Regression Stuff

EC 607, Set 05

Edward Rubin
Spring 2020

Prologue

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.†

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

Regression talk

Saturated models

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,

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

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 multi-valued variables, you need an indicator for each potential value.

Example$_2$ 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; β_j is the effect of j years of schooling.

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

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**.

Example[3] Regressing Wages on College Graduation and Gender.

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

Here, the uninteracted terms (β_1 & β_2) are called **main effects**; β_3 gives the effect of the **interaction**.

$$\begin{aligned}
\boldsymbol{E}[\text{Wages}_i|\text{College Graduate}_i = 0, \text{Female}_i = 0] &= \alpha \\
\boldsymbol{E}[\text{Wages}_i|\text{College Graduate}_i = 1, \text{Female}_i = 0] &= \alpha + \beta_1 \\
\boldsymbol{E}[\text{Wages}_i|\text{College Graduate}_i = 0, \text{Female}_i = 1] &= \alpha + \beta_2 \\
\boldsymbol{E}[\text{Wages}_i|\text{College Graduate}_i = 1, \text{Female}_i = 1] &= \alpha + \beta_1 + \beta_2 + \beta_3
\end{aligned}$$

Regression talk

Saturated models

The CEF can take on four possible values,

$$
\begin{aligned}
E[\text{Wages}_i | \text{College Graduate}_i = 0, \text{Female}_i = 0] &= \alpha \\
E[\text{Wages}_i | \text{College Graduate}_i = 1, \text{Female}_i = 0] &= \alpha + \beta_1 \\
E[\text{Wages}_i | \text{College Graduate}_i = 0, \text{Female}_i = 1] &= \alpha + \beta_2 \\
E[\text{Wages}_i | \text{College Graduate}_i = 1, \text{Female}_i = 1] &= \alpha + \beta_1 + \beta_2 + \beta_3
\end{aligned}
$$

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.

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 the there are many options in between—though some make less sense than others (*e.g.*, interactions without main effects).

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

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**?†

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

† *Hint:* There is no "`reg y x, causal`" command in Stata.

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.

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 requrire us to **hold many covariates constant** in order to achieve a valid causal interpretation

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*).

The conditional independence assumption

Definition(s)

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

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

To see how CIA eliminates selection bias...

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

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[†] buys us a causal interpretation, write out our old difference in means—but now condition on $\mathbf{X}_i$.

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

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

† 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**.

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

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.

$$\text{Y}_{si} \equiv f_i(s)$$

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

The conditional independence assumption

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

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

If we apply the CIA to $\text{Y}_{si} \equiv f_i(s)$, 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.*, $\text{Y}_i = f_i(s_i)$.

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

The conditional independence assumption

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

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

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

The conditional independence assumption

Example The causal effect of high-school graduation is

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

The conditional independence assumption

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

A First, remember what we just showed...

$$E[Y_i \mid X_i,\ s_i = 12] - E[Y_i \mid X_i,\ s_i = 11] = E[f_i(12) - f_i(11) \mid X_i]$$

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

$$E\Big(E[Y_i \mid X_i,\ s_i = 12] - E[Y_i \mid X_i,\ s_i = 11] \Big)$$

$$= E\Big(E[f_i(12) - f_i(11) \mid X_i] \Big)$$

$$= E[f_i(12) - f_i(11)]$$ (Iterating expectations)

The conditional independence assumption

Takeaways

1. Conditional independence gives our parameters **causal interpretations** (elminating 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 $(\mathbf{X}_i)$ 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.

From the CIA to regression

Conditional independence fits into our regression framework in two ways.

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

2. If we allow $f_i(s)$ to be nonlinear in s and heterogeneous across i, regression provides a weighted average of individual-specific differences $f_i(s) - f_i(s-1)$.[†]

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

† Leads to a matching-style estimator.

From the CIA to regression

Let $f_i(s)$ be linear in s and equal across i except for an error term, *e.g.*,

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

Substitute in our observed value of s_i and the outcome Y_i

$$\text{Y}_i = \alpha + \rho s_i + \eta_i \tag{B}$$

While ρ in (A) is explicitly causal, regression-based estimates of ρ 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…

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

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

$$\eta_i = \mathbf{X}_i'\gamma + \nu_i \tag{C}$$

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

Note Least-squares regression implies

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

From the CIA to regression

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

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

The CEF of $f_i(s_i)$ is linear, which means that the (right[†]) population regression will be the CEF.

