

# Regression III

Ye Wang

University of North Carolina at Chapel Hill

*Linear Methods in Causal Inference*

*POLI784*

# Review

- ▶ The EHW variance estimator is consistent for the true variance even under heteroscedasticity.
- ▶ There are multiple variants of the variance estimator.
- ▶ It enables us to test linear hypothesis.
- ▶ But the confidence interval suffers from the Behrens–Fisher problem.

## Regression and causality

- ▶ We often analyze experimental data with the OLS estimator.
- ▶ Now  $D_i$  is a binary variable.
- ▶ The Neyman-Rubin model does not justify either a linear relationship between  $D_i$  and  $Y_i$  or a constant treatment effect.
- ▶ Then, does it make sense to rely on the OLS estimator?
- ▶ Is  $\hat{\tau}_{OLS}$  consistent for  $\tau_{SATE}$ ?
- ▶ If so, does  $\widehat{Var}[\hat{\tau}_{OLS}]$  quantify the uncertainty of  $\hat{\tau}_{OLS}$  relative to  $\tau_{SATE}$ ?

## Regression and causality

- ▶ Luckily, the answers are yes and yes (Samii and Aronow 2012).
- ▶ Let's use the matrix representation of the OLS estimator:

$$\begin{aligned}\begin{pmatrix} \hat{\mu} \\ \hat{\tau} \end{pmatrix} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}\mathbf{Y}) \\ &= \begin{pmatrix} N & N_1 \\ N_1 & N_1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^N Y_i \\ \sum_{i=1}^N D_i Y_i \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{N_0} & -\frac{1}{N_0} \\ -\frac{1}{N_0} & \frac{N}{N_0 N_1} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^N Y_i \\ \sum_{i=1}^N D_i Y_i \end{pmatrix} \\ &= \begin{pmatrix} \frac{\sum_{i=1}^N (1-D_i) Y_i}{N_0} \\ \frac{\sum_{i=1}^N D_i Y_i}{N_1} - \frac{\sum_{i=1}^N (1-D_i) Y_i}{N_0} \end{pmatrix}.\end{aligned}$$

- ▶ Therefore,  $\hat{\tau}_{OLS} = \hat{\tau}_{HA}$ !

## Regression and causality

- ▶ In general, if the probability of being treated equals  $p_i$  for unit  $i$ , the Hajek estimator has the form

$$\hat{\tau}_{HA} = \frac{\sum_{i=1}^N D_i Y_i / p_i}{\sum_{i=1}^N D_i / p_i} - \frac{\sum_{i=1}^N (1 - D_i) Y_i / (1 - p_i)}{\sum_{i=1}^N (1 - D_i) / (1 - p_i)}$$

- ▶ It is equivalent to the weighted least squares (WLS) estimator based on

$$Y_i = \mu + \tau D_i + \varepsilon_i,$$

with the weight  $W_i = \frac{D_i}{p_i} + \frac{1-D_i}{1-p_i}$ .

- ▶ We can further prove that the Neyman variance estimator is exactly the same as the HC2 variance estimator in regression.
- ▶ We can estimate the ATE in experiments using regression.
- ▶ However, the Behrens–Fisher problem persists!

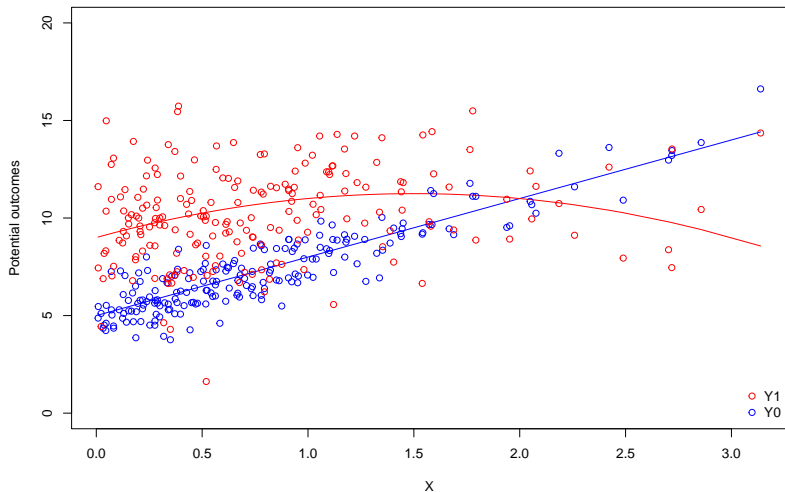
## Regression and causality

- ▶ Under the Bernoulli trial, the Hajek estimator is biased.
- ▶ But the Hajek estimator is identical to the OLS estimator.
- ▶ And we proved that the OLS estimator is unbiased!
- ▶ We can see that

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + \tau_i D_i \\ &= \bar{Y}(0) + \tau D_i + Y_i(0) - \bar{Y}(0) + (\tau_i - \tau) D_i. \end{aligned}$$

- ▶ We have the regression model by setting  $\mu = \bar{Y}(0)$  and  $\varepsilon_i = Y_i(0) - \bar{Y}(0) + (\tau_i - \tau) D_i$ .
- ▶ But in any finite sample,  $Y_i(0)$  and  $\tau_i$  are fixed numbers, hence  $E[Y_i(0) - \bar{Y}(0) + (\tau_i - \tau) D_i | D_i = d] = Y_i(0) - \bar{Y}(0) + (\tau_i - \tau) d \neq 0$ .
- ▶ The assumption holds when we consider an infinite population, where the Hajek estimator is indeed unbiased.

