Learning from Observational Data

EC 350: Labor Economics

Kyle Raze

Winter 2022

Learning from Observational Data

1. A taxonomy of data
    - Experimental *vs.* observational data
2. Direct acyclic graphs
    - Causal paths
    - Backdoor paths
    - Backdoor criterion
3. Regression discontinuity

A taxonomy of data

A taxonomy of data

Experimental

Data generated from a **randomized** experiment.

- Treatment assigned at **random**
- The **gold standard** of social science research
- Often difficult/impractical/unethical to conduct

Observational (non-experimental)

Data generated from the **decisions** of various individuals in the "real world."

- Sometimes treatment is randomly assigned (*e.g.*, in a lottery), but not usually **(non-random!)**
- Prone to selection bias and omitted-variable bias
- Must rely on natural experiments to identify causal relationships

A taxonomy of data

Example: Effect of job training on unemployment status

Experimental sample

	Unemployed? (= 1 if yes, = if no)			
	1	**2**	**3**	**4**
Training?	-0.111	-0.116	-0.115	-0.113
	(0.044)	(0.044)	(0.044)	(0.044)
Control mean	0.354	0.354	0.354	0.354
Demographics		✓	✓	✓
Education			✓	✓
Unemployed?$_{t-1}$				✓

Note: Standard errors in parentheses.

Non-experimental sample

	Unemployed? (= 1 if yes, = if no)			
	1	**2**	**3**	**4**
Training?	0.128	0.164	0.160	-0.182
	(0.025)	(0.027)	(0.027)	(0.027)
Control mean	0.115	0.115	0.115	0.115
Demographics		✓	✓	✓
Education			✓	✓
Unemployed?$_{t-1}$				✓

Note: Standard errors in parentheses.

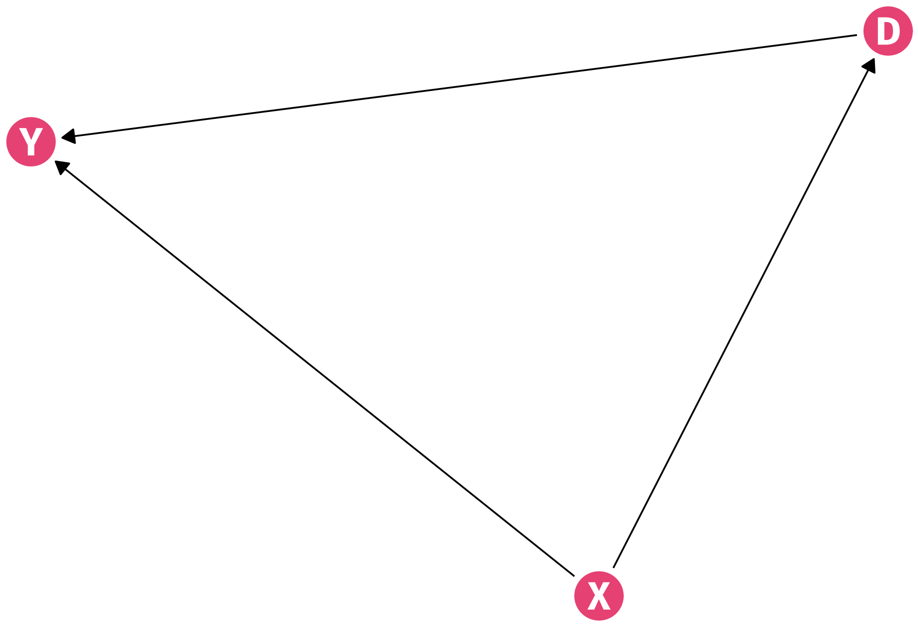
Direct acyclic graphs

Direct acyclic graphs

A direct acyclic graph (DAG) can help us visualize the assumptions necessary to estimate causal relationships using observational data.

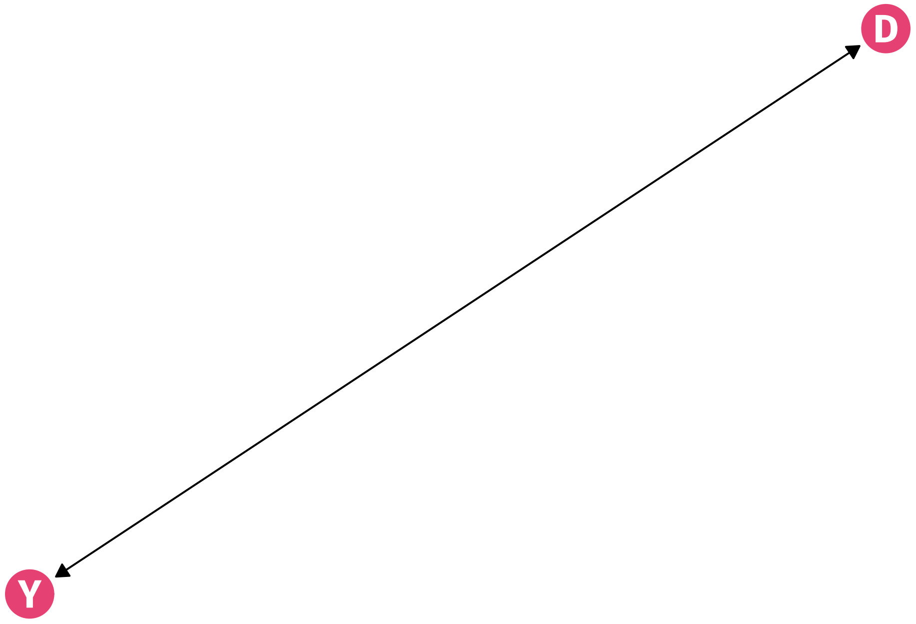
Nodes represent **variables**.

Arrows represent **causal relationships** between variables.



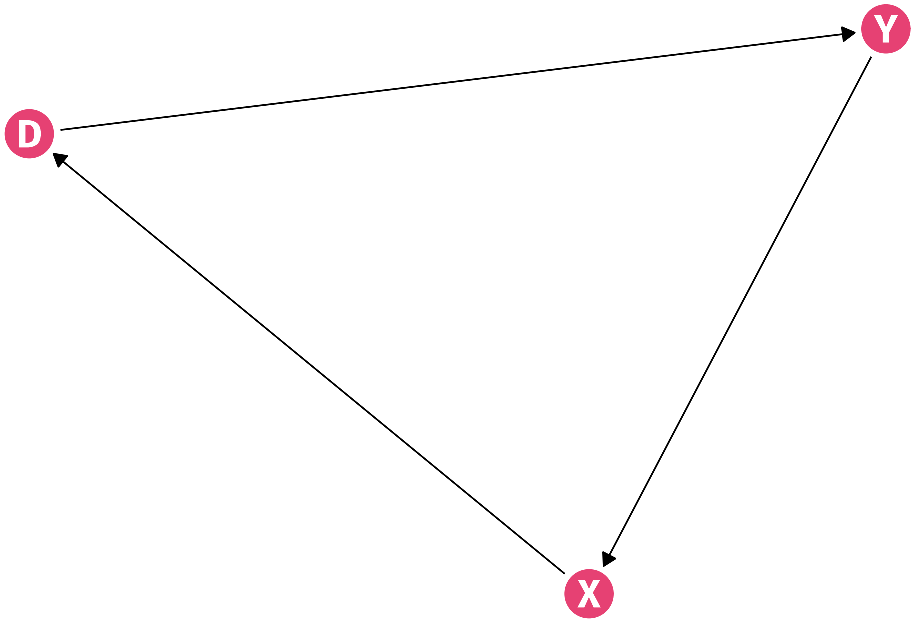
DAGs follow two rules

Rule 1 ("direct"): No bidirectional arrows!



Illegal!

Rule 2 ("acyclic"): No feedback loops!

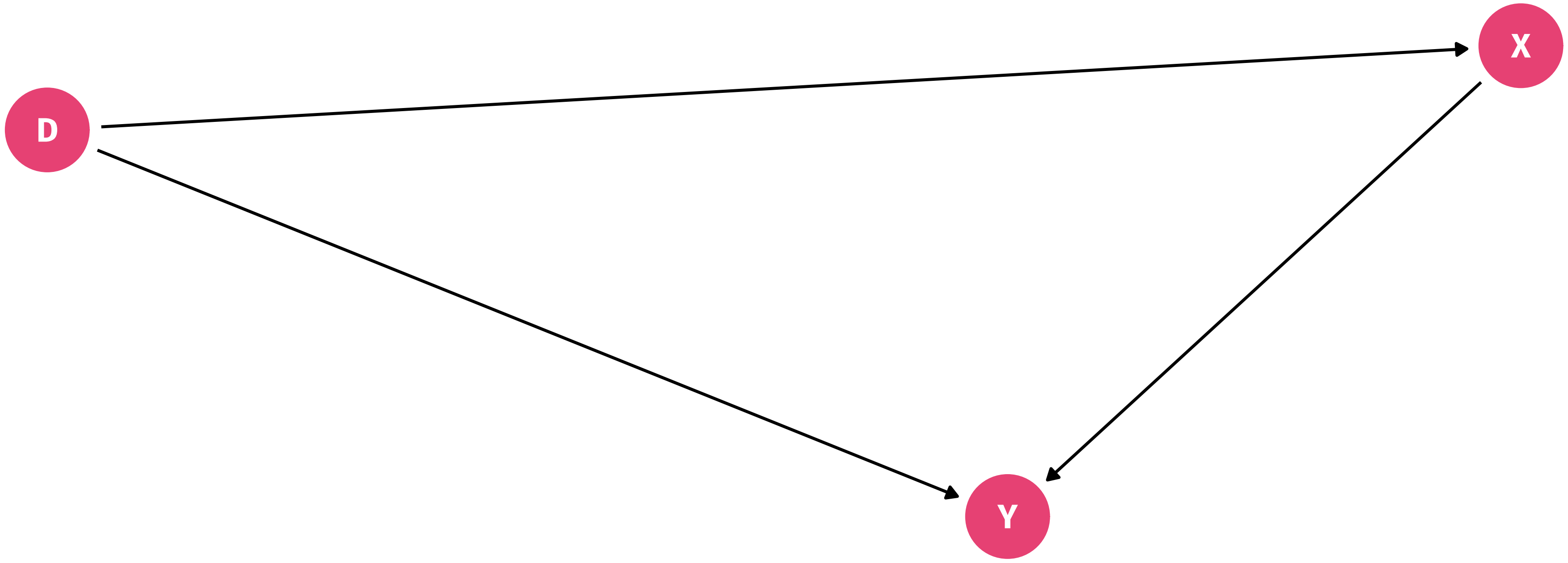


Illegal!

