What's The Point Of Statistical Modeling?

Applied Regression in R

Aleš Vomáčka

23. 03. 2026

Faculty of Arts, Charles University

So far, we have learn:

1. How linear regression is created
2. What regression coefficients mean
3. How to postprocess our models
4. How to quantify uncertainity
5. How to summarise model fit

All this is neccesary, but not sufficient to do research.

What Is a Model Anyway?

What Is a Model Anyway?

(Statistical) model are simplified approximations of reality.

What Is a Model Anyway?

(Statistical) model are simplified approximations of reality.

This is important:

(Statistical) model are simplified approximations of reality!!!!

What Is a Model Anyway?

Many people either over- or underestimate models usefulness.

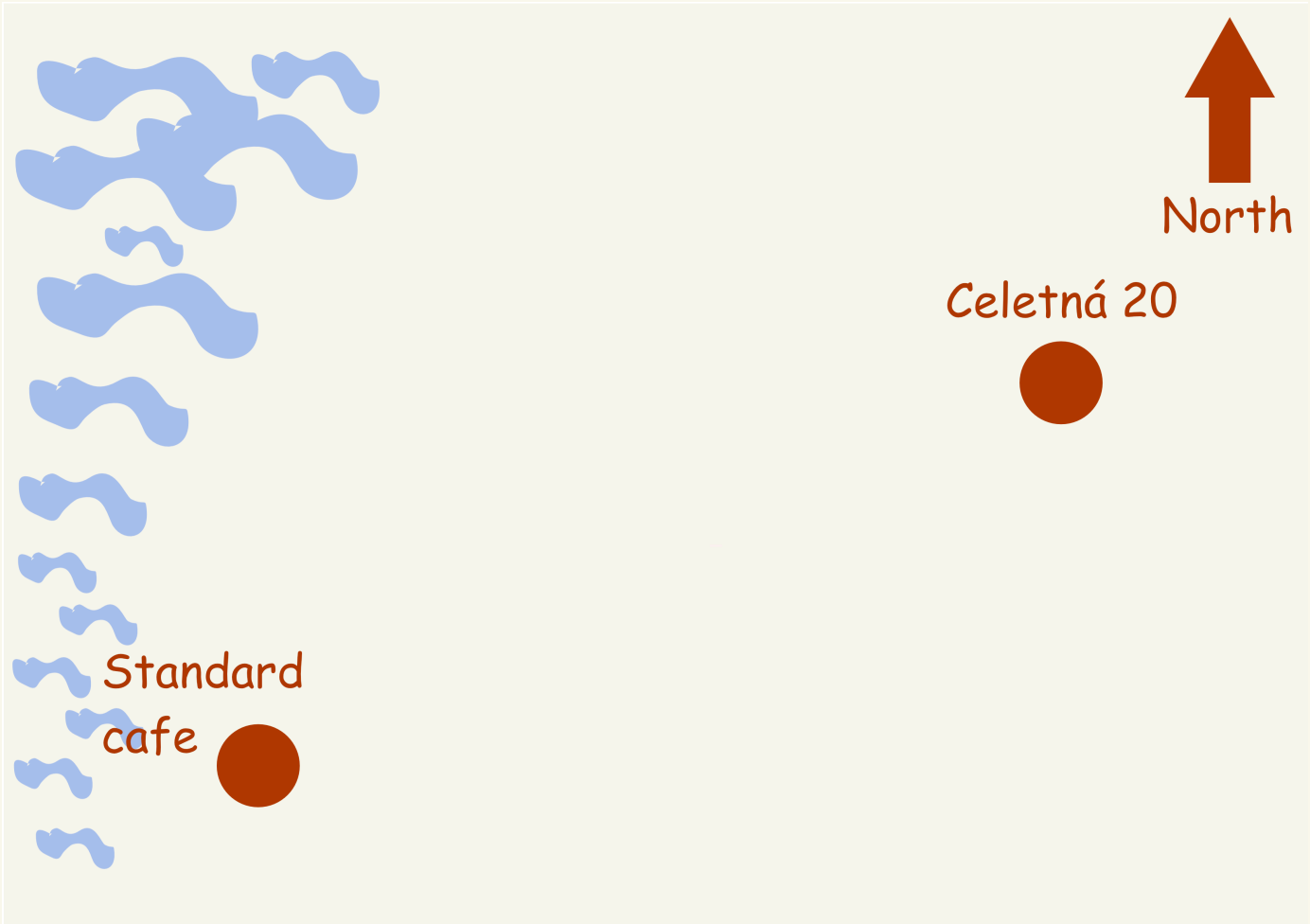
A good model can be useful. But makes a model good?

What Is a Model Anyway?

You have befriended a foreign student. They are currently at Celetná and you are supposed to meet Standard cafe.

They've asked you for directions. They are also a bit old-fashioned and prefer a map. What kind of map would be the best?

