

# Regression Stuff

EC 425/525, Set 5

Edward Rubin  
24 April 2019

# Prologue

# Schedule

## Last time: Inference and simulation

Let's review using a quote from *MHE*

We've chosen to start with the **asymptotic approach to inference** because modern empirical work typically leans heavily on the large-sample theory that lies behind robust variance formulas. The **payoff is valid inference under weak assumptions**, in particular, a framework that makes sense for our less-than-literal approach to regression models. On the other hand, the **large-sample approach is not without its dangers...**

*MHE*, p. 48 (emphasis added)

# Schedule

## Today

Regression and causality

*Read MHE 3.2*

## Upcoming

Assignment #1

**Advice** Make sure you're taking a few minutes for personal health.<sup>†</sup>

<sup>†</sup> *health* = physical, mental, and spiritual. Also: Do a better job than I do.

# Regression talk

*Saturated models*

# Regression talk

## Saturated models

A **saturated model** is a regression model that includes a discrete (indicator) variable for each set of values the explanatory variables can take.

# Regression talk

## Saturated models

A **saturated model** is a regression model that includes a discrete (indicator) variable for each set of values the explanatory variables can take.

For discrete regressors, saturated models are pretty straightforward.

# Regression talk

## Saturated models

A **saturated model** is a regression model that includes a discrete (indicator) variable for each set of values the explanatory variables can take.

For discrete regressors, saturated models are pretty straightforward.

Example For the relationship between *Wages* and *College Graduation*,

# Regression talk

## Saturated models

A **saturated model** is a regression model that includes a discrete (indicator) variable for each set of values the explanatory variables can take.

For discrete regressors, saturated models are pretty straightforward.

Example For the relationship between **Wages** and **College Graduation**,

$$\text{Wages}_i = \alpha + \beta \mathbb{I}\{\text{College Graduate}\}_i + \varepsilon_i$$

# Regression talk

## Saturated models

A **saturated model** is a regression model that includes a discrete (indicator) variable for each set of values the explanatory variables can take.

For continuous variables, you need an indicator for each potential value.

# Regression talk

## Saturated models

A **saturated model** is a regression model that includes a discrete (indicator) variable for each set of values the explanatory variables can take.

For continuous variables, you need an indicator for each potential value.

Example<sub>2</sub> Regressing Wages on Schooling ( $s_i \in \{0, 1, 2, \dots, T\}$ ).

# Regression talk

## Saturated models

A **saturated model** is a regression model that includes a discrete (indicator) variable for each set of values the explanatory variables can take.

For continuous variables, you need an indicator for each potential value.

Example<sub>2</sub> Regressing Wages on Schooling ( $s_i \in \{0, 1, 2, \dots, T\}$ ).

$$\text{Wages}_i = \alpha + \beta_1 \mathbb{I}\{s_i = 1\}_i + \beta_2 \mathbb{I}\{s_i = 2\}_i + \dots + \beta_T \mathbb{I}\{s_i = T\}_i + \varepsilon_i$$

# Regression talk

## Saturated models

A **saturated model** is a regression model that includes a discrete (indicator) variable for each set of values the explanatory variables can take.

For continuous variables, you need an indicator for each potential value.

Example<sub>2</sub> Regressing Wages on Schooling ( $s_i \in \{0, 1, 2, \dots, T\}$ ).

$$\text{Wages}_i = \alpha + \beta_1 \mathbb{I}\{s_i = 1\}_i + \beta_2 \mathbb{I}\{s_i = 2\}_i + \dots + \beta_T \mathbb{I}\{s_i = T\}_i + \varepsilon_i$$

Here,  $s_i = 0$  is our reference level;  $\beta_j$  is the effect of  $j$  years of schooling.

# Regression talk

## Saturated models

A **saturated model** is a regression model that includes a discrete (indicator) variable for each set of values the explanatory variables can take.

For continuous variables, you need an indicator for each potential value.

Example<sub>2</sub> Regressing Wages on Schooling ( $s_i \in \{0, 1, 2, \dots, T\}$ ).

$$\text{Wages}_i = \alpha + \beta_1 \mathbb{I}\{s_i = 1\}_i + \beta_2 \mathbb{I}\{s_i = 2\}_i + \cdots + \beta_T \mathbb{I}\{s_i = T\}_i + \varepsilon_i$$

Here,  $s_i = 0$  is our reference level;  $\beta_j$  is the effect of  $j$  years of schooling.

$$E[\text{Wages}_i \mid s_i = j] - E[\text{Wages}_i \mid s_i = 0] = \alpha + \beta_j - \alpha = \beta_j$$

# Regression talk

## Saturated models

Q Why focus on saturated models?

# Regression talk

## Saturated models

**Q** Why focus on saturated models?

**A Saturated models perfectly fit the CEF**

# Regression talk

## Saturated models

**Q** Why focus on saturated models?

**A** **Saturated models perfectly fit the CEF** because the CEF is a linear function of the dummy variables—a special case of the linear CEF theorem.

# Regression talk

## Saturated models

If you have multiple explanatory variables, you need **interactions**.

# Regression talk

## Saturated models

If you have multiple explanatory variables, you need **interactions**.

Example<sub>3</sub> Regressing **Wages** on **College Graduation** and **Gender**.

# Regression talk

## Saturated models

If you have multiple explanatory variables, you need **interactions**.

Example<sub>3</sub> Regressing Wages on College Graduation and Gender.

$$\begin{aligned} \text{Wages}_i = & \alpha + \beta_1 \mathbb{I}\{\text{College Graduate}\}_i + \beta_2 \mathbb{I}\{\text{Female}\}_i \\ & + \beta_3 \mathbb{I}\{\text{College Graduate}\}_i \times \mathbb{I}\{\text{Female}\}_i + \varepsilon_i \end{aligned}$$

# Regression talk

## Saturated models

If you have multiple explanatory variables, you need **interactions**.

Example<sub>3</sub> Regressing **Wages** on **College Graduation** and **Gender**.

$$\begin{aligned} \text{Wages}_i = & \alpha + \beta_1 \mathbb{I}\{\text{College Graduate}\}_i + \beta_2 \mathbb{I}\{\text{Female}\}_i \\ & + \beta_3 \mathbb{I}\{\text{College Graduate}\}_i \times \mathbb{I}\{\text{Female}\}_i + \varepsilon_i \end{aligned}$$

Here, the uninteracted terms ( $\beta_1$  &  $\beta_2$ ) are called **main effects**;  $\beta_3$  gives the effect of the **interaction**.

# Regression talk

## Saturated models

If you have multiple explanatory variables, you need **interactions**.