Causal paths

Our objective is to **identify the causal effect** of a treatment variable **D** on an outcome variable **Y**.

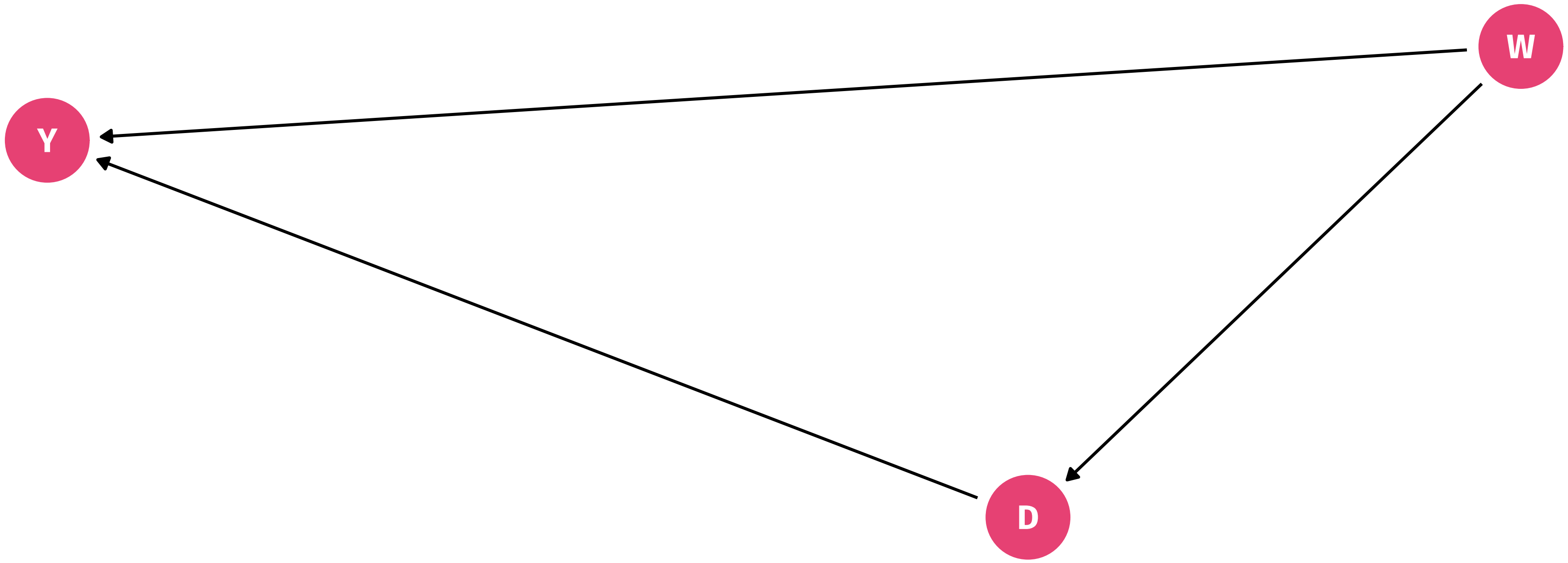
- The treatment could have a **direct effect** on the outcome: **D** $\longrightarrow$ **Y**.
- Alternatively, the treatment could have an **indirect effect** on the outcome through **X**, a mediator variable: **D** $\longrightarrow$ **X** $\longrightarrow$ **Y**.



Backdoor paths

The presence of a confounder variable **W** opens a **backdoor path** from the treatment to the outcome:

$$\mathbf{D} \longleftarrow \mathbf{W} \longrightarrow \mathbf{Y}$$



An open backdoor path creates a **spurious correlation** between the treatment and the outcome!

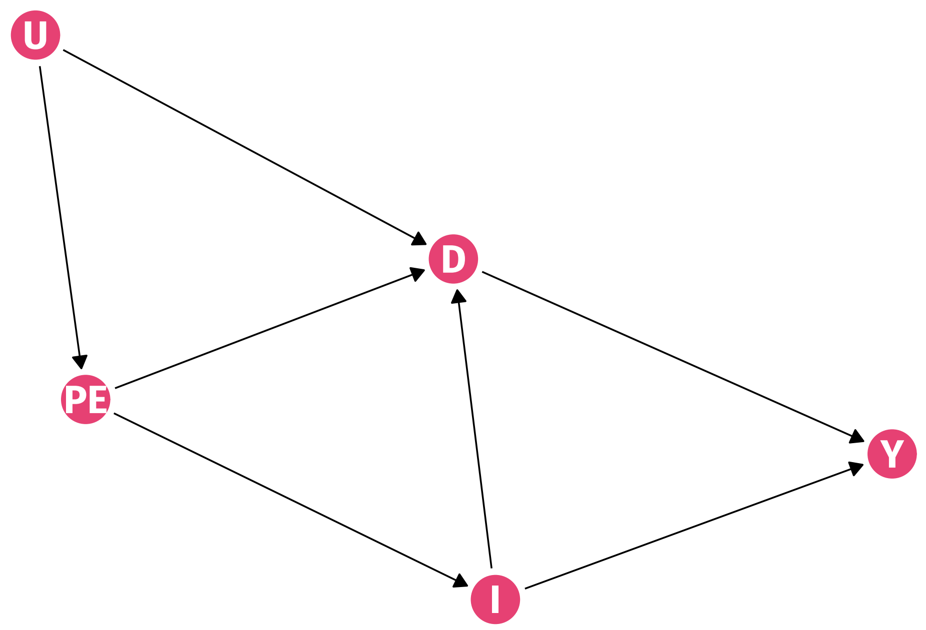
Backdoor paths

Example: Returns to education

Q: How does education affect earnings?

- **D** = Education (*e.g.*, going to college or not)
- **Y** = Earnings as an adult
- **PE** = Parental education
- **I** = Family income
- **U** = Unobserved characteristics (*e.g.*, family background)

The presence—*or absence*—of an arrow illustrates our **causal assumptions** about how education affects earnings!

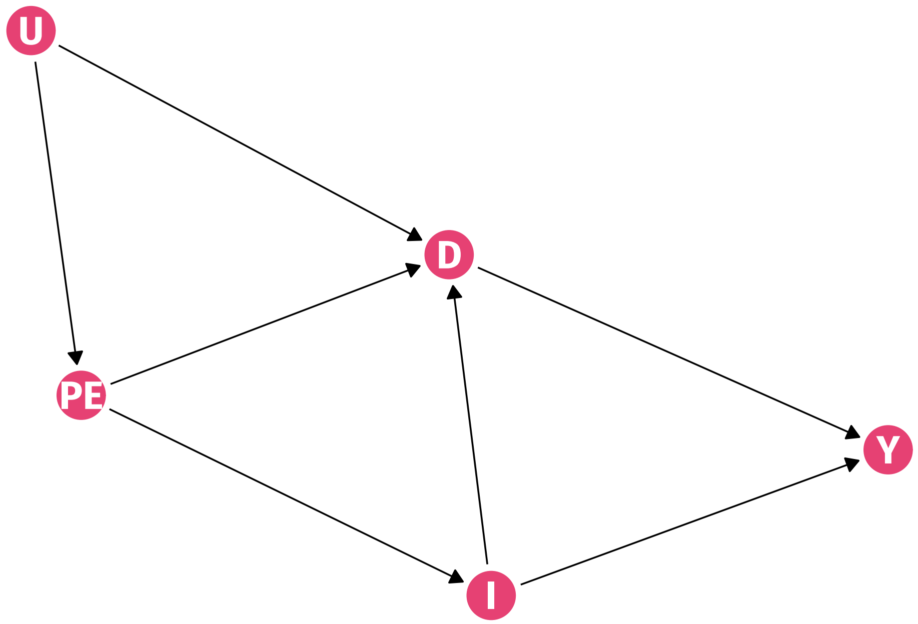


Backdoor paths

Example: Returns to education

Q: What are the paths through which education affects earnings?

- **D** $\longrightarrow$ **Y** (causal effect)
- **D** $\longleftarrow$ **I** $\longrightarrow$ **Y** (backdoor path)
- **D** $\longleftarrow$ **PE** $\longrightarrow$ **I** $\longrightarrow$ **Y** (backdoor path)
- **D** $\longleftarrow$ **U** $\longrightarrow$ **PE** $\longrightarrow$ **I** $\longrightarrow$ **Y** (backdoor path)



Backdoor paths

Backdoor criterion

> The observed correlation between **Y** and **D** isolates the causal effect of **D** on **Y** if and only if all backdoor paths from **D** to **Y** are closed.