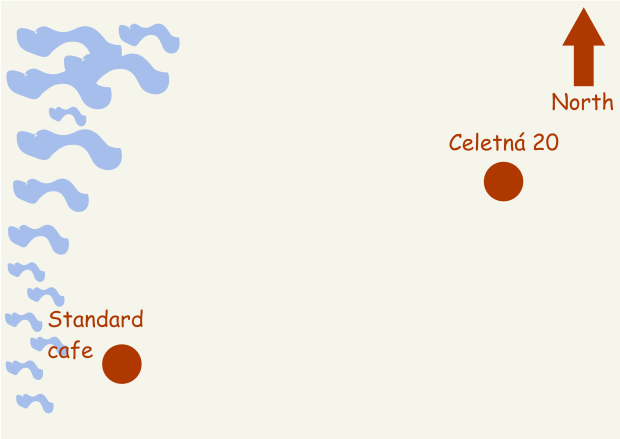
What Is a Model Anyway?



What Is a Model Anyway?

What Is a Model Anyway?

Too simple

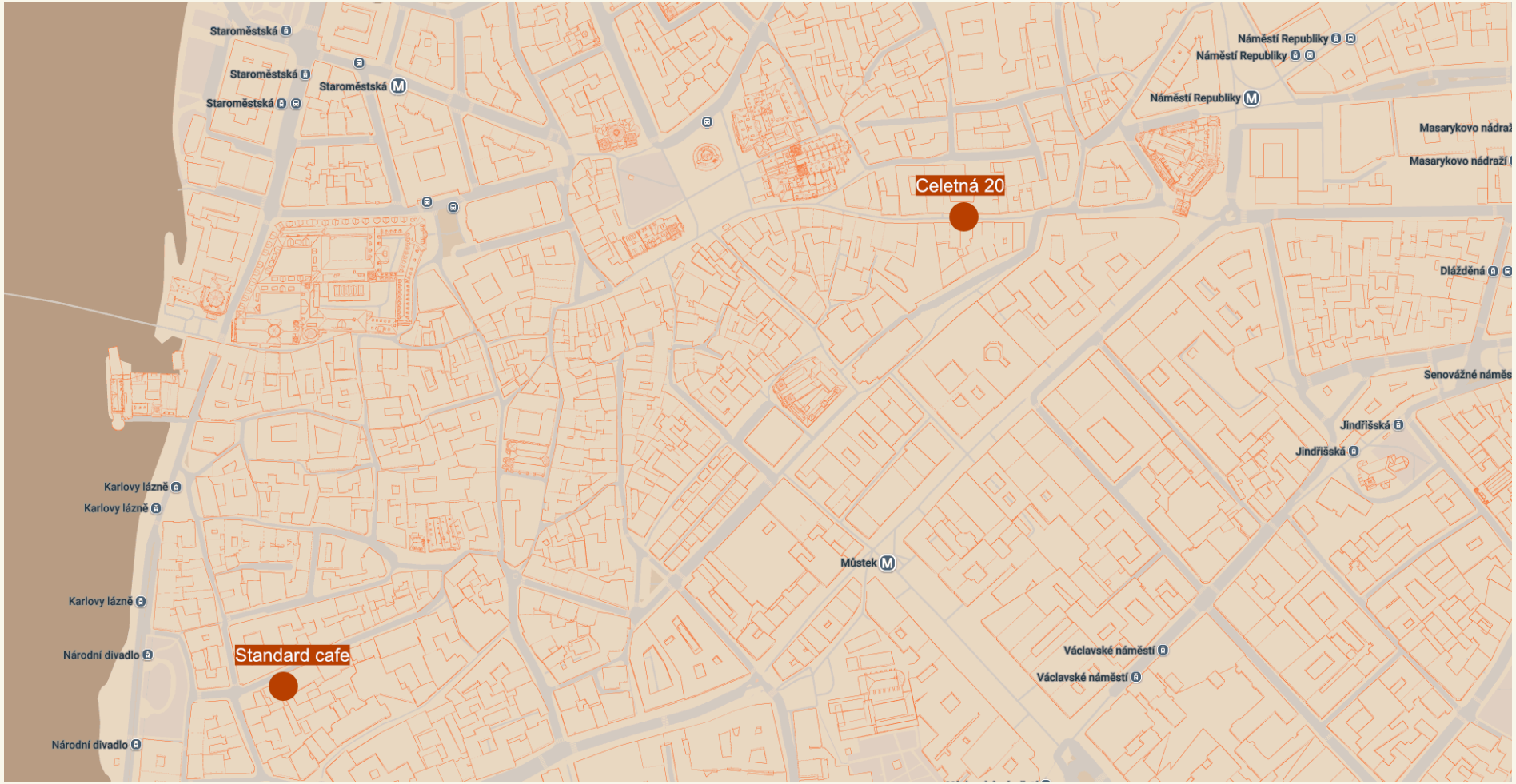


??

↔

Too complicated

What Is a Model Anyway?



What Is a Model Anyway?

A **good model contains all important details** neccesary to solve our problem and omits the rest.

But how can we tell which details are important?

Questions?

Goals of Statistical Modeling

Goals of Statistical Modeling

What kind of map is the best?

- When we are going on a hike → a map with hiking paths.
- When we are driving → a map with traffic information.
- When we are building a house → a topography map

What details matter depends on our goals.

Goals of Statistical Modeling

To argue whether a model is good or bad, we need to know what problem it's tryign to solve.