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

Regression and causality

From the CIA to regression

Thus, the linear causal (regression) model is

$$\mathrm{Y}_i = \alpha + \rho s_i + \mathrm{X}_i'\gamma + \nu_i$$

The residual ν_i is uncorrelated with

1. s_i (from the CIA)
2. $\mathbf{X}_i$ (from defining γ via the regression of η on $\mathbf{X}_i$)

The coefficient ρ gives the causal effect of s_i on Y_i.

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.

CIA example

Let's continue with the returns to graduation $(\mathbf{G}_i)$.

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.

CIA example

First, we need to generate some data.

```r
# 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)
)
```

CIA example

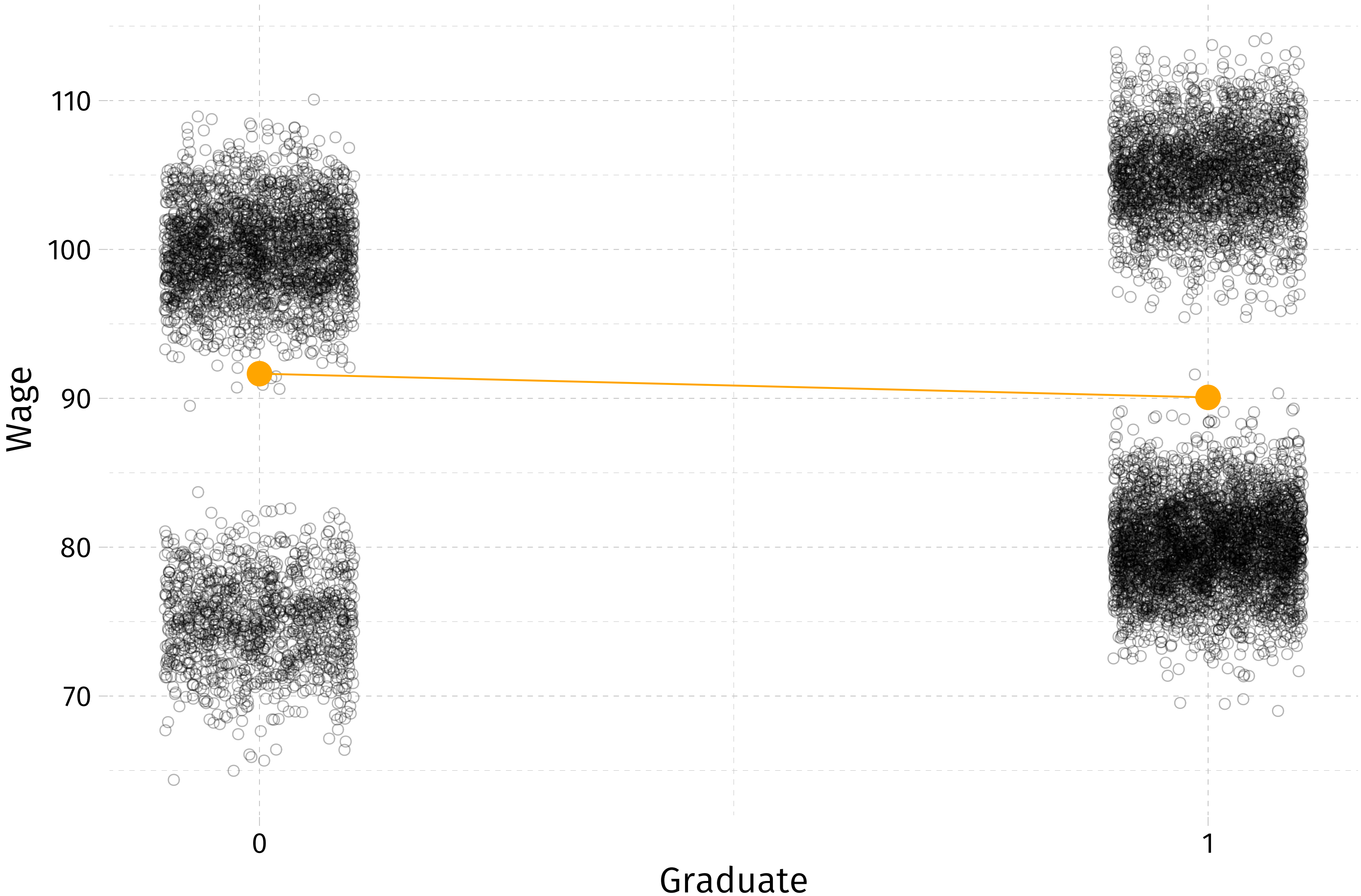
Now we can estimate our naïve regression