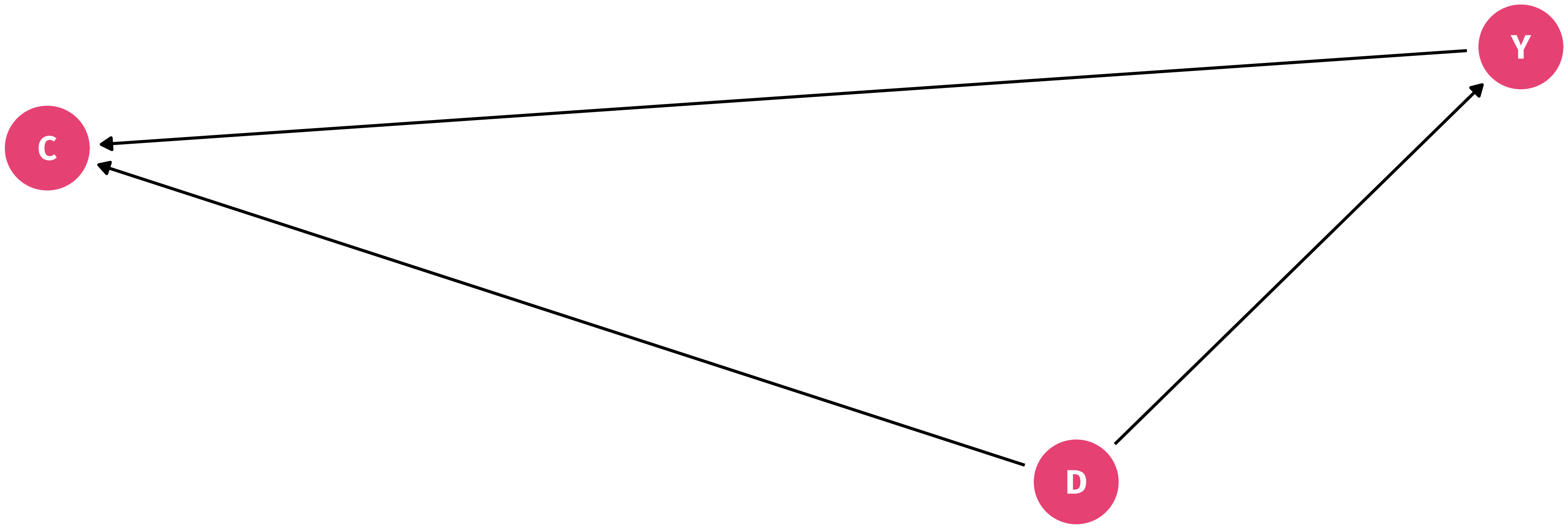
Q: What closes a backdoor path?

- **A_1:** *Conditioning* or *controlling for* the confounder variable on the path.
- **A_2:** The presence of a collider variable on the path.

Backdoor paths

The presence of a collider variable **C** closes a backdoor path from the treatment to the outcome:

$$\mathbf{D} \longrightarrow \mathbf{C} \longleftarrow \mathbf{Y}$$



The implication? We don't want to control for collider variables!

- Conditioning on a collider can open up new backdoor paths. (More on this later.)

Backdoor paths

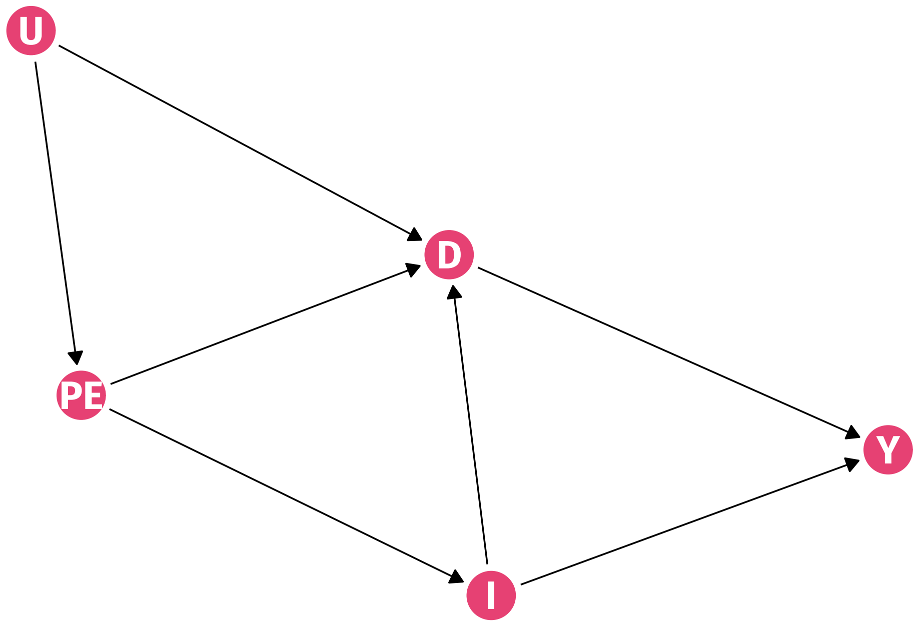
Example: Returns to education

Q: How could we satisfy the backdoor criterion given our assumptions about the effect of education on earnings?

A: Control for family income (**I**)

- **Why?** Family income appears as a non-collider on each backdoor path:

$$\mathbf{D} \longleftarrow \mathbf{I} \longrightarrow \mathbf{Y}$$
$$\mathbf{D} \longleftarrow \mathbf{PE} \longrightarrow \mathbf{I} \longrightarrow \mathbf{Y}$$
$$\mathbf{D} \longleftarrow \mathbf{U} \longrightarrow \mathbf{PE} \longrightarrow \mathbf{I} \longrightarrow \mathbf{Y}$$



Backdoor paths

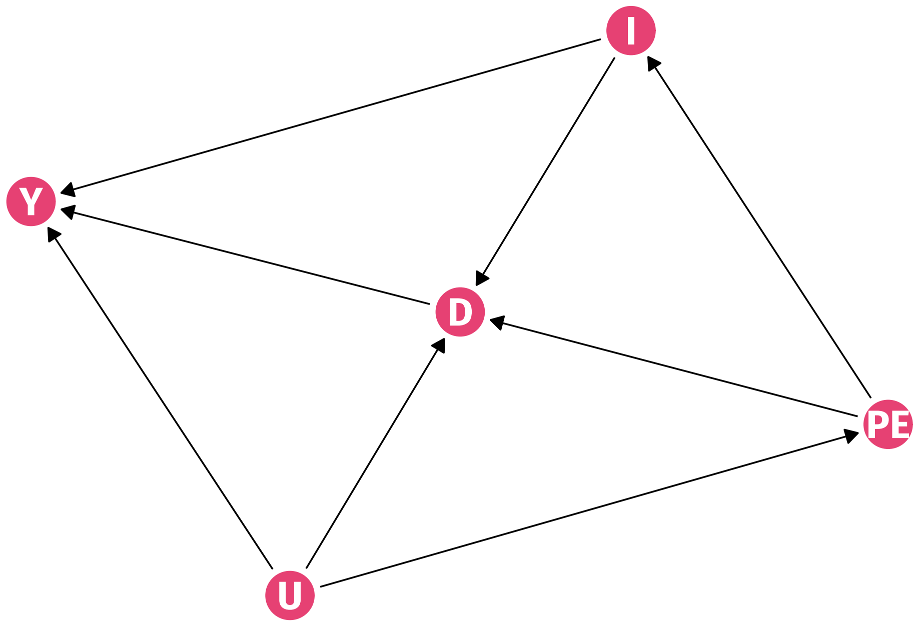
Example: Returns to education

Q: Would controlling for family income isolate the causal effect of education on earnings if unobserved family background (**U**) has a direct effect on earnings (**Y**)?

A: No!

- **U** is unobserved, so we can't control for it.
- The backdoor path **D** $\longleftarrow$ **U** $\longrightarrow$ **Y** would stay open.

The takeaway?
ALL causal inference is by assumption!



Regression discontinuity

Regression discontinuity

There are situations in the real world where treatment is assigned in a way that is **as good as random.**

- These situations can provide **valid comparison groups**, just like the ones you'd find in a randomized control trial!

Examples? When some arbitrary threshold triggers a change in treatment:

- Anti-discrimination laws only apply to firms with more than 15 employees.
- Prisoners are eligible for early parole if some score exceeds a threshold.
- An individual has legal access to alcohol if they are 21 or older.
- You get a ticket if your speed exceeds the speed limit.
- A candidate for governor wins if her vote share exceeds that of her competitors.

Economists can (and often do) use these situations to estimate causal effects.