Roughly speaking, there are three uses for statistical modeling.

Goals of Statistical Modeling

Descriptive models

- Goal is to summarize data (and quantify uncertainty).
- Few assumptions.

Predictive models

- Goal is to predict unobserved data.
- Estimating relationships less important.
- More assumptions.

Explanative models

- Goals is estimating causal relationships.
- Predictive power less important.
- Many assumptions.

Goals of Statistical Modeling

Descriptive models

- Pre-election models.
- What is the gender pay gap in Czechia?
- What is the correlation between social class and school performance?

Predictive models

- What will the voter turnout be next elections?
- What will the unemployment rate be next month?
- Which students will perform best if admitted?

Explanative models

- What is the causal effect of social class on school performance?
- What is the causal effect of teacher's attractivity on course ratings?

Goals of Statistical Modeling

At the beginning of every research project, you need to decide what type of modeling you want to do.

This will determine everything else, from the way we judge model quality to how we pick which variables to include.

Questions?

Variable Selection

Variable Selection

How you select predictors for you model depends on what kind of modeling you want to do.

- Descriptive modeling - pretty much anything goes.
- Predictive modeling - pick predictors that **maximase out-of sample predictive power** (e.g. adjusted R^2).
- Explanative modeling - pick predictors that **minimize bias** in regression coefficients.

Variable Selection

We'll focus on **variable selection for explanative models**.

Hardest (and in sociology, most common) problem you'll encounter.

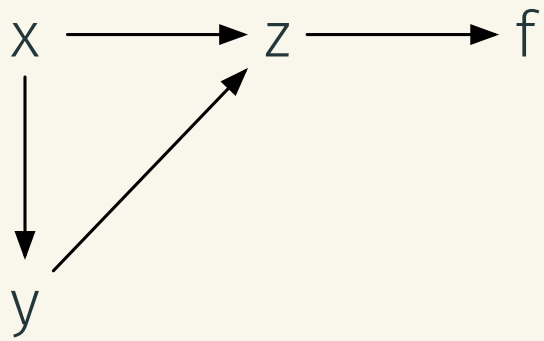
The goal is to pick „control" variables which allow you to interpret the relationship between „main" predictor and outcome as causal effect.

Variable Selection

Many, many ways to do causal inference. A good way to start are **directed acyclic graphs** (DAGs).

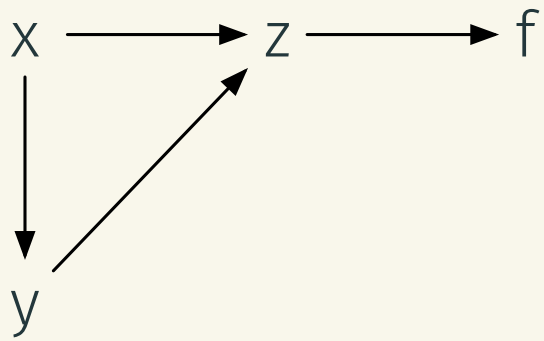
Variable Selection

This is a directed acyclic graph.



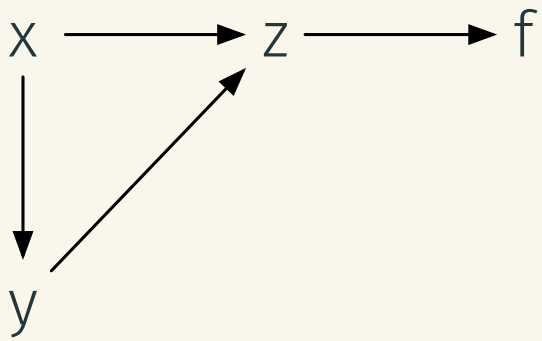
Variable Selection

It's a graph - a set of **nodes** connected by **edges** (lines)



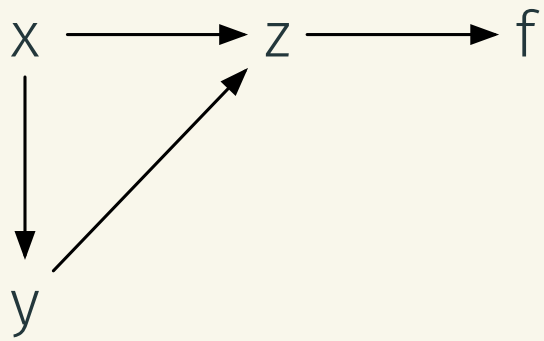
Variable Selection

It's directed - every node either **causes or is caused** by another.



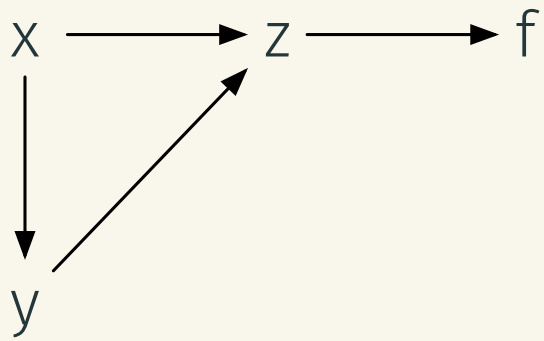
Variable Selection

It's acyclic - **no node** can be **it's own cause**.