Example<sub>3</sub> Regressing Wages on College Graduation and Gender.

$$\begin{aligned} \text{Wages}_i = & \alpha + \beta_1 \mathbb{I}\{\text{College Graduate}\}_i + \beta_2 \mathbb{I}\{\text{Female}\}_i \\ & + \beta_3 \mathbb{I}\{\text{College Graduate}\}_i \times \mathbb{I}\{\text{Female}\}_i + \varepsilon_i \end{aligned}$$

Here, the uninteracted terms ( $\beta_1$  &  $\beta_2$ ) are called **main effects**;  $\beta_3$  gives the effect of the **interaction**.

$$E[\text{Wages}_i | \text{College Graduate}_i = 0, \text{Female}_i = 0] = \alpha$$

$$E[\text{Wages}_i | \text{College Graduate}_i = 1, \text{Female}_i = 0] = \alpha + \beta_1$$

$$E[\text{Wages}_i | \text{College Graduate}_i = 0, \text{Female}_i = 1] = \alpha + \beta_2$$

$$E[\text{Wages}_i | \text{College Graduate}_i = 1, \text{Female}_i = 1] = \alpha + \beta_1 + \beta_2 + \beta_3$$

# Regression talk

## Saturated models

The CEF can take on four possible values,

$$E[\text{Wages}_i | \text{College Graduate}_i = 0, \text{Female}_i = 0] = \alpha$$

$$E[\text{Wages}_i | \text{College Graduate}_i = 1, \text{Female}_i = 0] = \alpha + \beta_1$$

$$E[\text{Wages}_i | \text{College Graduate}_i = 0, \text{Female}_i = 1] = \alpha + \beta_2$$

$$E[\text{Wages}_i | \text{College Graduate}_i = 1, \text{Female}_i = 1] = \alpha + \beta_1 + \beta_2 + \beta_3$$

and the specification of our saturated regression model

$$\begin{aligned} \text{Wages}_i = & \alpha + \beta_1 \mathbb{I}\{\text{College Graduate}\}_i + \beta_2 \mathbb{I}\{\text{Female}\}_i \\ & + \beta_3 \mathbb{I}\{\text{College Graduate}\}_i \times \mathbb{I}\{\text{Female}\}_i + \varepsilon_i \end{aligned}$$

does not restrict the CEF at all.

# Regression talk

## Model specification

*Saturated models* sit at one extreme of the model-specification spectrum, with *linear, uninteracted models* occupying the opposite extreme.

### Saturated models

- Fit CEF (+)
- Complex (−)
  - Many dummies
  - Many interactions

### Plain, linear models

- Linear approximations (−)
- Simple (+)

# Regression talk

## Model specification

*Saturated models* sit at one extreme of the model-specification spectrum, with *linear, uninteracted models* occupying the opposite extreme.

### Saturated models

- Fit CEF (+)
- Complex (−)
  - Many dummies
  - Many interactions

### Plain, linear models

- Linear approximations (−)
- Simple (+)

Don't forget there are many options in between—though some make less sense than others (e.g., interactions without main effects).

# Regression talk

## Model specification

*Note* Saturated models perfectly fit the CEF regardless of  $\mathbf{Y}_i$ 's distribution.

# Regression talk

## Model specification

*Note* Saturated models perfectly fit the CEF regardless of  $\mathbf{Y}_i$ 's distribution.

Continuous, linear probability, logged, non-negative—it works for all.

Now back to causality...

# Regression and causality

# Regression and causality

## The return of causality

We've spent the last few lectures developing properties/understanding of  
**(1)** the CEF and **(2)** least-squares regression.

Let's return to our main goal of the course...

Q When can we actually interpret a regression as **causal**?<sup>†</sup>

<sup>†</sup> Hint: There is no "reg y x, causal" command in Stata.

# Regression and causality

## The return of causality

We've spent the last few lectures developing properties/understanding of  
**(1)** the CEF and **(2)** least-squares regression.

Let's return to our main goal of the course...

**Q** When can we actually interpret a regression as **causal**?<sup>†</sup>

**A** A regression is causal when the CEF it approximates is causal.

<sup>†</sup> Hint: There is no "reg y x, causal" command in Stata.

# Regression and causality

## The return of causality

Great... thanks.

Q So when is a CEF causal?

# Regression and causality

## The return of causality

Great... thanks.

**Q** So when is a CEF causal?

**A** First, return to the potential-outcomes framework, describing hypothetical outcomes.

A CEF is causal when it describes **differences in average potential outcomes** for a fixed reference population.

*MHE, p. 52 (emphasis added)*

# Regression and causality

## The return of causality

Great... thanks.

**Q** So when is a CEF causal?

**A** First, return to the potential-outcomes framework, describing hypothetical outcomes.

A CEF is causal when it describes **differences in average potential outcomes** for a fixed reference population.

*MHE, p. 52 (emphasis added)*

Let's work through this "definition" of causal CEFs with an example.

# Regression and causality

## Causal CEFs

Example The (causal) effect of schooling on income.

# Regression and causality

## Causal CEFs

Example The (causal) effect of schooling on income.

The causal effect of schooling for individual  $i$  would tell us how  $i$ 's **earnings**  $Y_i$  would change if we varied  $i$ 's **level of schooling**  $s_i$ .

# Regression and causality

## Causal CEFs

Example The (causal) effect of schooling on income.

The causal effect of schooling for individual  $i$  would tell us how  $i$ 's **earnings**  $Y_i$  would change if we varied  $i$ 's **level of schooling**  $s_i$ .

Previously, we discussed how experiments randomly assign treatment to ensure the variable of interest is independent of potential outcomes.

# Regression and causality

## Causal CEFs

Example The (causal) effect of schooling on income.

The causal effect of schooling for individual  $i$  would tell us how  $i$ 's **earnings**  $Y_i$  would change if we varied  $i$ 's **level of schooling**  $s_i$ .

Previously, we discussed how experiments randomly assign treatment to ensure the variable of interest is independent of potential outcomes.

Now we would like to **extend this framework** to

1. variables that take on **more than two values**
2. situations that require us to **hold many covariates constant** in order to achieve a valid causal interpretation

# Regression and causality

## Causal CEFs

The idea of *holding (many) covariates constant* brings us to one of the cornerstones of applied econometrics: the **conditional independence assumption (CIA)** (also called *selection on observables*).

# Regression and causality

## The conditional independence assumption

*Definition(s)*

- Conditional on some set of covariates  $\mathbf{X}_i$ , selection bias disappears.