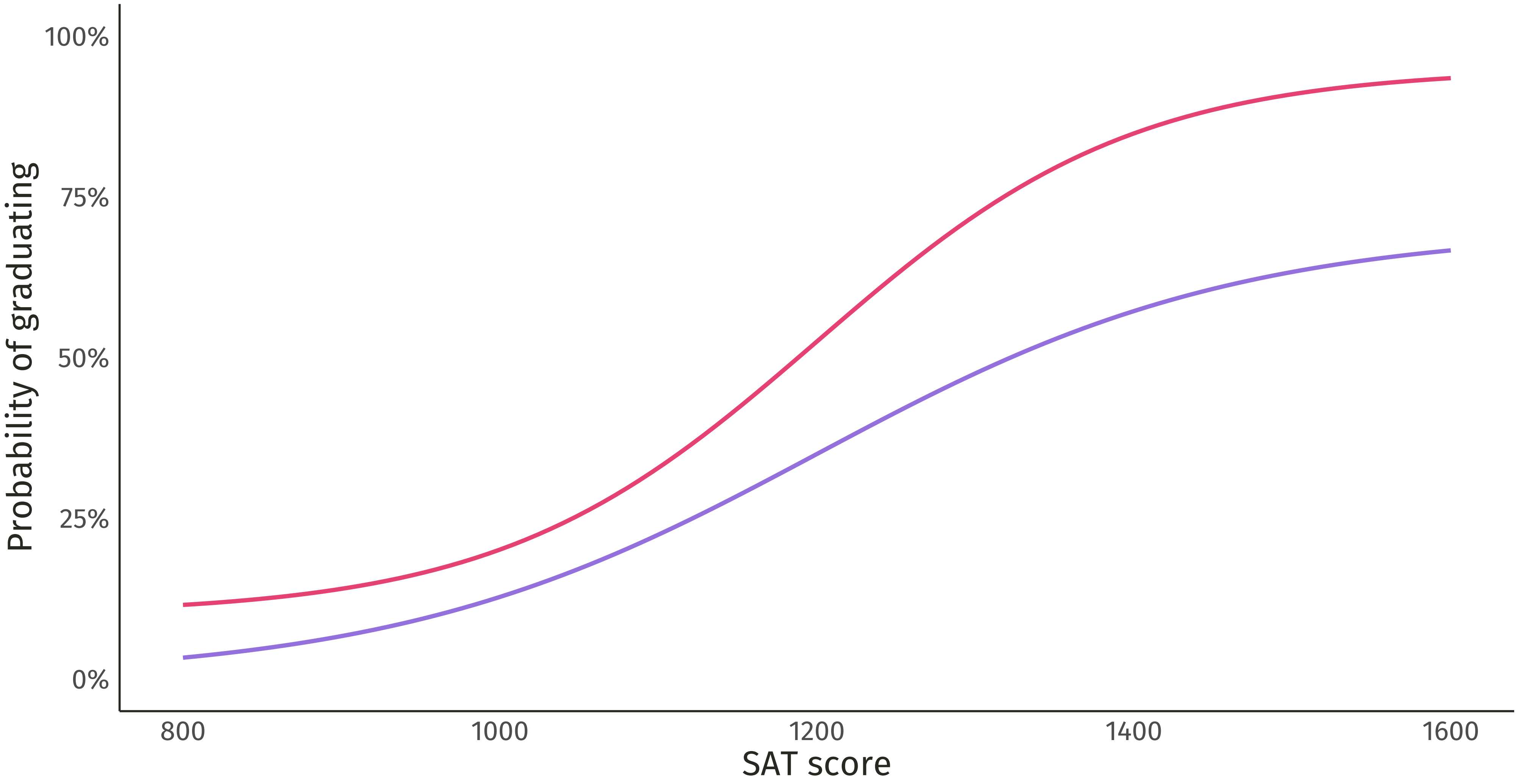
Regression discontinuity

Example: Effect of merit scholarships on graduation

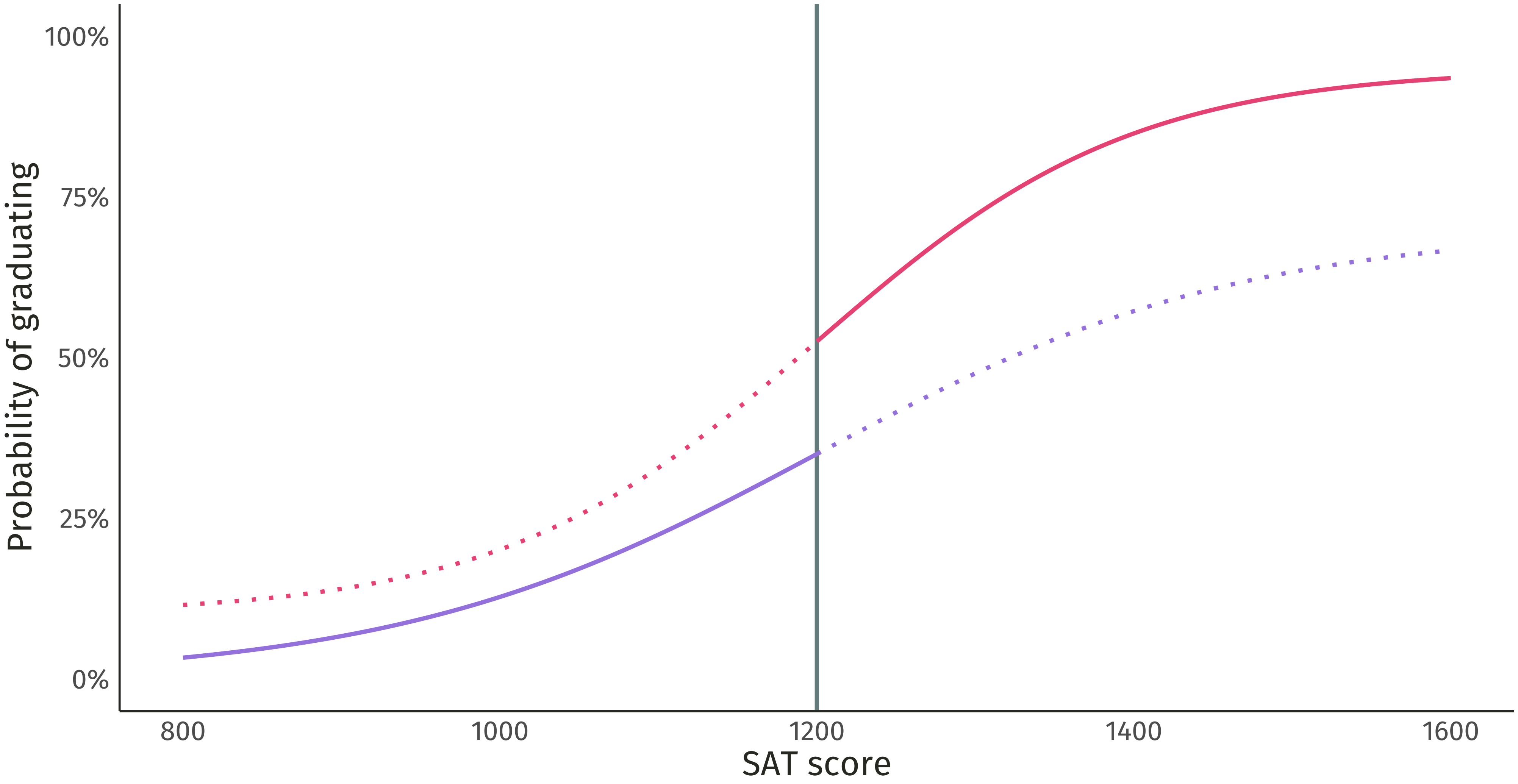
- Outcome variable = probability of graduation
- Treatment = scholarship money
- "Assignment variable" = admissions test score (*e.g.*, the SAT)
- "Cutoff/threshold" = minimum score for getting a scholarship (*e.g.*, SAT score of 1200 or higher)

Assumption: Students *just below* the cutoff are comparable to those *just above* the cutoff.

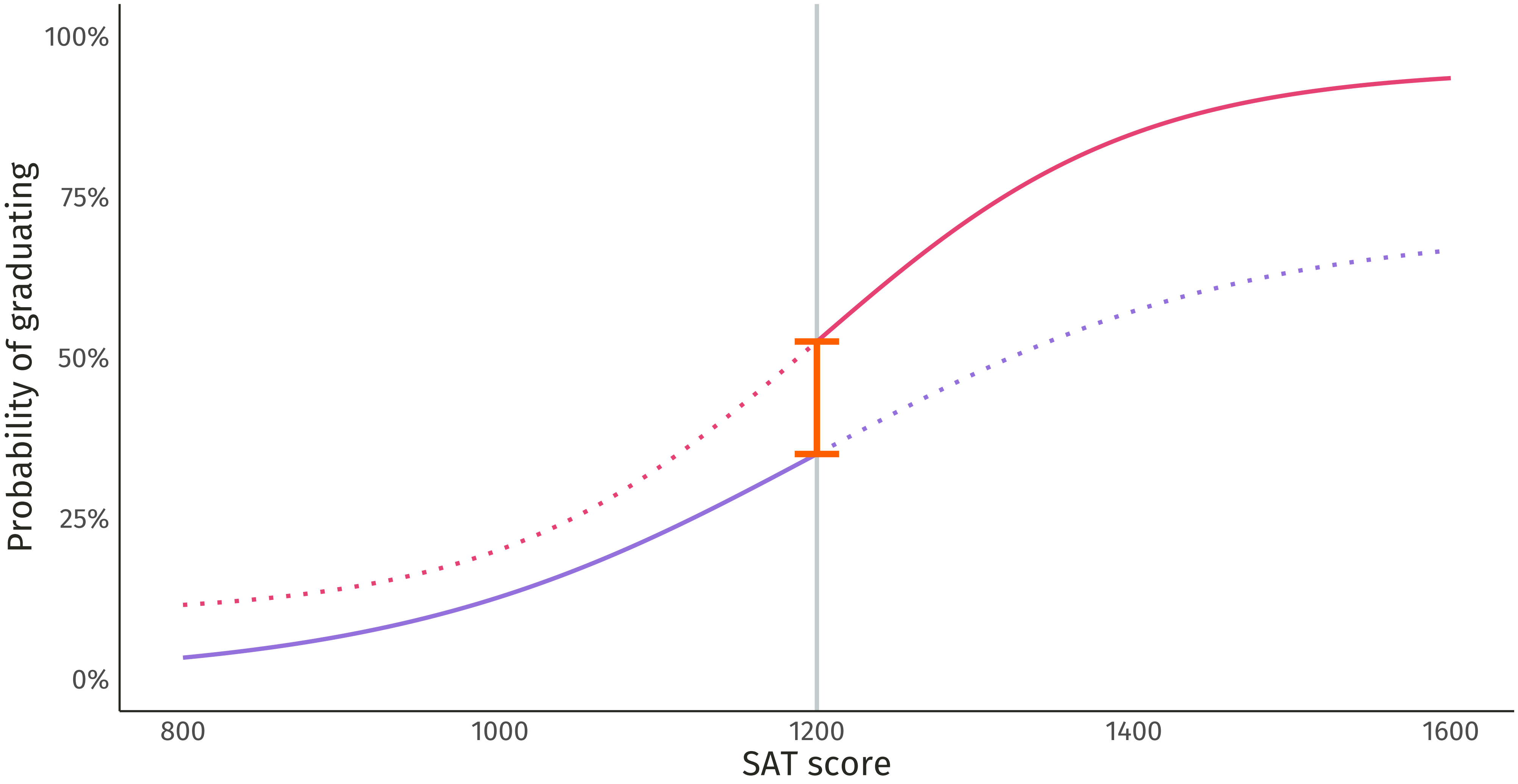
Let's start with potential graduation rates: $E[\mathrm{Y}_{0,i} \mid \mathrm{SAT}_i]$ and $E[\mathrm{Y}_{1,i} \mid \mathrm{SAT}_i]$.