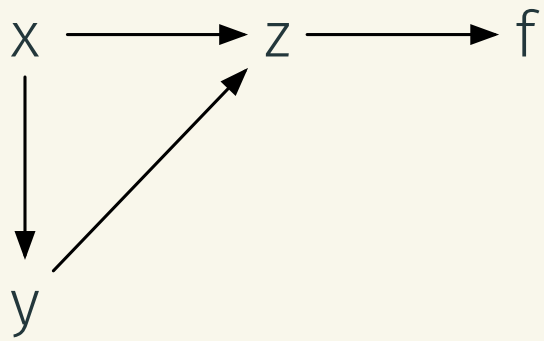
Variable Selection

Parents are nodes which cause other nodes. E.g. x is a parent to y and z.



Variable Selection

Children are nodes caused by other nodes. E.g. z is a child to x and y.

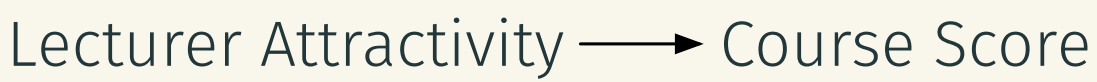


Questions?

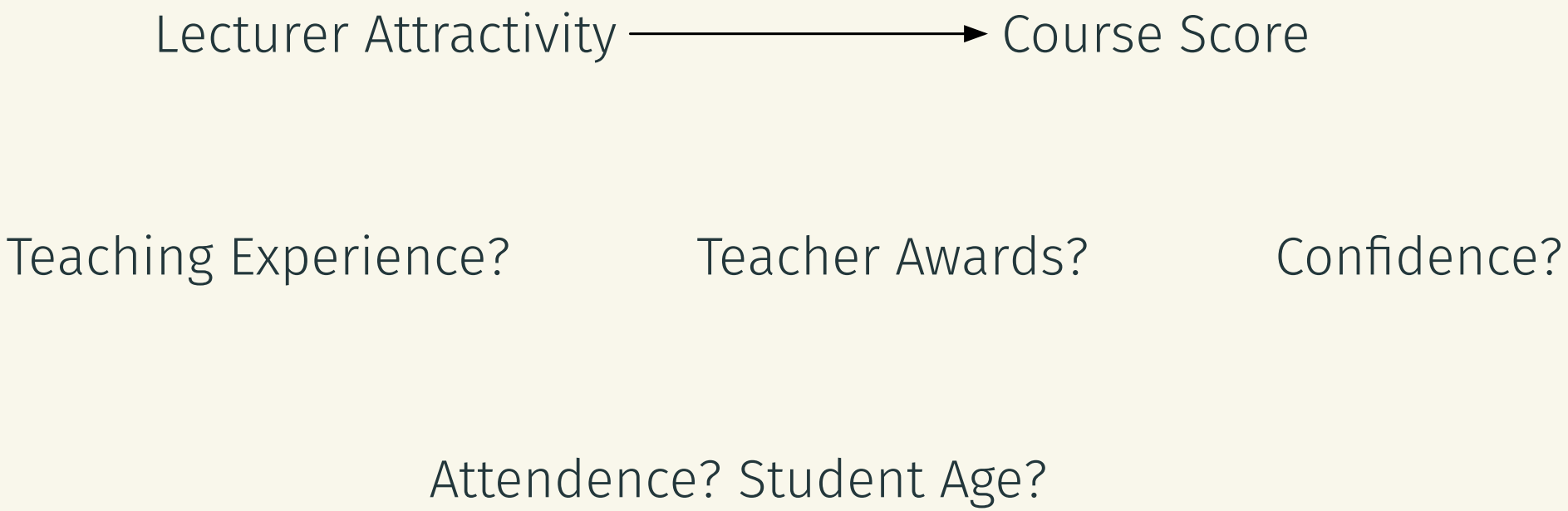
DAGs Example

DAGs Example

What is the causal effect of teacher attractivity on their course scores?



DAGs Example



DAGs Example

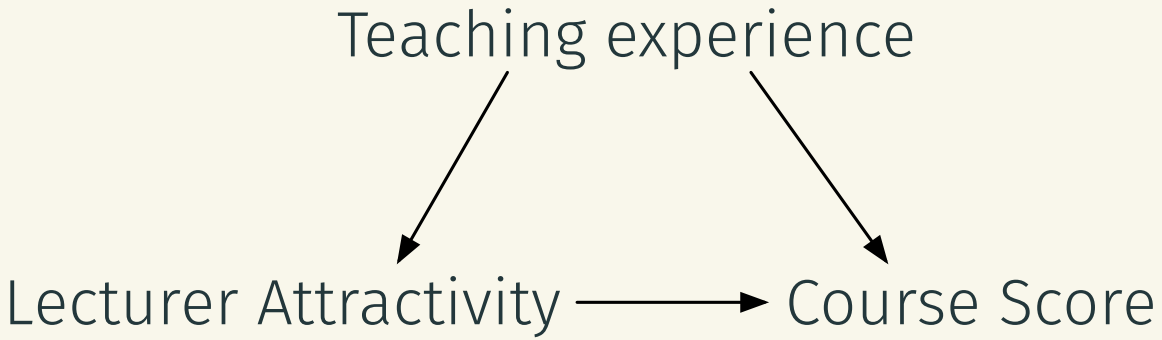
There four types of intervening variables.