# Regression and causality

## The conditional independence assumption

*Definition(s)*

- Conditional on some set of covariates  $\mathbf{X}_i$ , selection bias disappears.
- Conditional on  $\mathbf{X}_i$ , potential outcomes ( $\mathbf{Y}_{0i}$ ,  $\mathbf{Y}_{1i}$ ) are independent of treatment status ( $\mathbf{D}_i$ ).

$$\{\mathbf{Y}_{0i}, \mathbf{Y}_{1i}\} \perp\!\!\!\perp \mathbf{D}_i | \mathbf{X}_i$$

# Regression and causality

## The conditional independence assumption

*Definition(s)*

- Conditional on some set of covariates  $\mathbf{X}_i$ , selection bias disappears.
- Conditional on  $\mathbf{X}_i$ , potential outcomes ( $\mathbf{Y}_{0i}$ ,  $\mathbf{Y}_{1i}$ ) are independent of treatment status ( $\mathbf{D}_i$ ).

$$\{\mathbf{Y}_{0i}, \mathbf{Y}_{1i}\} \perp\!\!\!\perp \mathbf{D}_i | \mathbf{X}_i$$

To see how CIA eliminates selection bias...

$$\text{Selection bias} = E[\mathbf{Y}_{0i} | \mathbf{X}_i, \mathbf{D}_i = 1] - E[\mathbf{Y}_{0i} | \mathbf{X}_i, \mathbf{D}_i = 0]$$

# Regression and causality

## The conditional independence assumption

*Definition(s)*

- Conditional on some set of covariates  $\mathbf{X}_i$ , selection bias disappears.
- Conditional on  $\mathbf{X}_i$ , potential outcomes ( $\mathbf{Y}_{0i}$ ,  $\mathbf{Y}_{1i}$ ) are independent of treatment status ( $\mathbf{D}_i$ ).

$$\{\mathbf{Y}_{0i}, \mathbf{Y}_{1i}\} \perp\!\!\!\perp \mathbf{D}_i | \mathbf{X}_i$$

To see how CIA eliminates selection bias...

$$\begin{aligned}\text{Selection bias} &= E[\mathbf{Y}_{0i} | \mathbf{X}_i, \mathbf{D}_i = 1] - E[\mathbf{Y}_{0i} | \mathbf{X}_i, \mathbf{D}_i = 0] \\ &= E[\mathbf{Y}_{0i} | \mathbf{X}_i] - E[\mathbf{Y}_{0i} | \mathbf{X}_i]\end{aligned}$$

# Regression and causality

## The conditional independence assumption

*Definition(s)*

- Conditional on some set of covariates  $\mathbf{X}_i$ , selection bias disappears.
- Conditional on  $\mathbf{X}_i$ , potential outcomes ( $\mathbf{Y}_{0i}$ ,  $\mathbf{Y}_{1i}$ ) are independent of treatment status ( $\mathbf{D}_i$ ).

$$\{\mathbf{Y}_{0i}, \mathbf{Y}_{1i}\} \perp\!\!\!\perp \mathbf{D}_i | \mathbf{X}_i$$

To see how CIA eliminates selection bias...

$$\begin{aligned}\text{Selection bias} &= E[\mathbf{Y}_{0i} | \mathbf{X}_i, \mathbf{D}_i = 1] - E[\mathbf{Y}_{0i} | \mathbf{X}_i, \mathbf{D}_i = 0] \\ &= E[\mathbf{Y}_{0i} | \mathbf{X}_i] - E[\mathbf{Y}_{0i} | \mathbf{X}_i] \\ &= 0\end{aligned}$$

# Regression and causality

## The conditional independence assumption

Another way you'll hear CIA: After controlling for some set of variables  $\mathbf{X}_i$ , treatment assignment is ***as good as random.***

# Regression and causality

## The conditional independence assumption

Another way you'll hear CIA: After controlling for some set of variables  $\mathbf{X}_i$ , treatment assignment is ***as good as random.***

To see how this assumption<sup>†</sup> buys us a causal interpretation, write out our old difference in means—but now condition on  $\mathbf{X}_i$ .

<sup>†</sup> Another way to think about econometric assumptions is as requirements.

# Regression and causality

## The conditional independence assumption

Another way you'll hear CIA: After controlling for some set of variables  $\mathbf{X}_i$ , treatment assignment is ***as good as random.***

To see how this assumption<sup>†</sup> buys us a causal interpretation, write out our old difference in means—but now condition on  $\mathbf{X}_i$ .

$$\begin{aligned} & E[\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{D}_i = 1] - E[\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{D}_i = 0] \\ &= E[\mathbf{Y}_{1i} \mid \mathbf{X}_i] - E[\mathbf{Y}_{0i} \mid \mathbf{X}_i] \\ &= E[\mathbf{Y}_{1i} - \mathbf{Y}_{0i} \mid \mathbf{X}_i] \end{aligned}$$

<sup>†</sup> Another way to think about econometric assumptions is as requirements.

# Regression and causality

## The conditional independence assumption

Another way you'll hear CIA: After controlling for some set of variables  $\mathbf{X}_i$ , treatment assignment is ***as good as random.***

To see how this assumption<sup>†</sup> buys us a causal interpretation, write out our old difference in means—but now condition on  $\mathbf{X}_i$ .

$$\begin{aligned} & E[\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{D}_i = 1] - E[\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{D}_i = 0] \\ &= E[\mathbf{Y}_{1i} \mid \mathbf{X}_i] - E[\mathbf{Y}_{0i} \mid \mathbf{X}_i] \\ &= E[\mathbf{Y}_{1i} - \mathbf{Y}_{0i} \mid \mathbf{X}_i] \end{aligned}$$

Even randomized experiments need the CIA—e.g., the STAR experiment's *within-school* randomization.

<sup>†</sup> Another way to think about econometric assumptions is as requirements.

# Regression and causality

## The conditional independence assumption

Now let's extend this framework to **multi-valued explanatory variables**.

# Regression and causality

## The conditional independence assumption

Now let's extend this framework to **multi-valued explanatory variables**.

Example continued Schooling ( $s_i$ ) takes on integers  $\in \{0, 1, \dots, T\}$ .

We want to know the effect of an individual's **schooling** on her **wages** ( $Y_i$ ).

# Regression and causality

## The conditional independence assumption

Now let's extend this framework to **multi-valued explanatory variables**.

Example continued Schooling ( $s_i$ ) takes on integers  $\in \{0, 1, \dots, T\}$ .

We want to know the effect of an individual's schooling on her wages ( $Y_i$ ).

