56)	−14.94	(3.49)

The LRC-CPPT cholesterol data

A bit explanation about compliance

- If we compare between control and treatment group, we are evaluating the causal effect of “being assigned”, not the causal effect of actually taking the drug
- Compliance lower in the treatment group possibly due to the side effect of the drug
- Can we just throw away individuals who do not follow the treatment and estimate the causal effect of taking the drug based on the rest individuals? No
- Will discuss more about compliance in later lectures

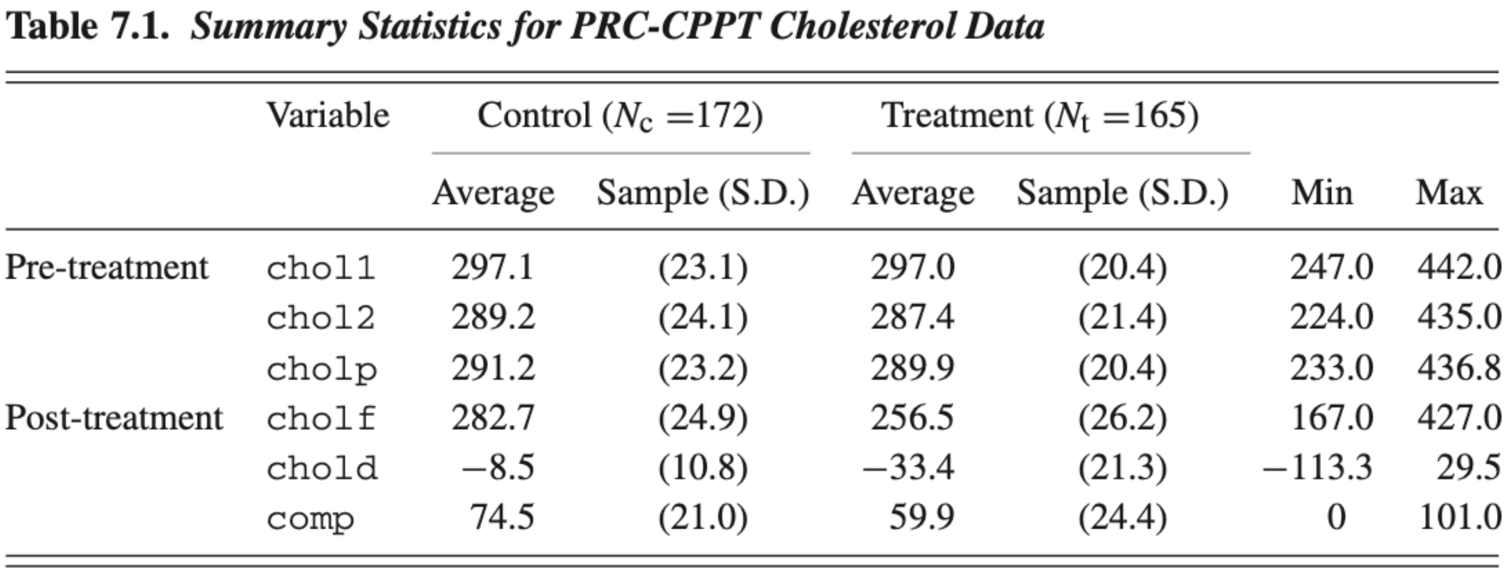


Table 7.1. ***Summary Statistics for PRC-CPPT Cholesterol Data***

	Variable	Control (N_c =172)		Treatment (N_t =165)			
		Average	Sample (S.D.)	Average	Sample (S.D.)	Min	Max
Pre-treatment	`chol1`	297.1	(23.1)	297.0	(20.4)	247.0	442.0
	`chol2`	289.2	(24.1)	287.4	(21.4)	224.0	435.0
	`cholp`	291.2	(23.2)	289.9	(20.4)	233.0	436.8
Post-treatment	`cholf`	282.7	(24.9)	256.5	(26.2)	167.0	427.0
	`chold`	−8.5	(10.8)	−33.4	(21.3)	−113.3	29.5
	`comp`	74.5	(21.0)	59.9	(24.4)	0	101.0

Why do we use linear regression in randomized experiments?

- Covariate adjustment can be used to improve efficiency in randomized experiments
  - Always add interaction terms (between each covariate and treatment) to guarantee power improvement

- In completely randomized experiments
  - No need to worry about model misspecification
  - Treatment and covariates are independent