- Confounders
- Colliders
- Mediators
- Moderators

Confounders

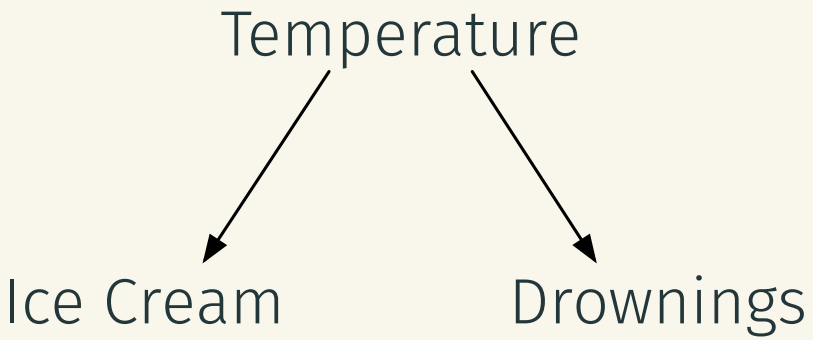
Confounders

Confounder is a **common parent** of two (or more) nodes.



Confounders

Confoundeers create correlation even without causal relationships. We **always control for colliders**.

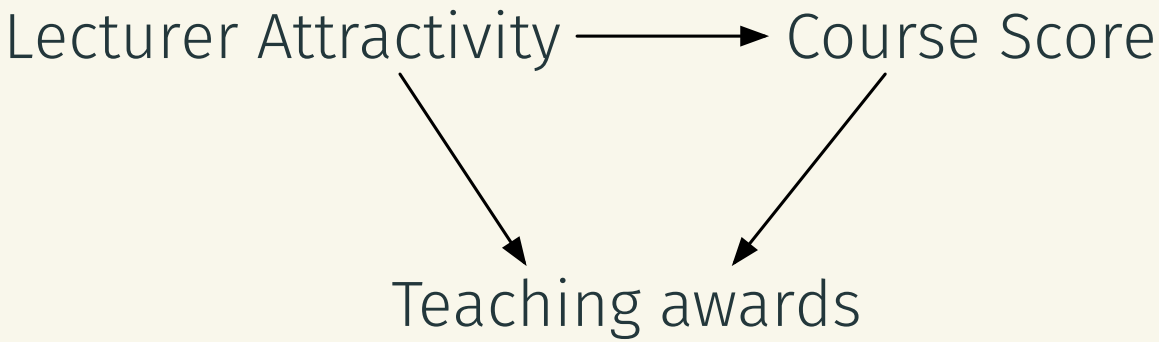


Questions?

Colliders

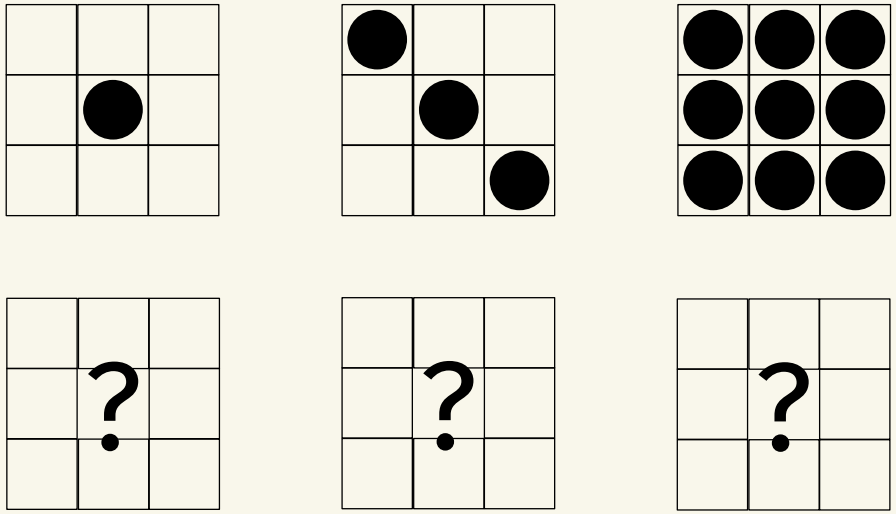
Colliders

Colliders are **common children** of two (or more) variables.