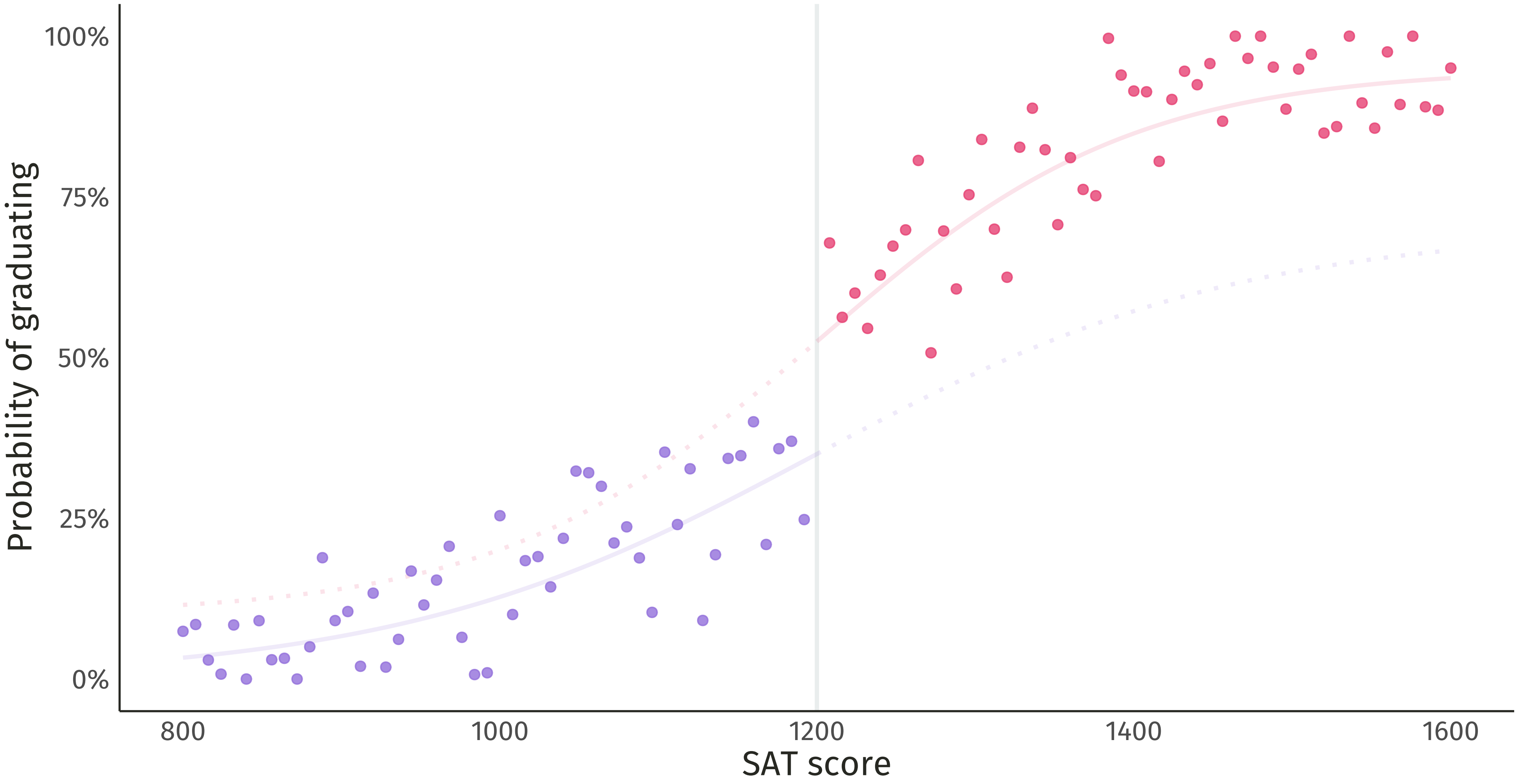
You only get a scholarship if if your **SAT score exceeds the cutoff score**.



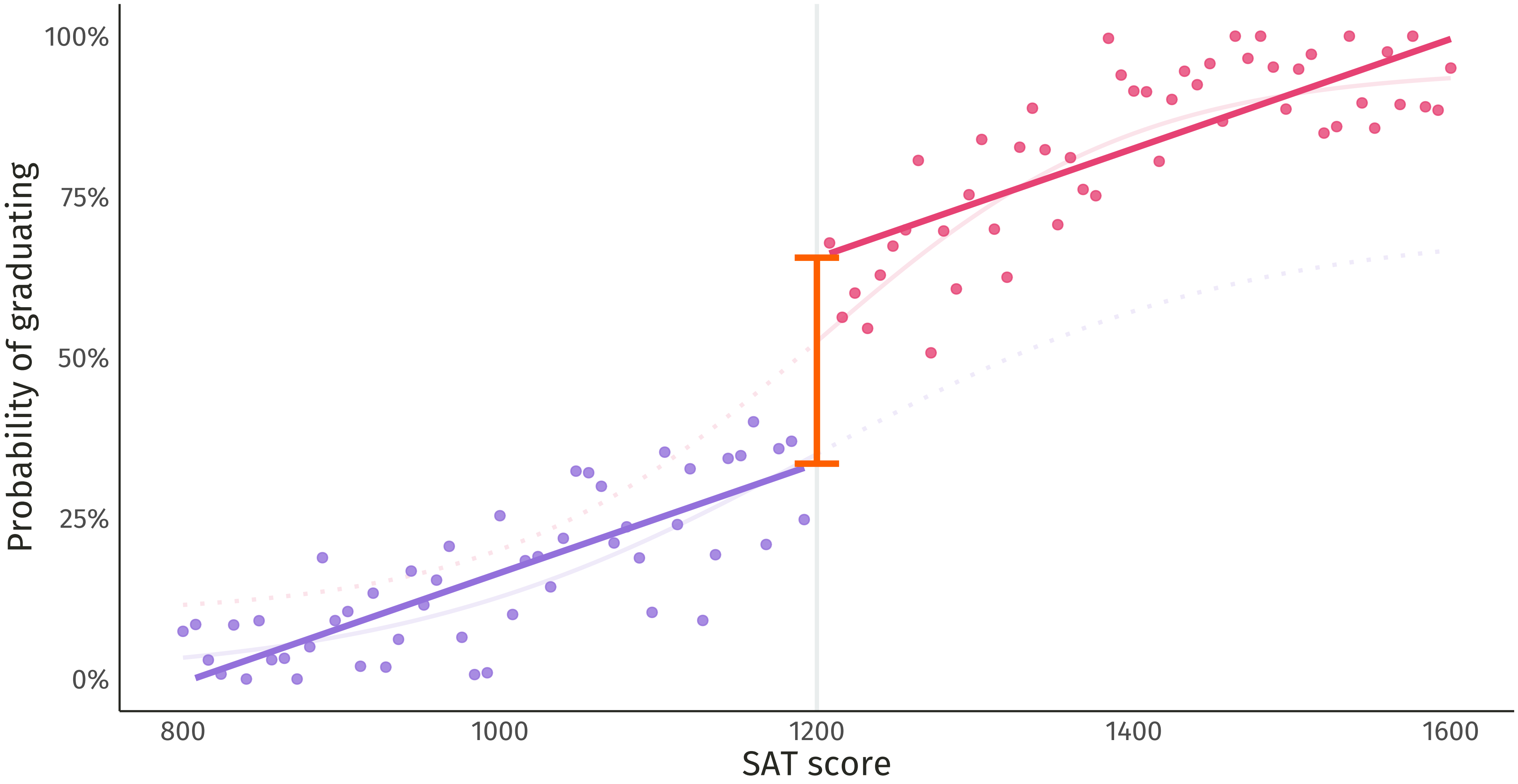
$E[\mathrm{Y}_{1,i} \mid \mathrm{SAT}_i = 1200] - E[\mathrm{Y}_{0,i} \mid \mathrm{SAT}_i = 1200]$ gives the **causal effect at the cutoff**.



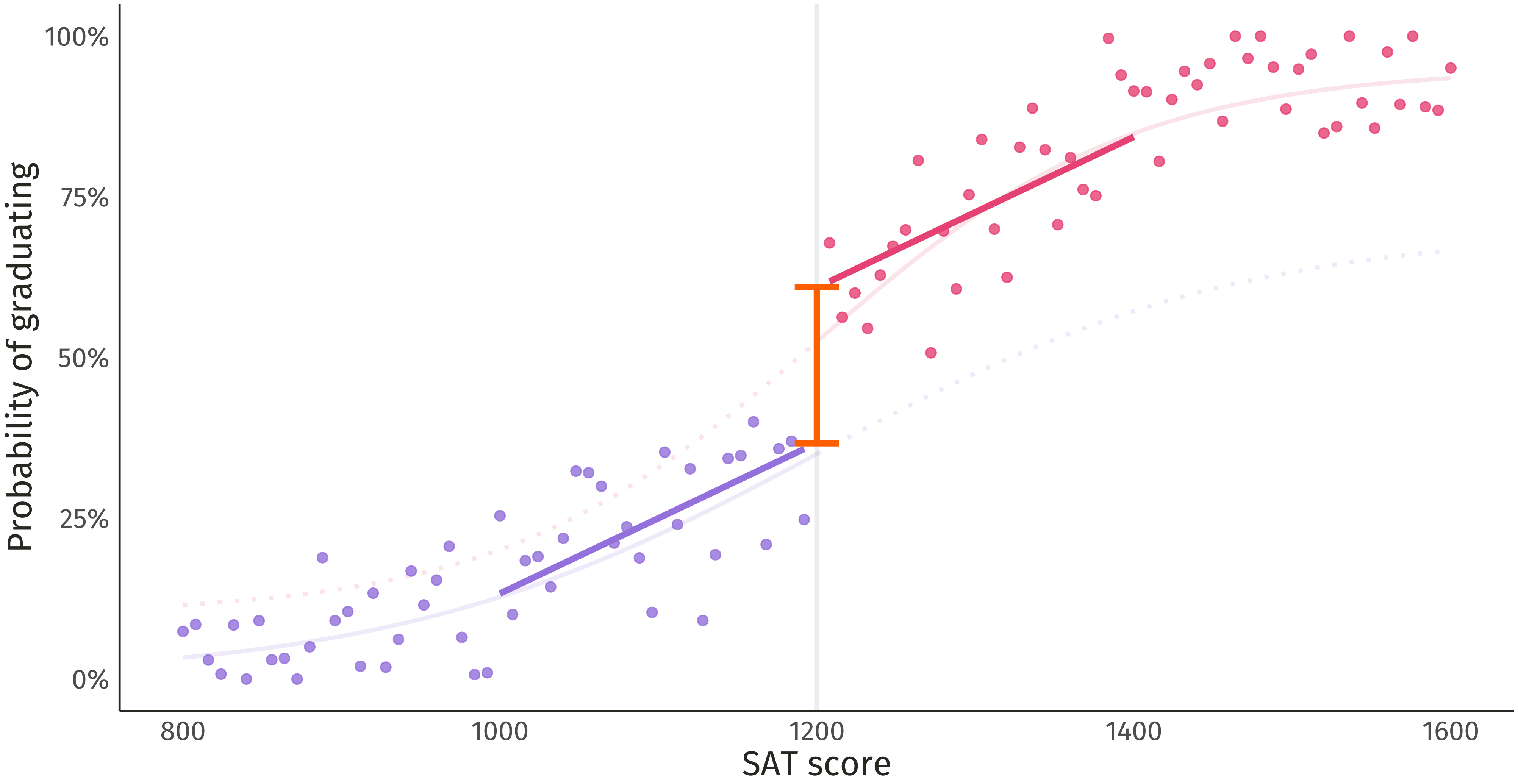
Using real data, researchers have to estimate $E[\mathrm{Y}_{1,i} \mid \mathrm{SAT}_i]$ and $E[\mathrm{Y}_{0,i} \mid \mathrm{SAT}_i]$.