Previously,  $Y_{1i}$  denoted individual  $i$ 's outcome under treatment.

Now,  $Y_{si}$  denotes individual  $i$ 's outcome with  $s$  years of schooling.

# Regression and causality

## The conditional independence assumption

Now let's extend this framework to **multi-valued explanatory variables**.

Example continued Schooling ( $s_i$ ) takes on integers  $\in \{0, 1, \dots, T\}$ .

We want to know the effect of an individual's **schooling** on her **wages** ( $Y_i$ ).

Previously,  $Y_{1i}$  denoted individual  $i$ 's outcome under treatment.

Now,  $Y_{si}$  denotes individual  $i$ 's outcome **with  $s$  years of schooling**.

Let each individual have her own function between **schooling** and **earnings**.

$$Y_{si} \equiv f_i(s)$$

$f_i(s)$  answers exactly the type of causal questions that we want to answer.

# Regression and causality

## The conditional independence assumption

Extending the CIA to this multi-valued setting...

$$Y_{si} \perp\!\!\!\perp s_i \mid X_i \text{ for all } s$$

# Regression and causality

## The conditional independence assumption

Extending the CIA to this multi-valued setting...

$$Y_{si} \perp\!\!\!\perp s_i \mid X_i \text{ for all } s$$

If we apply the CIA to  $Y_{si} \equiv f_i(s)$ , we define the *average causal effect* of a one-year increase in **schooling** as

$$E[f_i(s) - f_i(s-1) \mid X_i]$$

# Regression and causality

## The conditional independence assumption

Extending the CIA to this multi-valued setting...

$$\mathbf{Y}_{si} \perp\!\!\!\perp s_i \mid \mathbf{X}_i \text{ for all } s$$

If we apply the CIA to  $\mathbf{Y}_{si} \equiv f_i(s)$ , we define the *average causal effect* of a one-year increase in **schooling** as

$$E[f_i(s) - f_i(s-1) \mid \mathbf{X}_i]$$

However, the data only contain one realization of  $f_i(s)$  per  $i$ —we only see  $f_i(s)$  evaluated at exactly one value of  $s$  per  $i$ , i.e.,  $\mathbf{Y}_i = f_i(s_i)$ .

# Regression and causality

## The conditional independence assumption

Extending the CIA to this multi-valued setting...

$$\mathbf{Y}_{si} \perp\!\!\!\perp s_i \mid \mathbf{X}_i \text{ for all } s$$

If we apply the CIA to  $\mathbf{Y}_{si} \equiv f_i(s)$ , we define the *average causal effect* of a one-year increase in **schooling** as

$$E[f_i(s) - f_i(s-1) \mid \mathbf{X}_i]$$

However, the data only contain one realization of  $f_i(s)$  per  $i$ —we only see  $f_i(s)$  evaluated at exactly one value of  $s$  per  $i$ , i.e.,  $\mathbf{Y}_i = f_i(s_i)$ .

The CIA to the rescue!

# Regression and causality

## The conditional independence assumption

Extending the CIA to this multi-valued setting...

$$Y_{si} \perp\!\!\!\perp s_i \mid X_i \text{ for all } s$$

If we apply the CIA to  $Y_{si} \equiv f_i(s)$ , we define the *average causal effect* of a one-year increase in **schooling** as

$$E[f_i(s) - f_i(s-1) \mid X_i]$$

However, the data only contain one realization of  $f_i(s)$  per  $i$ —we only see  $f_i(s)$  evaluated at exactly one value of  $s$  per  $i$ , i.e.,  $Y_i = f_i(s_i)$ .

The CIA to the rescue! Conditional on  $X_i$ ,  $Y_{si}$  and  $s_i$  are independent.

# Regression and causality

## The conditional independence assumption

The CIA to the rescue! Conditional on  $\mathbf{X}_i$ ,  $\mathbf{Y}_{si}$  and  $s_i$  are independent.

$$\begin{aligned} & E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = s] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = s - 1] \\ &= E[\mathbf{Y}_{si} \mid \mathbf{X}_i, s_i = s] - E[\mathbf{Y}_{(s-1)i} \mid \mathbf{X}_i, s_i = s - 1] \\ &= E[\mathbf{Y}_{si} \mid \mathbf{X}_i] - E[\mathbf{Y}_{(s-1)i} \mid \mathbf{X}_i] \\ &= E[\mathbf{Y}_{si} - \mathbf{Y}_{(s-1)i} \mid \mathbf{X}_i] \\ &= E[f_i(s) - f_i(s - 1) \mid \mathbf{X}_i] \end{aligned}$$

# Regression and causality

## The conditional independence assumption

The CIA to the rescue! Conditional on  $\mathbf{X}_i$ ,  $\mathbf{Y}_{si}$  and  $s_i$  are independent.

$$\begin{aligned} & E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = s] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = s - 1] \\ &= E[\mathbf{Y}_{si} \mid \mathbf{X}_i, s_i = s] - E[\mathbf{Y}_{(s-1)i} \mid \mathbf{X}_i, s_i = s - 1] \\ &= E[\mathbf{Y}_{si} \mid \mathbf{X}_i] - E[\mathbf{Y}_{(s-1)i} \mid \mathbf{X}_i] \\ &= E[\mathbf{Y}_{si} - \mathbf{Y}_{(s-1)i} \mid \mathbf{X}_i] \\ &= E[f_i(s) - f_i(s - 1) \mid \mathbf{X}_i] \end{aligned}$$

With the CIA, a difference in conditional averages allows causal interpretations.

# Regression and causality

## The conditional independence assumption

Example The causal effect of high-school graduation is

# Regression and causality

## The conditional independence assumption

Example The causal effect of high-school graduation is

$$E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 12] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 11]$$

# Regression and causality

## The conditional independence assumption

Example The causal effect of high-school graduation is

$$\begin{aligned} & E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 12] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 11] \\ &= E[f_i(12) \mid \mathbf{X}_i, s_i = 12] - E[f_i(11) \mid \mathbf{X}_i, s_i = 11] \end{aligned}$$

# Regression and causality

## The conditional independence assumption

Example The causal effect of high-school graduation is

$$\begin{aligned} & E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 12] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 11] \\ &= E[f_i(12) \mid \mathbf{X}_i, s_i = 12] - E[f_i(11) \mid \mathbf{X}_i, s_i = 11] \\ &= E[f_i(12) \mid \mathbf{X}_i, s_i = 12] - E[f_i(11) \mid \mathbf{X}_i, s_i = 12] \quad (\text{from CIA}) \end{aligned}$$