# Regression and causality



## The SATE is 2.913636

## Regression and causality

## The HA estimate is 3.350671

## The OLS estimate is 3.350671

## The HA SE estimate is 0.330464

## The OLS SE estimate is 0.330464

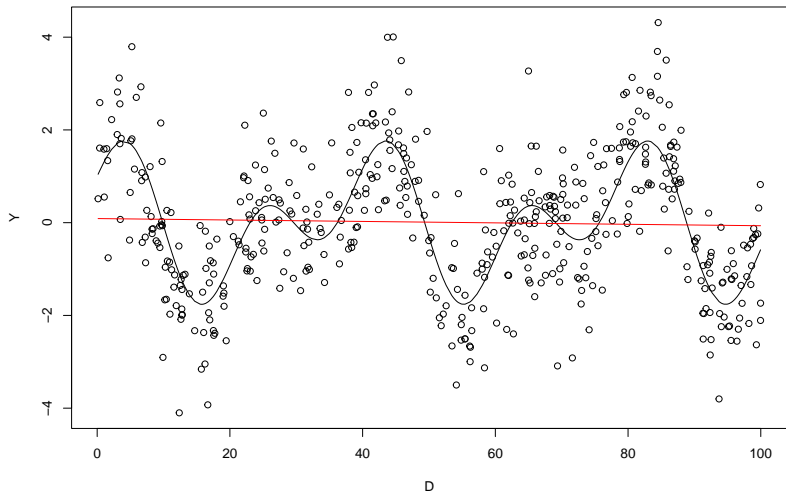
## Regression as projection

- ▶ Suppose  $D_i$  is continuous rather than binary.
- ▶ The effect may vary by the dosage of  $D_i$ .
- ▶ Yet the regression model assumes that the effect is constant.
- ▶ The regression estimate usually does not have a causal interpretation.
- ▶ It can be understood as a projection.
- ▶ The following is always correct due to Taylor expansion:

$$Y_i = \mu + \tau_1 D_i + \tau_2 D_i^2 + \cdots + \tau_k D_i^k + \cdots + \varepsilon_i,$$
$$E[\varepsilon_i | D_i] = 0.$$

- ▶ The OLS estimator now provides the best linear approximation of the conditional expectation  $E[Y_i | D_i]$ .

## Regression as projection



## Covariate adjustment

- ▶ We often want to control for variables other than the treatment in regression and fit the following model:

$$Y_i = \mu + \tau D_i + \mathbf{X}'_i \beta + \varepsilon_i,$$

- ▶ Two reasons: 1. gain efficiency; 2. investigate heterogeneity in treatment effects.
- ▶ The first reason is justified if the outcome is linear in the regressors.
- ▶ We can understand this point through the Frisch-Waugh-Lovell (FWL) theorem.

## Covariate adjustment

- ▶ We can get the OLS estimate  $\hat{\tau}$  via the following algorithm:
  - ▶ Regress  $Y_i$  on  $\mathbf{X}_i$  and save the residuals  $\hat{\epsilon}_{Y_i}$ ;
  - ▶ Regress  $D_i$  on  $\mathbf{X}_i$  and save the residuals  $\hat{\epsilon}_{D_i}$ ;
  - ▶ Regress  $\hat{\epsilon}_{Y_i}$  on  $\hat{\epsilon}_{D_i}$  and save the coefficients.
- ▶ The coefficient for  $\hat{\epsilon}_{D_i}$  will be  $\hat{\tau}$ .
- ▶ Intuitively, we tease out the influence of  $\mathbf{X}_i$  on  $Y_i$  and  $D_i$  separately to isolate the partial effect of  $D_i$  on  $Y_i$ .
- ▶ If the influence of  $\mathbf{X}_i$  on  $Y_i$  and  $D_i$  is linear, then it is guaranteed that controlling  $\mathbf{X}_i$  increases efficiency.
- ▶ Otherwise, we may increase the standard errors of  $\hat{\tau}$  by doing so (Freedman 2008).

## Proof of the FWL theorem

- ▶ Let's write the regression model as

$$\mathbf{Y} = \tau \mathbf{D} + \mathbf{X}\beta + \varepsilon.$$

- ▶ Then, we multiply  $\mathbf{Q} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  to both sides and get

$$\begin{aligned}\mathbf{QY} &= \tau \mathbf{QD} + \mathbf{QX}\beta + \mathbf{QE} \\ &= \tau \mathbf{QD} + \tilde{\varepsilon}.\end{aligned}$$

- ▶ The theorem is proved since  $\mathbf{QY} = \hat{\varepsilon}_Y$  and  $\mathbf{QD} = \hat{\varepsilon}_D$ .