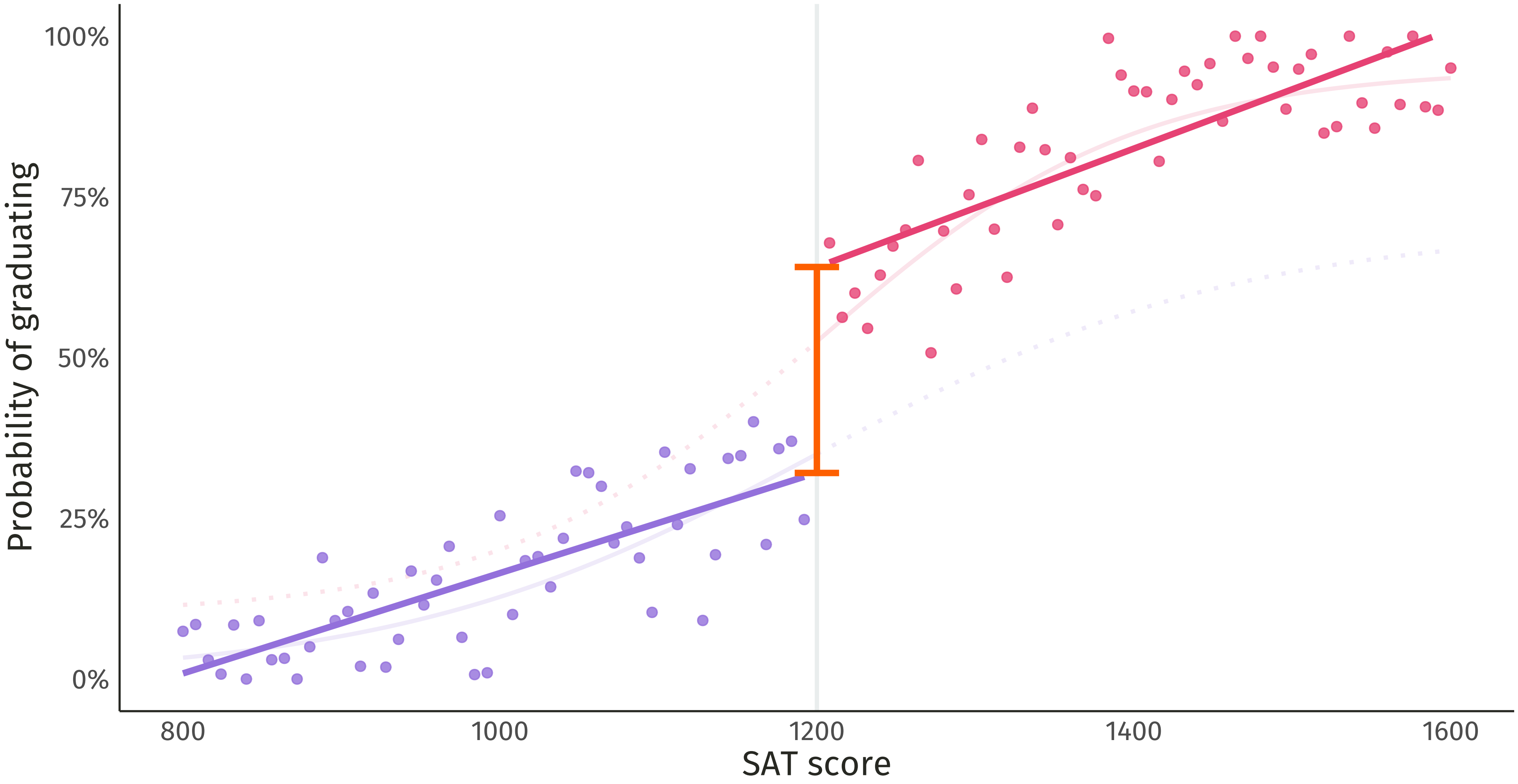
One way to estimate the **jump** is to estimate a regression on each side of the cutoff.



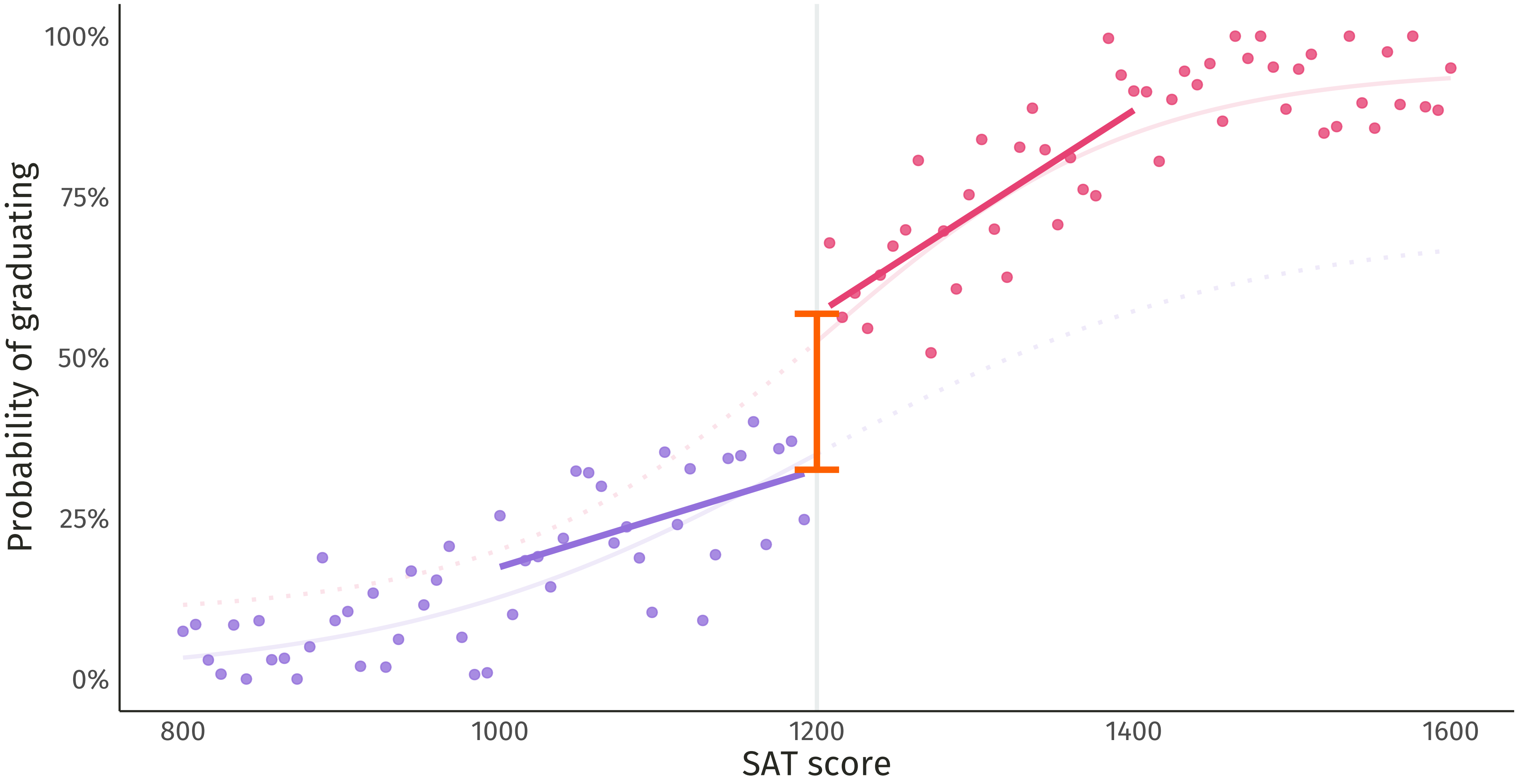
Another way is to estimate regressions using only data closer to the cutoff.