# Regression and causality

## The conditional independence assumption

Example The causal effect of high-school graduation is

$$\begin{aligned} & E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 12] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 11] \\ &= E[f_i(12) \mid \mathbf{X}_i, s_i = 12] - E[f_i(11) \mid \mathbf{X}_i, s_i = 11] \\ &= E[f_i(12) \mid \mathbf{X}_i, s_i = 12] - E[f_i(11) \mid \mathbf{X}_i, s_i = 12] \quad (\text{from CIA}) \\ &= E[f_i(12) - f_i(11) \mid \mathbf{X}_i, s_i = 12] \end{aligned}$$

# Regression and causality

## The conditional independence assumption

Example The causal effect of high-school graduation is

$$\begin{aligned} & E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 12] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 11] \\ &= E[f_i(12) \mid \mathbf{X}_i, s_i = 12] - E[f_i(11) \mid \mathbf{X}_i, s_i = 11] \\ &= E[f_i(12) \mid \mathbf{X}_i, s_i = 12] - E[f_i(11) \mid \mathbf{X}_i, s_i = 12] \quad (\text{from CIA}) \\ &= E[f_i(12) - f_i(11) \mid \mathbf{X}_i, s_i = 12] \\ &= \text{The average causal effect of graduation for graduates} \end{aligned}$$

# Regression and causality

## The conditional independence assumption

Example The causal effect of high-school graduation is

$$\begin{aligned} & E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 12] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 11] \\ &= E[f_i(12) \mid \mathbf{X}_i, s_i = 12] - E[f_i(11) \mid \mathbf{X}_i, s_i = 11] \\ &= E[f_i(12) \mid \mathbf{X}_i, s_i = 12] - E[f_i(11) \mid \mathbf{X}_i, s_i = 12] \quad (\text{from CIA}) \\ &= E[f_i(12) - f_i(11) \mid \mathbf{X}_i, s_i = 12] \\ &= \text{The average causal effect of graduation for graduates} \\ &= E[f_i(12) - f_i(11) \mid \mathbf{X}_i] \quad (\text{CIA again}) \end{aligned}$$

# Regression and causality

## The conditional independence assumption

Example The causal effect of high-school graduation is

$$\begin{aligned} & E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 12] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 11] \\ &= E[f_i(12) \mid \mathbf{X}_i, s_i = 12] - E[f_i(11) \mid \mathbf{X}_i, s_i = 11] \\ &= E[f_i(12) \mid \mathbf{X}_i, s_i = 12] - E[f_i(11) \mid \mathbf{X}_i, s_i = 12] \quad (\text{from CIA}) \\ &= E[f_i(12) - f_i(11) \mid \mathbf{X}_i, s_i = 12] \\ &= \text{The average causal effect of graduation for graduates} \\ &= E[f_i(12) - f_i(11) \mid \mathbf{X}_i] \quad (\text{CIA again}) \\ &= \text{The (conditional) average causal effect of graduation at } X_i \end{aligned}$$

# Regression and causality

## The conditional independence assumption

Q What about the **unconditional** average causal effect of graduation?

# Regression and causality

## The conditional independence assumption

**Q** What about the **unconditional** average causal effect of graduation?

**A** First, remember what we just showed...

# Regression and causality

## The conditional independence assumption

**Q** What about the **unconditional** average causal effect of graduation?

**A** First, remember what we just showed...

$$E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 12] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 11] = E[f_i(12) - f_i(11) \mid \mathbf{X}_i]$$

# Regression and causality

## The conditional independence assumption

**Q** What about the **unconditional** average causal effect of graduation?

**A** First, remember what we just showed...

$$E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 12] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 11] = E[f_i(12) - f_i(11) \mid \mathbf{X}_i]$$

Now take the expected value of both sides and apply the LIE.

$$\begin{aligned} & E\left( E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 12] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 11] \right) \\ &= E\left( E[f_i(12) - f_i(11) \mid \mathbf{X}_i] \right) \end{aligned}$$

# Regression and causality

## The conditional independence assumption

**Q** What about the **unconditional** average causal effect of graduation?

**A** First, remember what we just showed...

$$E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 12] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 11] = E[f_i(12) - f_i(11) \mid \mathbf{X}_i]$$

Now take the expected value of both sides and apply the LIE.

$$\begin{aligned} & E\left(E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 12] - E[\mathbf{Y}_i \mid \mathbf{X}_i, s_i = 11]\right) \\ &= E\left(E[f_i(12) - f_i(11) \mid \mathbf{X}_i]\right) \\ &= E[f_i(12) - f_i(11)] \quad (\text{Iterating expectations}) \end{aligned}$$

# Regression and causality

## The conditional independence assumption

### *Takeaways*

1. Conditional independence gives our parameters **causal interpretations** (eliminating selection bias).

# Regression and causality

## The conditional independence assumption

### *Takeaways*

1. Conditional independence gives our parameters **causal interpretations** (eliminating selection bias).
2. The interpretation changes slightly—without iterating expectations, we have **conditional average treatment effects**.

# Regression and causality

## The conditional independence assumption

### Takeaways

1. Conditional independence gives our parameters **causal interpretations** (eliminating selection bias).
2. The interpretation changes slightly—without iterating expectations, we have **conditional average treatment effects**.
3. The CIA is challenging—you need to know which set of covariates ( $X_i$ ) leads to **as-good-as-random residual variation in your treatment**.

# Regression and causality

## The conditional independence assumption

### Takeaways

1. Conditional independence gives our parameters **causal interpretations** (eliminating selection bias).
2. The interpretation changes slightly—without iterating expectations, we have **conditional average treatment effects**.
3. The CIA is challenging—you need to know which set of covariates ( $X_i$ ) leads to **as-good-as-random residual variation in your treatment**.
4. The idea of conditioning on observables to match *comparable* individuals introduces us to **matching estimators**—comparing groups of individuals with the same covariate values.

# Regression and causality

## From the CIA to regression

Conditional independence fits into our regression framework in two ways.

# Regression and causality

## From the CIA to regression

Conditional independence fits into our regression framework in two ways.

1. If we assume  $f_i(\textcolor{red}{s})$  is **(A)** linear in  $\textcolor{red}{s}$  and **(B)** equal across all individuals except for an additive error, linear regression estimates  $f(\textcolor{red}{s})$ .

# Regression and causality

## From the CIA to regression

Conditional independence fits into our regression framework in two ways.