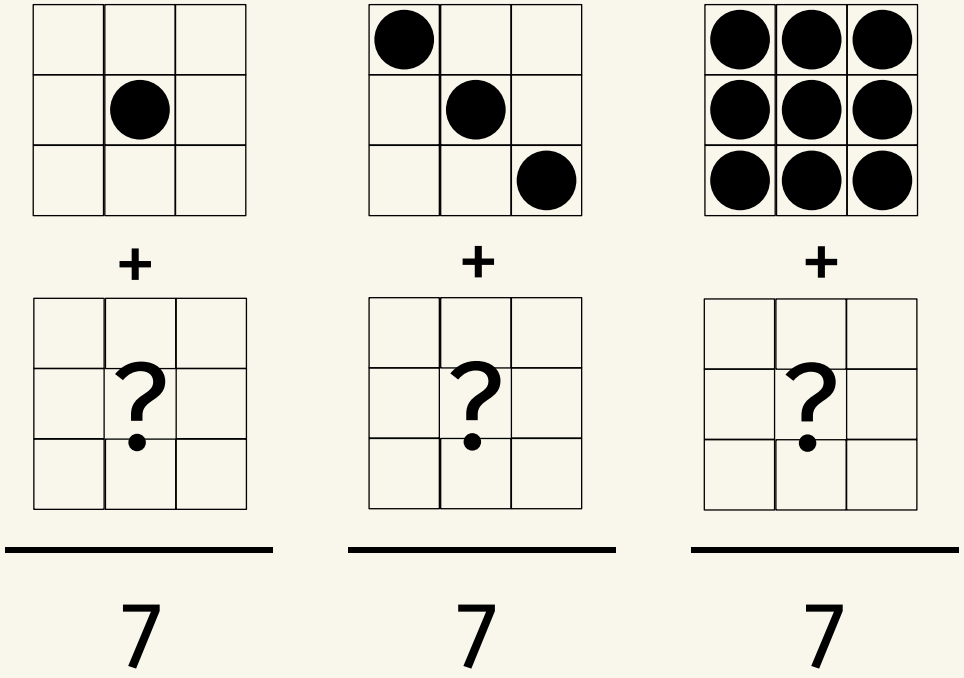
Colliders

We are throwing two dice. What is the value on the second one?



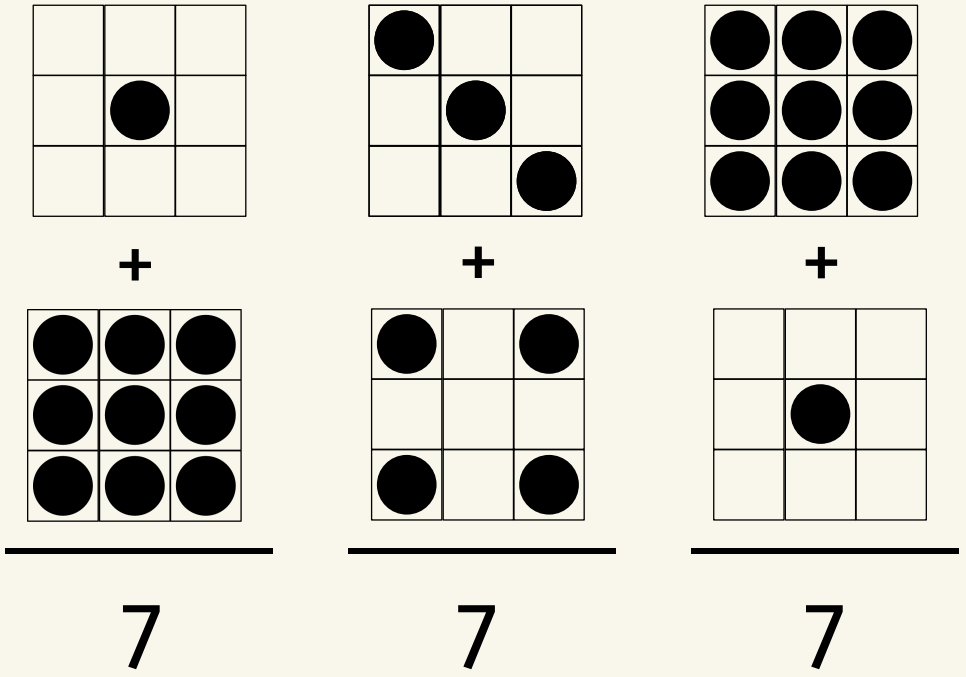
Colliders

We are throwing two dice. What is the value on the second one?