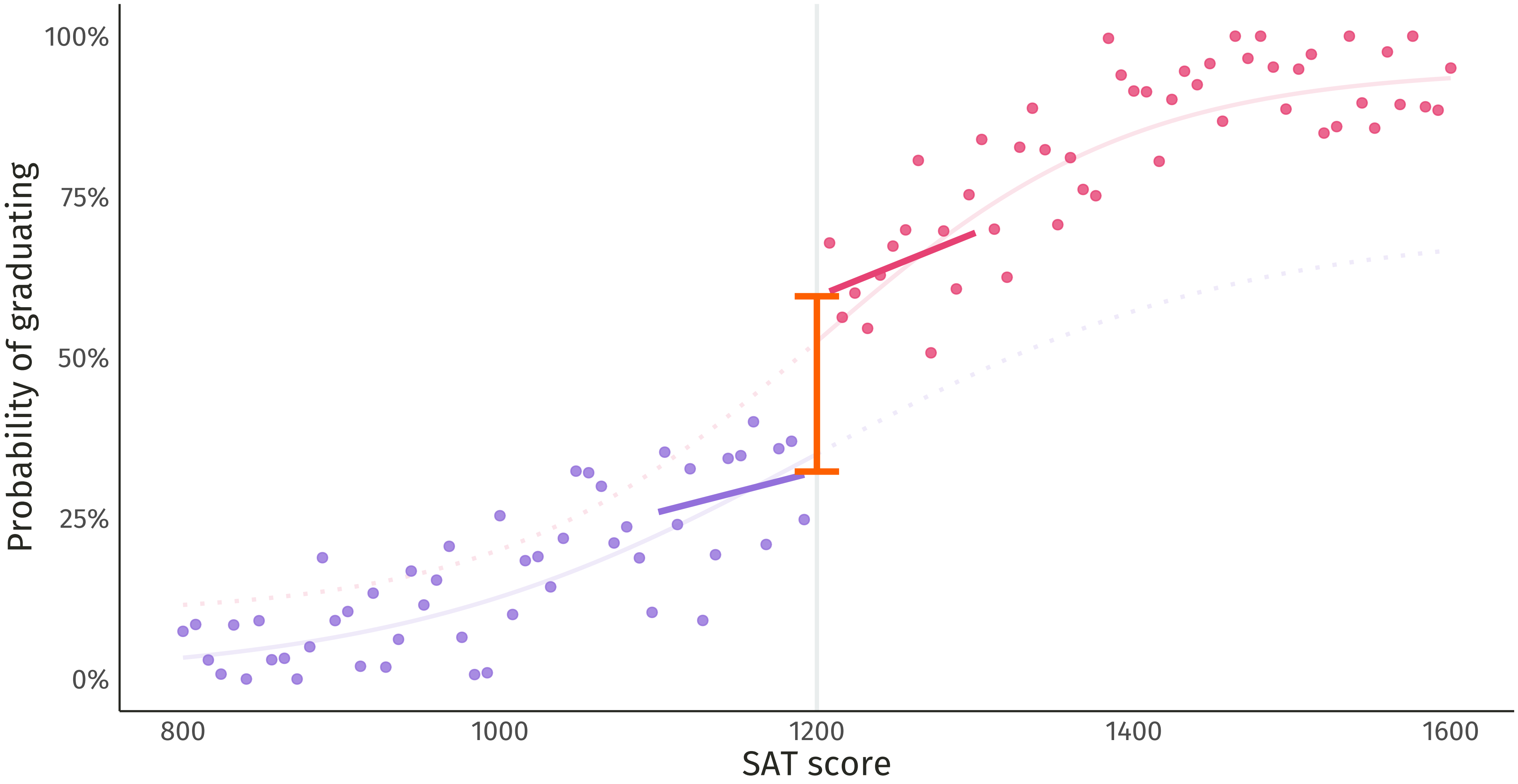
Different choices of samples and models can lead to different estimates of the treatment effect!



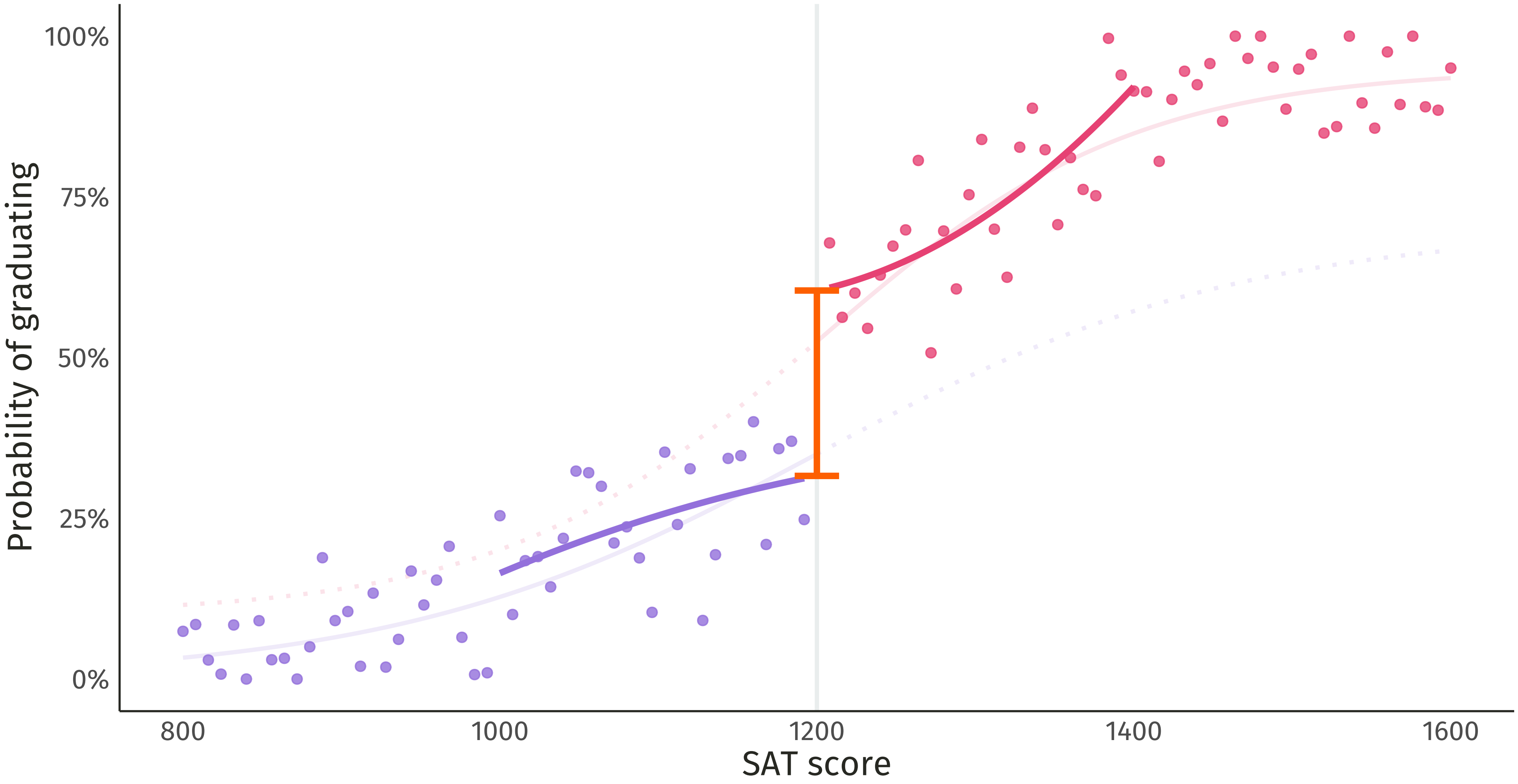
Different choices of samples and models can lead to different estimates of the treatment effect!



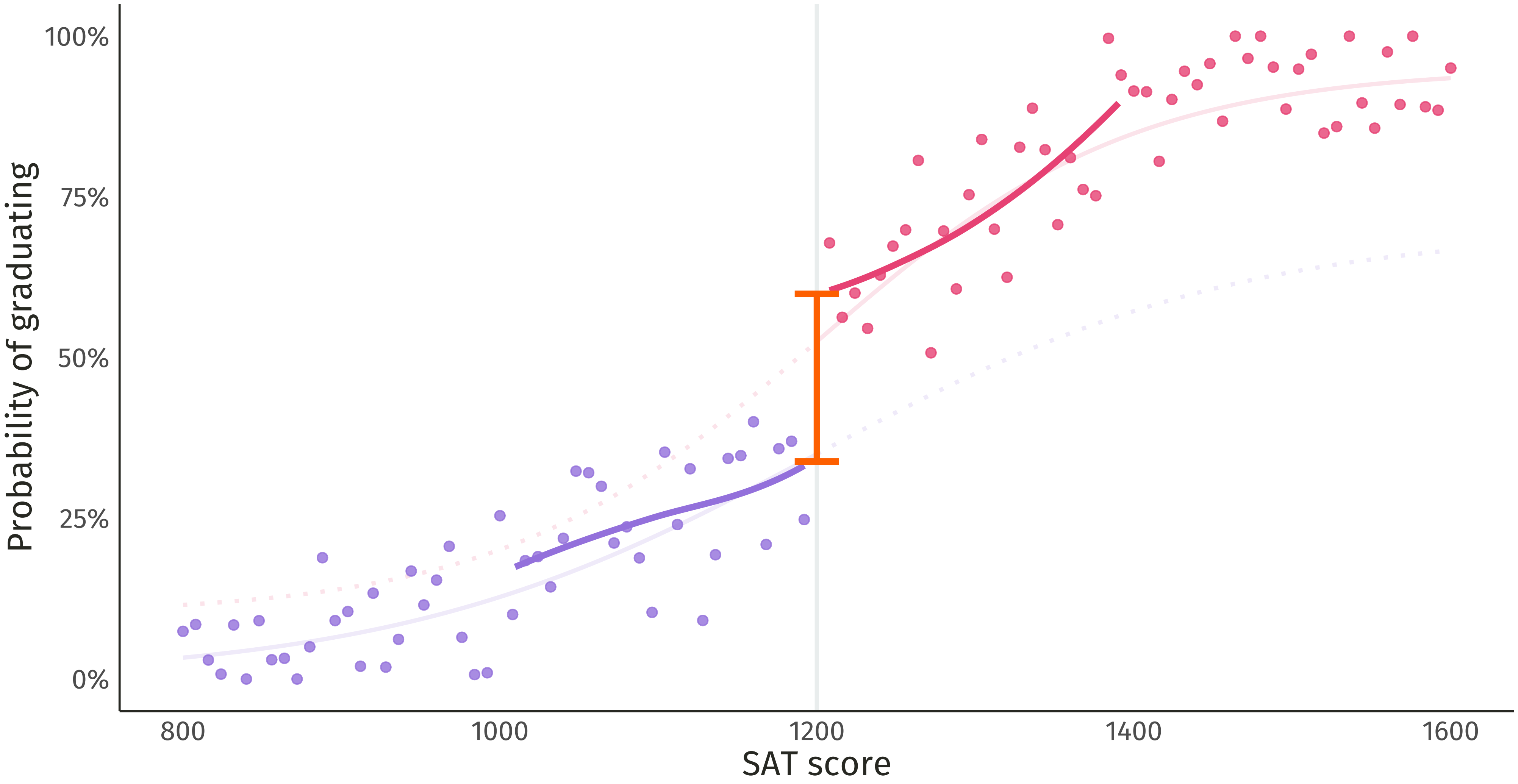
Different choices of samples and models can lead to different estimates of the treatment effect!