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

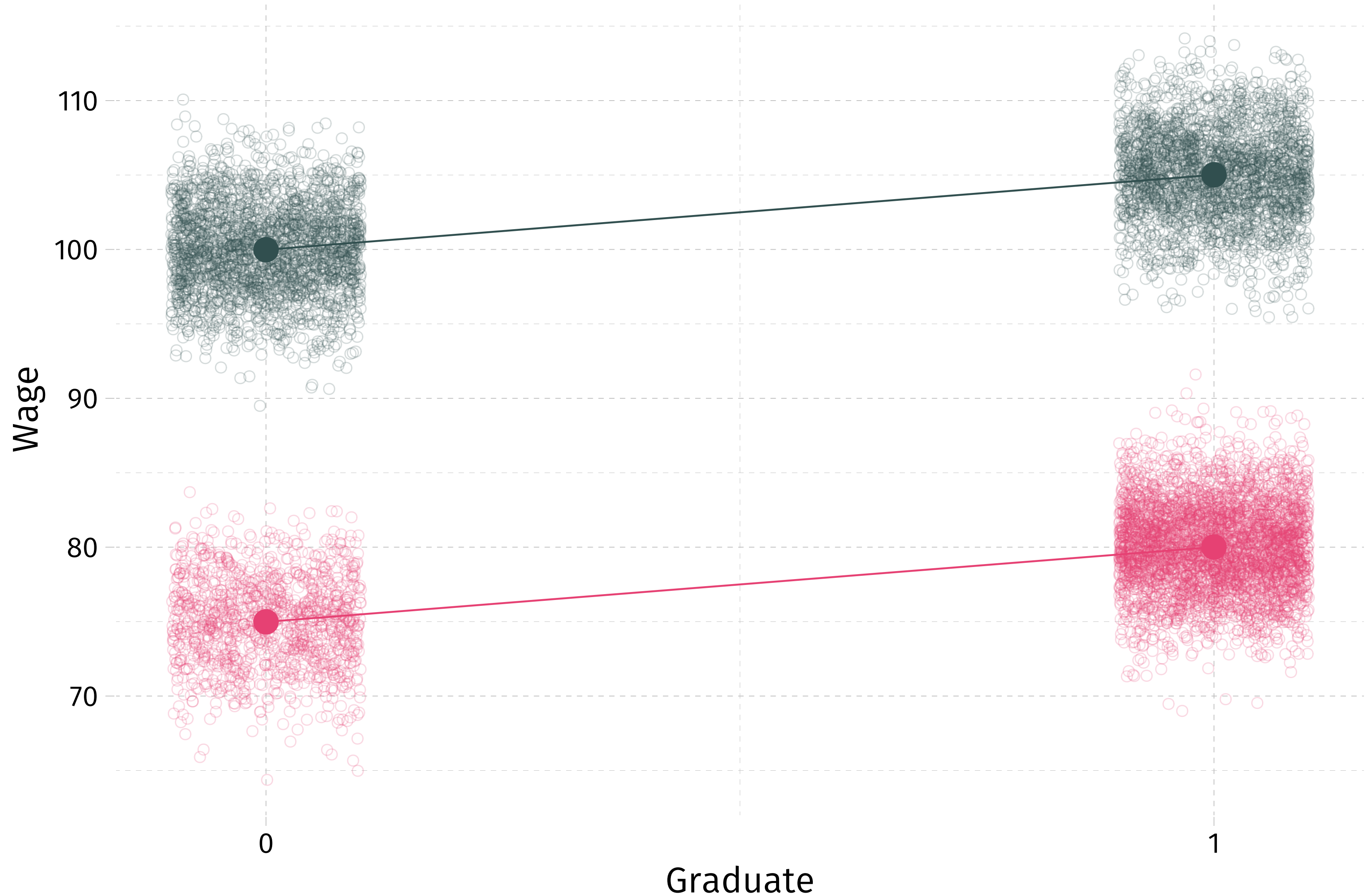
```
lm(wage ~ grad, data = ex_df)
```

	Coef.	**S.E.**	**t stat**
Intercept	91.65	0.20	447.70
Graduate	-1.59	0.26	-6.18

Maybe we should have plotted our data…


110
100
90
80
70
Wage
0
1
Graduate

We're still missing something...



110
100
90
80
70
Wage
0
1
Graduate

CIA example

Now we can estimate our causal regression

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

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

	Coef.	**S.E.**	**t stat**
Intercept	99.98	0.05	1868.81
Graduate	5.03	0.06	78.23
Female	-25.00	0.06	-402.64

Table of contents