1. If we assume  $f_i(\mathbf{s})$  is **(A)** linear in  $\mathbf{s}$  and **(B)** equal across all individuals except for an additive error, linear regression estimates  $f(\mathbf{s})$ .
2. If we allow  $f_i(\mathbf{s})$  to be nonlinear in  $\mathbf{s}$  and heterogeneous across  $i$ , regression provides a weighted average of individual-specific differences  $f_i(\mathbf{s}) - f_i(\mathbf{s} - \mathbf{1})$ .<sup>†</sup>

<sup>†</sup> Leads to a matching-style estimator.

# Regression and causality

## From the CIA to regression

Conditional independence fits into our regression framework in two ways.

1. If we assume  $f_i(\textcolor{red}{s})$  is **(A)** linear in  $\textcolor{red}{s}$  and **(B)** equal across all individuals except for an additive error, linear regression estimates  $f(\textcolor{red}{s})$ .
2. If we allow  $f_i(\textcolor{red}{s})$  to be nonlinear in  $\textcolor{red}{s}$  and heterogeneous across  $i$ , regression provides a weighted average of individual-specific differences  $f_i(\textcolor{red}{s}) - f_i(\textcolor{red}{s} - 1)$ .<sup>†</sup>

Let's start with the 'easier' case: a linear, constant-effects (causal) model.

<sup>†</sup> Leads to a matching-style estimator.

# Regression and causality

## From the CIA to regression

Let  $f_i(\textcolor{red}{s})$  be linear in  $\textcolor{red}{s}$  and equal across  $i$  except for an error term, e.g.,

$$f_i(\textcolor{red}{s}) = \alpha + \rho \textcolor{red}{s} + \eta_i \quad (\text{A})$$

# Regression and causality

## From the CIA to regression

Let  $f_i(\textcolor{red}{s})$  be linear in  $\textcolor{red}{s}$  and equal across  $i$  except for an error term, e.g.,

$$f_i(\textcolor{red}{s}) = \alpha + \rho \textcolor{red}{s} + \eta_i \tag{A}$$

Substitute in our observed value of  $\textcolor{red}{s}_i$  and the outcome  $\textcolor{blue}{Y}_i$

$$\textcolor{blue}{Y}_i = \alpha + \rho \textcolor{red}{s}_i + \eta_i \tag{B}$$

# Regression and causality

## From the CIA to regression

Let  $f_i(\textcolor{red}{s})$  be linear in  $\textcolor{red}{s}$  and equal across  $i$  except for an error term, e.g.,

$$f_i(\textcolor{red}{s}) = \alpha + \rho \textcolor{red}{s} + \eta_i \tag{A}$$

Substitute in our observed value of  $\textcolor{red}{s}_i$  and the outcome  $\textcolor{blue}{Y}_i$

$$\textcolor{blue}{Y}_i = \alpha + \rho \textcolor{red}{s}_i + \eta_i \tag{B}$$

While  $\rho$  in (A) is explicitly causal, regression-based estimates of  $\rho$  in (B) need not be causal (selection/OVB for endogenous  $s_i$ ).

# Regression and causality

## From the CIA to regression

Continuing with our linear, constant-effect causal model...

$$f_i(\textcolor{red}{s}) = \alpha + \rho \textcolor{red}{s} + \eta_i \quad (\text{A})$$

Now impose the conditional independence assumption for covariates  $\mathbf{X}_i$ .

$$\eta_i = \mathbf{X}'_i \gamma + \nu_i \quad (\text{C})$$

where  $\gamma$  is a vector of population coefficients from regressing  $\eta_i$  on  $\mathbf{X}_i$ .

# Regression and causality

## From the CIA to regression

Continuing with our linear, constant-effect causal model...

$$f_i(\textcolor{red}{s}) = \alpha + \rho \textcolor{red}{s} + \eta_i \quad (\text{A})$$

Now impose the conditional independence assumption for covariates  $\mathbf{X}_i$ .

$$\eta_i = \mathbf{X}'_i \gamma + \nu_i \quad (\text{C})$$

where  $\gamma$  is a vector of population coefficients from regressing  $\eta_i$  on  $\mathbf{X}_i$ .

Note Least-squares regression implies

1.  $E[\eta_i | \mathbf{X}_i] = \mathbf{X}'_i \gamma$
2.  $\mathbf{X}_i$  is uncorrelated with  $\nu_i$ .

# Regression and causality

## From the CIA to regression

Now write out the conditional expectation function of  $f_i(\textcolor{red}{s})$  on  $\mathbf{X}_i$  and  $\textcolor{red}{s}_i$ .

$$\begin{aligned} & E[f_i(\textcolor{red}{s}) \mid \mathbf{X}_i, \textcolor{red}{s}_i] \\ &= E[f_i(\textcolor{red}{s}) \mid \mathbf{X}_i] \quad (\text{CIA}) \end{aligned}$$

# Regression and causality

## From the CIA to regression

Now write out the conditional expectation function of  $f_i(\textcolor{red}{s})$  on  $\mathbf{X}_i$  and  $\textcolor{red}{s}_i$ .

$$\begin{aligned} & E[f_i(\textcolor{red}{s}) \mid \mathbf{X}_i, \textcolor{red}{s}_i] \\ &= E[f_i(\textcolor{red}{s}) \mid \mathbf{X}_i] \quad (\text{CIA}) \\ &= E[\alpha + \rho \textcolor{red}{s}_i + \eta_i \mid \mathbf{X}_i] \end{aligned}$$

# Regression and causality

## From the CIA to regression

Now write out the conditional expectation function of  $f_i(\textcolor{red}{s})$  on  $\mathbf{X}_i$  and  $\textcolor{red}{s}_i$ .

$$\begin{aligned} & E[f_i(\textcolor{red}{s}) \mid \mathbf{X}_i, \textcolor{red}{s}_i] \\ &= E[f_i(\textcolor{red}{s}) \mid \mathbf{X}_i] \quad (\text{CIA}) \\ &= E[\alpha + \rho \textcolor{red}{s}_i + \eta_i \mid \mathbf{X}_i] \\ &= \alpha + \rho \textcolor{red}{s}_i + E[\eta_i \mid \mathbf{X}_i] \end{aligned}$$

# Regression and causality

## From the CIA to regression

Now write out the conditional expectation function of  $f_i(\textcolor{red}{s})$  on  $\mathbf{X}_i$  and  $\textcolor{red}{s}_i$ .

$$E[f_i(\textcolor{red}{s}) \mid \mathbf{X}_i, \textcolor{red}{s}_i]$$

$$= E[f_i(\textcolor{red}{s}) \mid \mathbf{X}_i] \quad (\text{CIA})$$

$$= E[\alpha + \rho \textcolor{red}{s}_i + \eta_i \mid \mathbf{X}_i]$$

$$= \alpha + \rho \textcolor{red}{s}_i + E[\eta_i \mid \mathbf{X}_i]$$

$$= \alpha + \rho \textcolor{red}{s}_i + \mathbf{X}'_i \gamma \quad (\text{Least-squares regression})$$