## Covariate adjustment

- ▶ Fortunately, there is a solution proposed by Lin (2013)!
- ▶ We estimate the following model instead:

$$Y_i = \mu + \tau D_i + (\mathbf{X}'_i - \bar{\mathbf{X}})\beta + \delta D_i * (\mathbf{X}'_i - \bar{\mathbf{X}}) + \varepsilon_i,$$

- ▶ It has two features: 1) demeaned covariates, and 2) interaction between the treatment and the covariates.
- ▶ Lin proved that this approach always reduces the standard error of  $\hat{\tau}$ !
- ▶ **Always** use Lin's regression in experimental analysis!

## Covariate adjustment

- ▶ We use the deviation of the covariates to predict that of the outcome.
- ▶ The model is “correct” when  $\mathbf{X}_i = \bar{\mathbf{X}}$ .
- ▶ Bias caused by misspecification becomes local and negligible in variance estimation.
- ▶ Our task: measure the average surface area of a large bag of leaves (e.g., 10000):  $\bar{Y}_P$ .
- ▶ We take a sample of 100 leaves, calculating the sample average:  $\bar{Y}_S = \frac{1}{100} \sum_{i=1}^{100} Y_i$ .
- ▶ The estimate is unbiased and consistent; yet we can do better.
- ▶ The weight of each leaf is correlated with its surface area.
- ▶ Let's measure the weight of each leaf in the sample,  $X_i$ , and the average weight of all leaves in the bag,  $\bar{X}_P$ .
- ▶ A better prediction:  $\bar{Y}_S + (\bar{X}_P - \bar{X}_S) \hat{\beta}$ , where  $\hat{\beta}$  is obtained by regressing  $Y_i$  on  $X_i$ .

## Covariate adjustment

## The OLS estimate is 2.885288

## The naive regression estimate is 2.788625

## The Lin regression estimate is 2.789173

## The OLS SE estimate is 0.3069953

## The naive regression SE estimate is 0.2526677

## The Lin regression SE estimate is 0.244859

## Model-assisted causal inference

- ▶ Lin's regression is a good example of what we call "model-assisted causal inference."
- ▶ Is  $Y_i$  or  $D_i$  linear in  $\mathbf{X}_i$ ? Probably not.
- ▶ Under the conventional perspective, our model is misspecified hence we should be in trouble.
- ▶ Yet we can still use this linear specification to obtain consistent estimates of the target parameter.
- ▶ Causal identification is ensured by the fact that  $D_i$  is randomly assigned.
- ▶ The linear model just assists us to increase the efficiency of the estimator.

## Misconceptions about regression

- ▶ We can say more about the OLS estimator under stricter restrictions on model specification.
- ▶ E.g., the Gauss-Markov theorem: the OLS estimator  $\hat{\beta}$  has the smallest variance among all the linear unbiased estimators for  $\beta$  if

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

$$E[\varepsilon_i | \mathbf{X}_i] = 0,$$

$$E[\varepsilon_i^2 | \mathbf{X}_i] = \sigma^2.$$

- ▶ Therefore, the OLS estimator is known as the best linear unbiased estimator (BLUE).
- ▶ But conditions for the theorem to hold are too strong to be realistic.

## Misconceptions about regression

- ▶ The convention also includes tests for the correct specification, such as homoscedasticity.
- ▶ If you believe in the existence of a true regression model, then heteroscedasticity is concerning (King and Roberts 2015).
- ▶ But in experiments, heteroscedasticity always exists when treatment effects are heterogeneous (Aronow 2016).
- ▶ The existence of a “true model” is probably wishful thinking.
- ▶ There are also methods that help you detect the “correct specification” by examining the fitness ( $R^2$  plus a penalty term) of your model (AIC, BIC, etc.).
- ▶ But we care about the accuracy of estimating  $\tau$  rather than maximizing fitness.
- ▶ Machine learning (ML) algorithms can do a better job in prediction.

## References I

- Aronow, Peter M. 2016. "A Note on" How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do about It"." *arXiv Preprint arXiv:1609.01774*.
- Freedman, David A. 2008. "On Regression Adjustments in Experiments with Several Treatments." *The Annals of Applied Statistics* 2 (1): 176–96.
- King, Gary, and Margaret E Roberts. 2015. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do about It." *Political Analysis* 23 (2): 159–79.
- Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *The Annals of Applied Statistics* 7 (1): 295–318.
- Samii, Cyrus, and Peter M Aronow. 2012. "On Equivalencies Between Design-Based and Regression-Based Variance Estimators for Randomized Experiments." *Statistics & Probability Letters* 82 (2): 365–70.