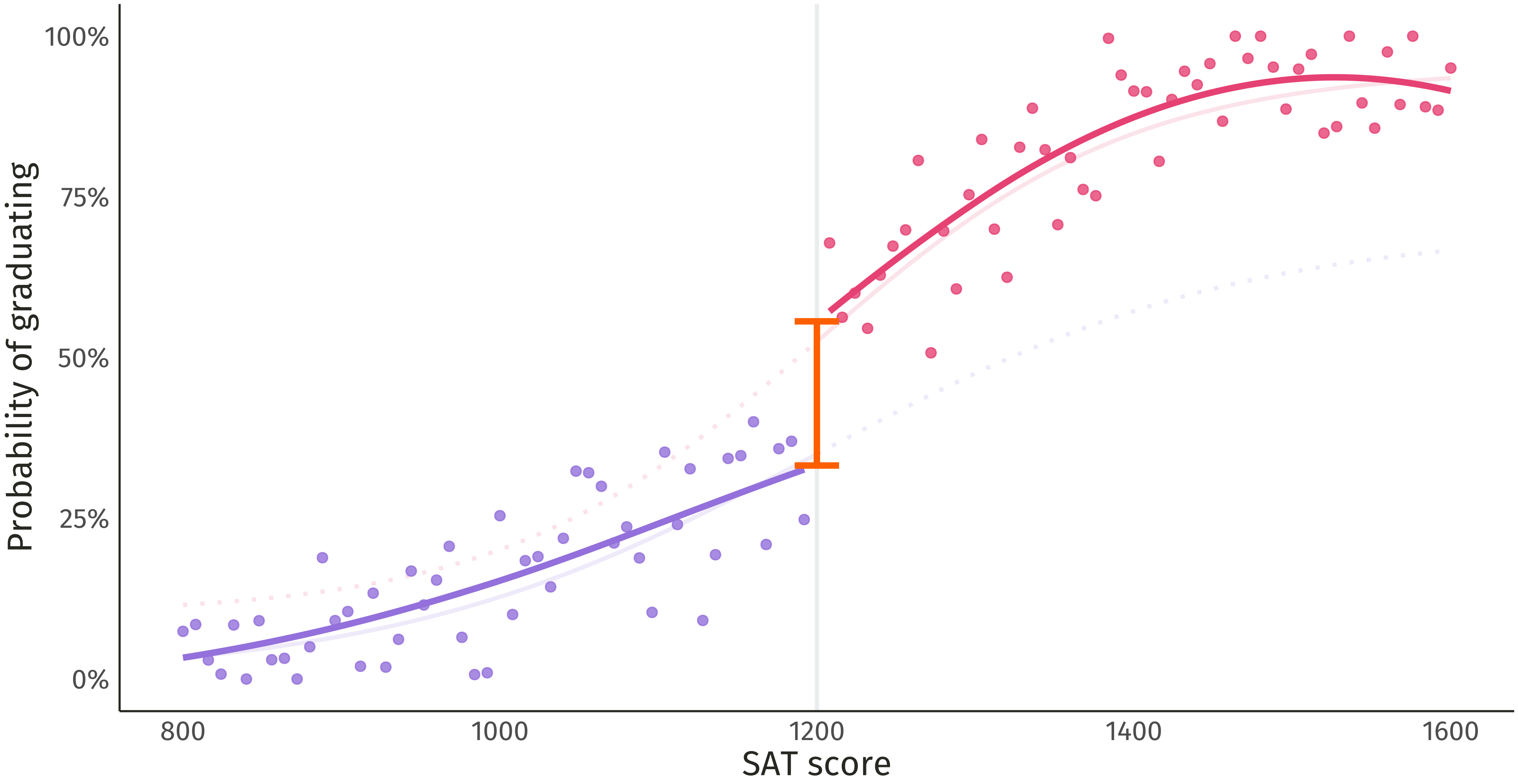
Different choices of samples and models can lead to different estimates of the treatment effect!



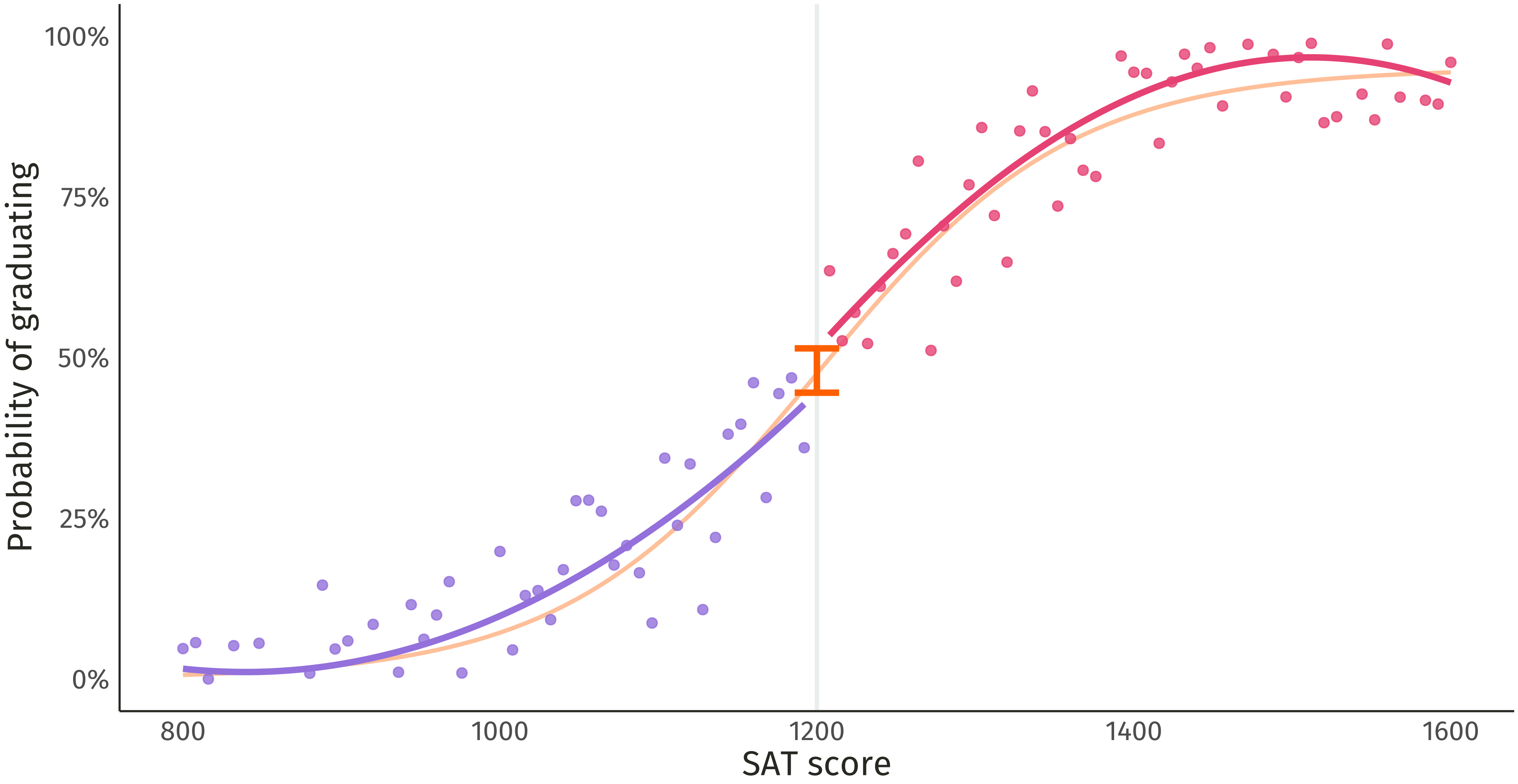
Different choices of samples and models can lead to different estimates of the treatment effect!



Different choices of samples and models can lead to different estimates of the treatment effect!



Some modeling choices can find an effect even if none exists!



Regression discontinuity

Q: When should we trust a regression discontinuity comparison?

- When is the comparison *internally valid*?

A: When we believe that **treatment is the only thing that changes** (other than observed outcomes) at the cutoff.

1. We don't want to see evidence of people **bunching** on one side of the threshold.
    - This could mean that people are **manipulating the assignment variable** near the cutoff so that they get the treatment.
    - Example: cheating among students who anticipate being close to the cutoff as a way to increase their score just enough to get the scholarship.
2. We don't want to see a **"jump" in other variables** at the cutoff.
    - This would mean that people on one side of the cutoff are **no longer comparable** to people on the other side!

Regression discontinuity

Q: How can we tell if the treatment actually has a causal effect on the outcome?

A: The treatment has an effect if **all three** of the statements below are true.

1. We believe that the regression discontinuity comparison is **internally valid.**
2. We can see that the **outcome variable "jumps"** at the cutoff ***when we look at the raw data.***
3. The estimate of the "jump" is **precise enough** to conclude that the effect is statistically significant.