# Regression and causality

## From the CIA to regression

Now write out the conditional expectation function of  $f_i(\textcolor{red}{s})$  on  $\mathbf{X}_i$  and  $\textcolor{red}{s}_i$ .

$$\begin{aligned} & E[f_i(\textcolor{red}{s}) \mid \mathbf{X}_i, \textcolor{red}{s}_i] \\ &= E[f_i(\textcolor{red}{s}) \mid \mathbf{X}_i] \quad (\text{CIA}) \\ &= E[\alpha + \rho \textcolor{red}{s}_i + \eta_i \mid \mathbf{X}_i] \\ &= \alpha + \rho \textcolor{red}{s}_i + E[\eta_i \mid \mathbf{X}_i] \\ &= \alpha + \rho \textcolor{red}{s}_i + \mathbf{X}'_i \gamma \quad (\text{Least-squares regression}) \end{aligned}$$

The CEF of  $f_i(\textcolor{red}{s}_i)$  is linear, which means that the (right<sup>†</sup>) population regression will be the CEF.

<sup>†</sup> Here, "right" means conditional on  $\mathbf{X}_i$ .

# Regression and causality

## From the CIA to regression

Thus, the linear causal (regression) model is

$$Y_i = \alpha + \rho s_i + X'_i \gamma + \nu_i$$

The residual  $\nu_i$  is uncorrelated with

1.  $s_i$  (from the CIA)
2.  $X_i$  (from defining  $\gamma$  via the regression of  $\eta$  on  $X_i$ )

The coefficient  $\rho$  gives the causal effect of  $s_i$  on  $Y_i$ .

# Regression and causality

## From the CIA to regression

As Angrist and Pischke note, this **conditional-independence assumption** (a.k.a. the selection-on-observables assumption) is the cornerstone of modern empirical work in economics—and many other disciplines.

Nearly any empirical application that wants a causal interpretation involves a (sometimes implicit) argument that **conditional on some set of covariates, treatment is as-good-as random.**

# Regression and causality

## From the CIA to regression

As Angrist and Pischke note, this **conditional-independence assumption** (a.k.a. the selection-on-observables assumption) is the cornerstone of modern empirical work in economics—and many other disciplines.

Nearly any empirical application that wants a causal interpretation involves a (sometimes implicit) argument that **conditional on some set of covariates, treatment is as-good-as random.**

Part of our job: Reasoning through the validity of this assumption.

# Regression and causality

## CIA example

Let's continue with the returns to graduation ( $G_i$ ).

Let's imagine

1. Women are more likely to graduate.
2. Everyone receives the same return to graduation.
3. Women receive lower wages across the board.

# Regression and causality

## CIA example

First, we need to generate some data.

```
# Set seed
set.seed(12345)
# Set sample size
n ← 1e4
# Generate data
ex_df ← tibble(
  female = rep(c(0, 1), each = n/2),
  grad = runif(n, min = female/3, max = 1) %>% round(0),
  wage = 100 - 25 * female + 5 * grad + rnorm(n, sd = 3)
)
```

# Regression and causality

## CIA example

Now we can estimate our naïve regression

$$\text{Wage}_i = \alpha + \beta \text{Grad}_i + \varepsilon_i$$

# Regression and causality

## CIA example

Now we can estimate our naïve regression

$$\text{Wage}_i = \alpha + \beta \text{Grad}_i + \varepsilon_i$$

```
lm(wage ~ grad, data = ex_df)
```

	<b>Coef.</b>	<b>S.E.</b>	<b>t stat</b>
Intercept	91.65	0.20	447.70
Graduate	-1.59	0.26	-6.18

# Regression and causality

## CIA example

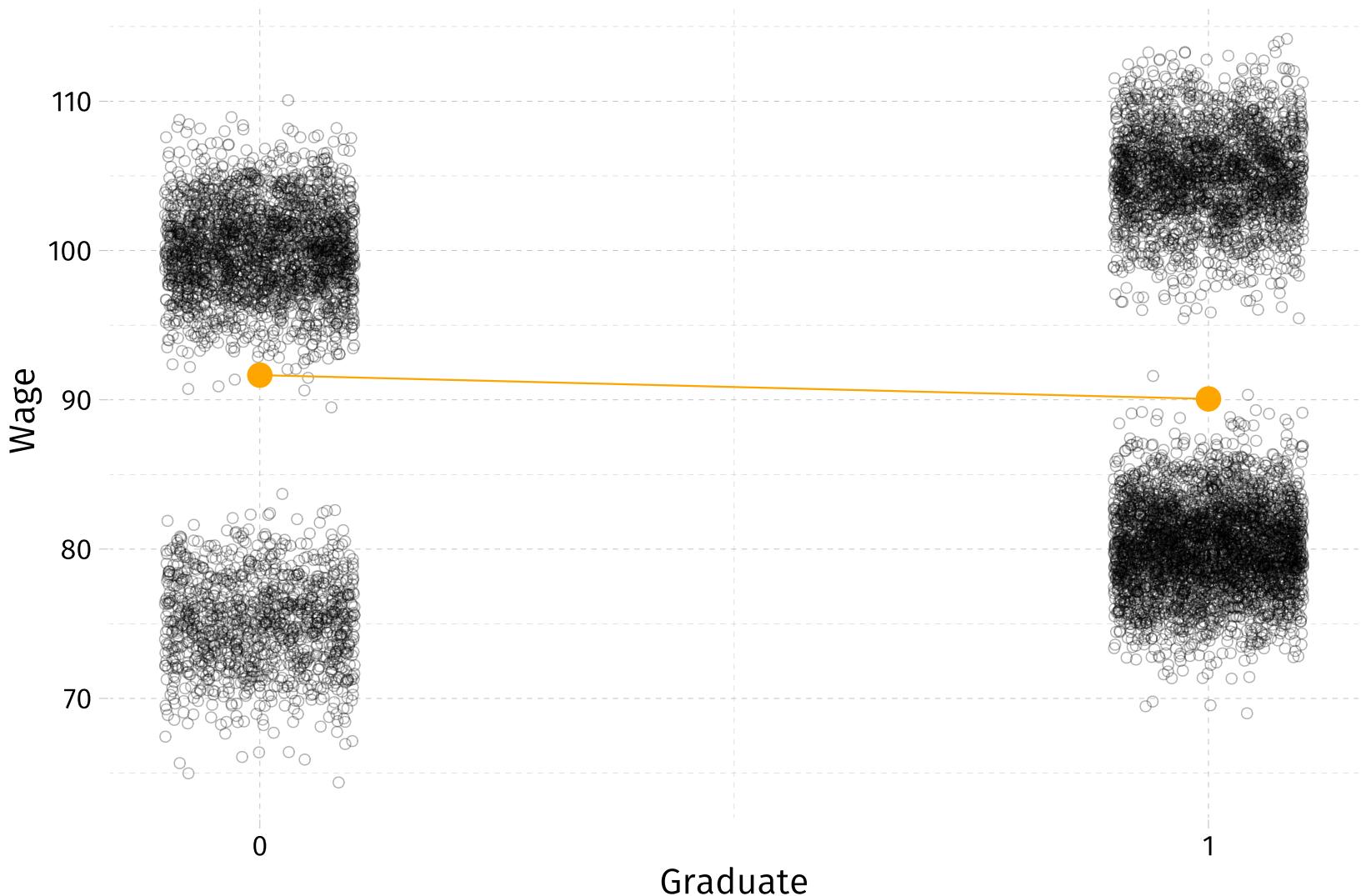
Now we can estimate our naïve regression