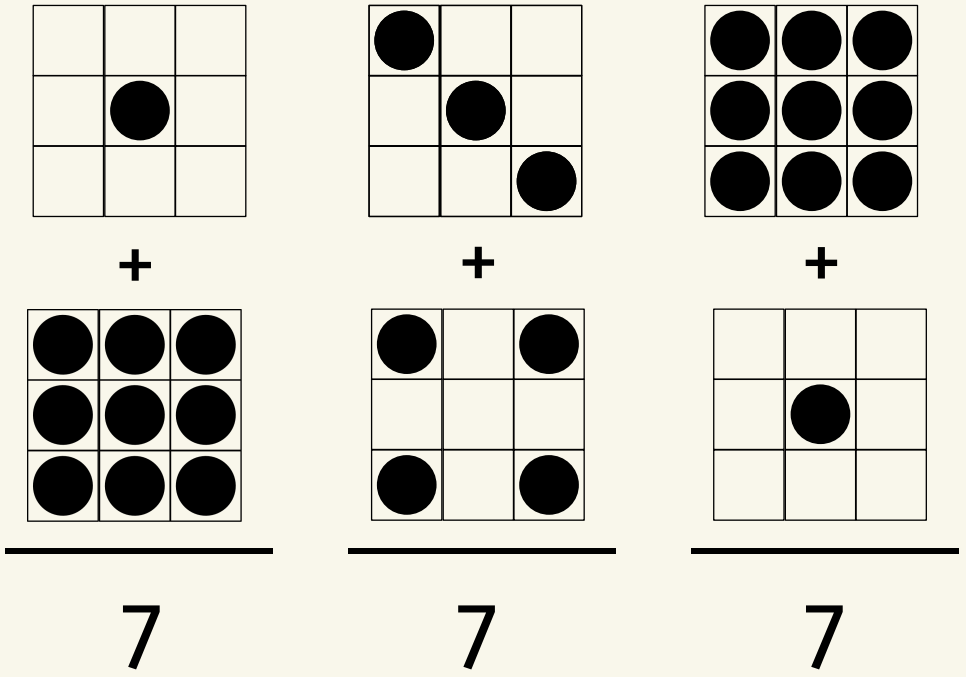
Colliders

Once we fix the sum of both die, we know that the higher the value on the first one, the lower value must be on the second.

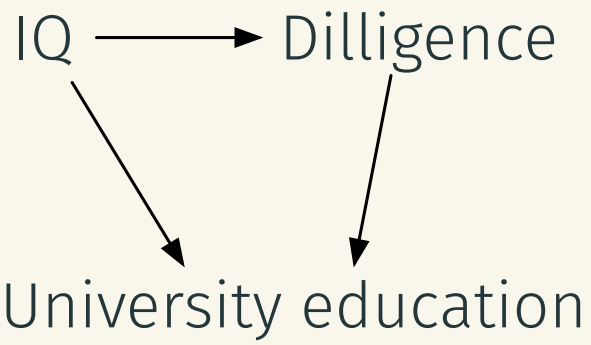


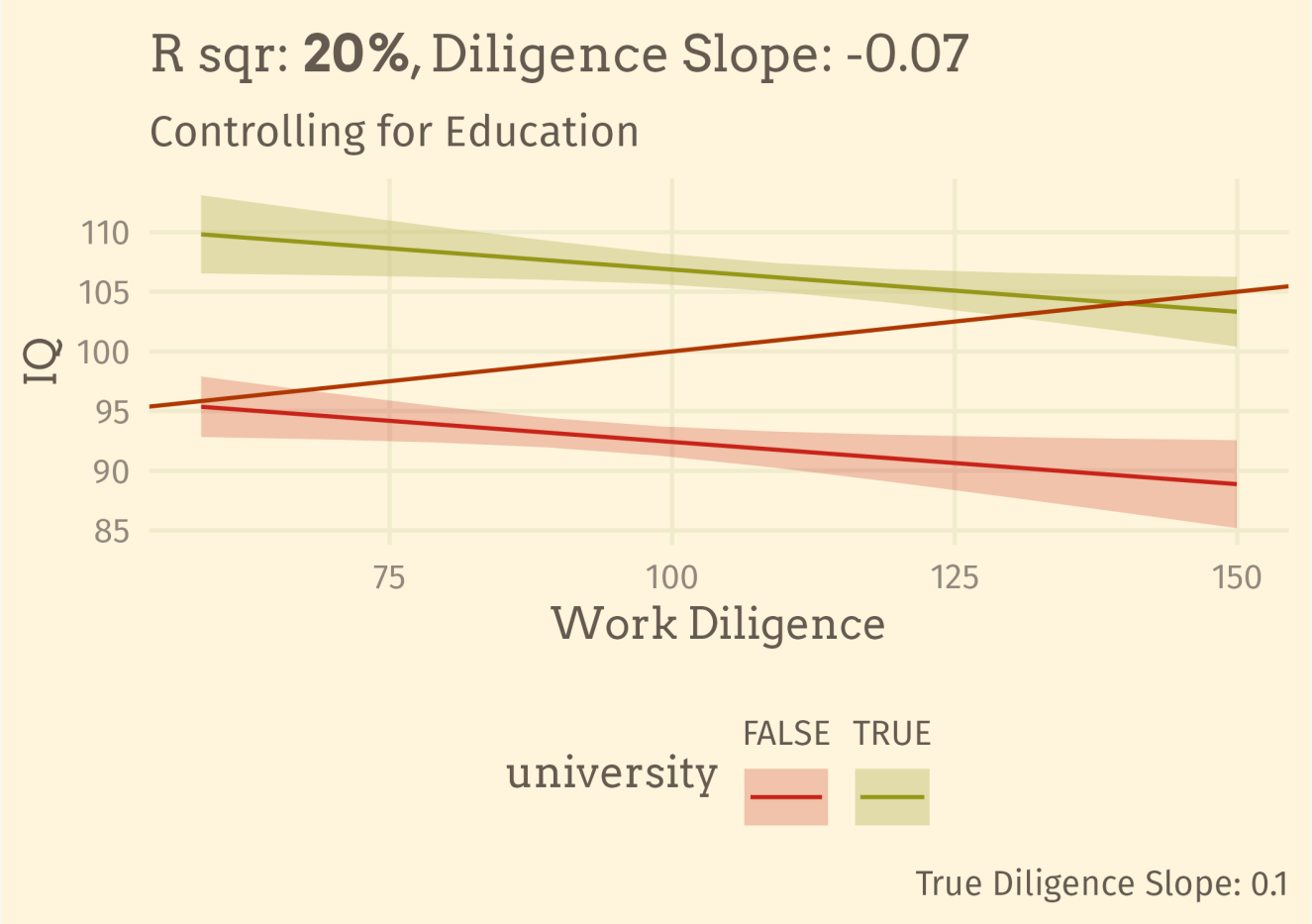
Colliders

More generally, **controlling for a collider creates a correlation** even if no causal relationship exists!