$$\text{Wage}_i = \alpha + \beta \text{Grad}_i + \varepsilon_i$$

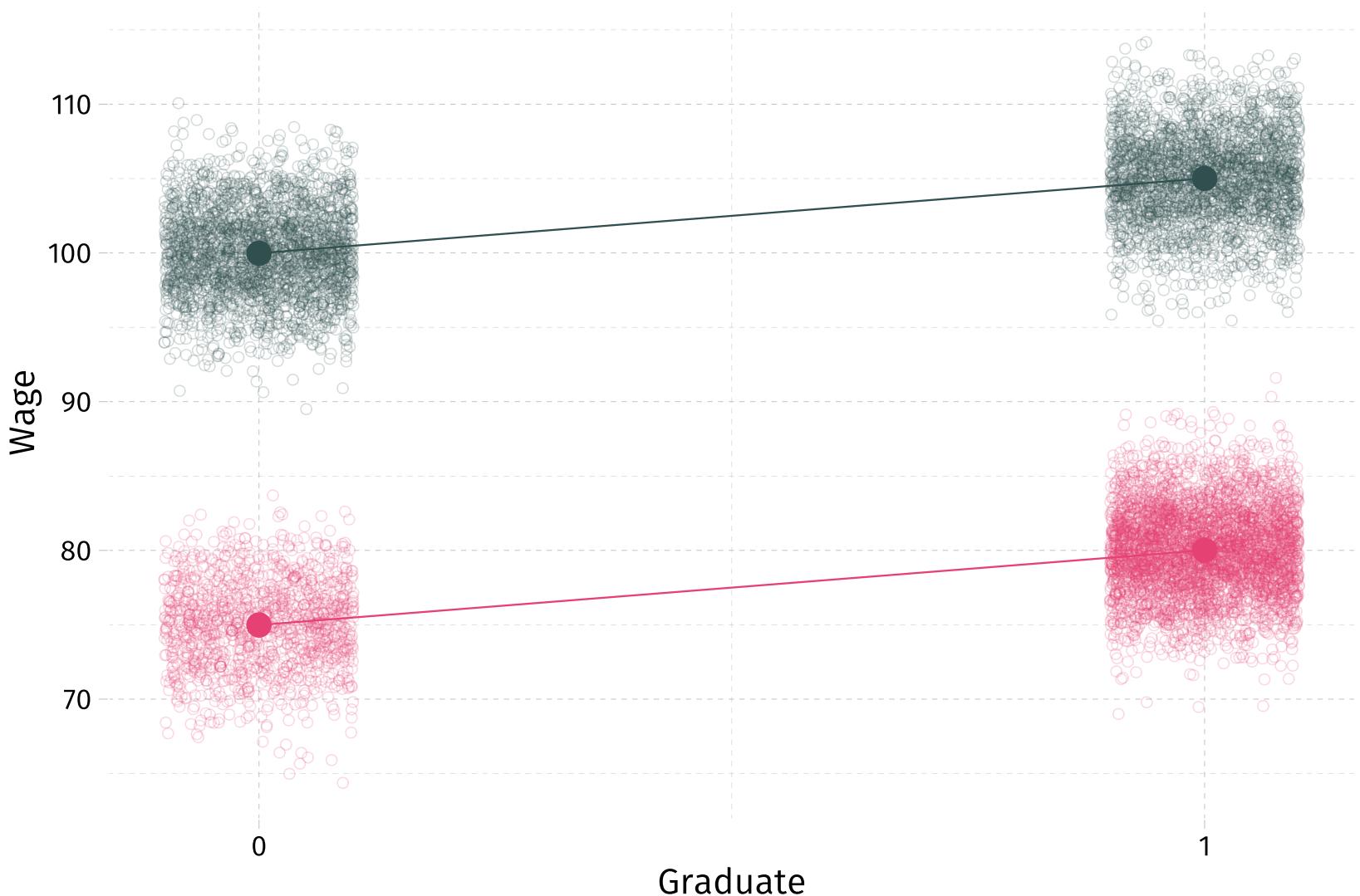
```
lm(wage ~ grad, data = ex_df)
```

	<b>Coef.</b>	<b>S.E.</b>	<b>t stat</b>
Intercept	91.65	0.20	447.70
Graduate	-1.59	0.26	-6.18

Maybe we should have plotted our data...



We're still missing something...



# Regression and causality

## CIA example

Now we can estimate our causal regression

$$\text{Wage}_i = \alpha + \beta_1 \text{Grad}_i + \beta_2 \text{Female}_i + \varepsilon_i$$

# Regression and causality

## CIA example

Now we can estimate our causal regression

$$\text{Wage}_i = \alpha + \beta_1 \text{Grad}_i + \beta_2 \text{Female}_i + \varepsilon_i$$

```
lm(wage ~ grad + female, data = ex_df)
```

	<b>Coef.</b>	<b>S.E.</b>	<b>t stat</b>
Intercept	99.98	0.05	1868.81
Graduate	5.03	0.06	78.23
Female	-25.00	0.06	-402.64

# Table of contents

## Admin

1. Last time
2. Schedule
3. Advice

## Regression

1. Saturated models
2. Model specification
3. Causal regressions
4. Causal CEFs
5. Conditional independence assumption
  - Binary treatment
  - Multi-valued treatment
  - Regression
  - Example