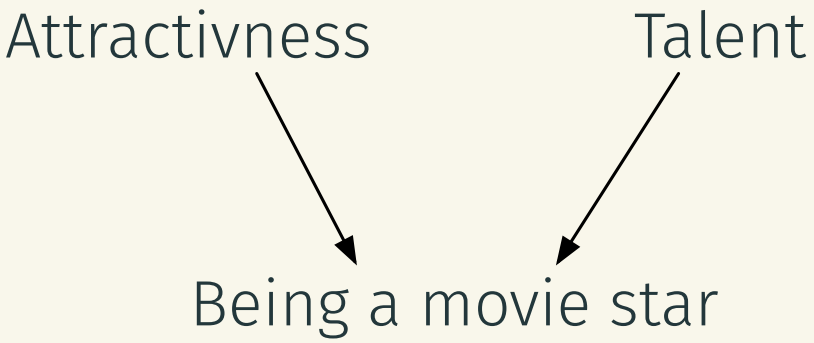


Colliders

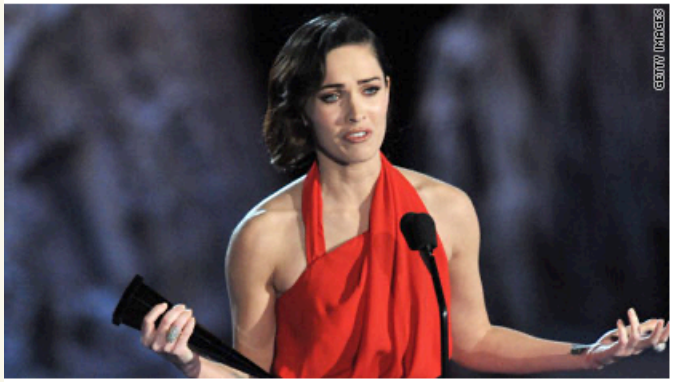

IQ
Dilligence
University education


R sqr: 20%, Diligence Slope: -0.07
Controlling for Education
IQ
110
105
100
95
90
85
75
100
125
150
Work Diligence
university
FALSE
TRUE
True Diligence Slope: 0.1

Colliders



Megan Fox voted worst - but sexiest - actress of 2009



Movie fans have mixed feelings about this year's "Transformers: Revenge of the Fallen."

On one hand, the action-packed sequel, which grossed more than $830 million in ticket sales worldwide, was voted the worst movie of the year by readers of the Web site Moviefone.com.

But in a bout of what can only be described as voter schizophrenia, "Transformers" won out in the poll's category for best action movie, beating "Star Trek," "Avatar" and "District 9."

"Transformers" leading lady Megan Fox was voted the sexiest star of 2009, but was also voted the actress who gave the worst performance.

This poll leads me to believe that moviegoers would have been just as happy if "Transformers" just had action-packed scenes of stuff blowing up along with Megan Fox in a non-speaking part. (Make your own joke here – ed.)

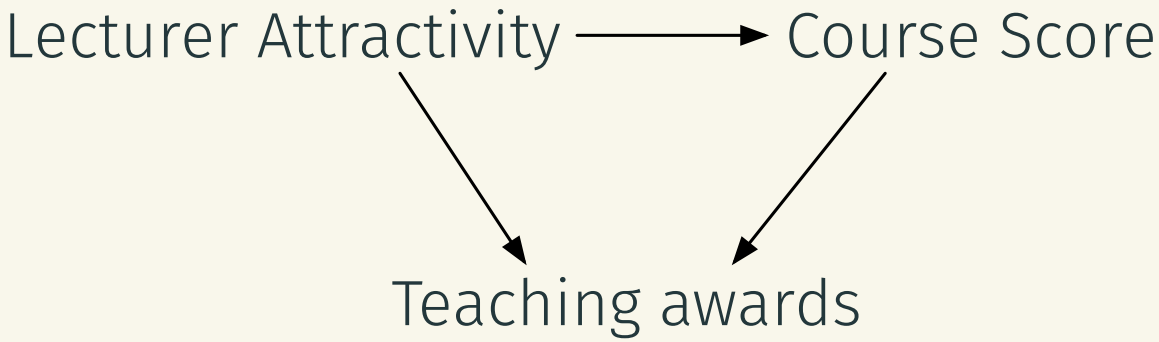
Fans offered no such ambivalence when it came to their love affair with "The Twilight Saga: New Moon." The vampire romance won out for the best movie of the year and best chick flick and piqued the interest of future movie-goers.

And the movie folks are most excited to see in 2010? "The Twilight Saga: Eclipse."

Post by: Jo Piazza, Special to CNN
Filed under: Megan Fox

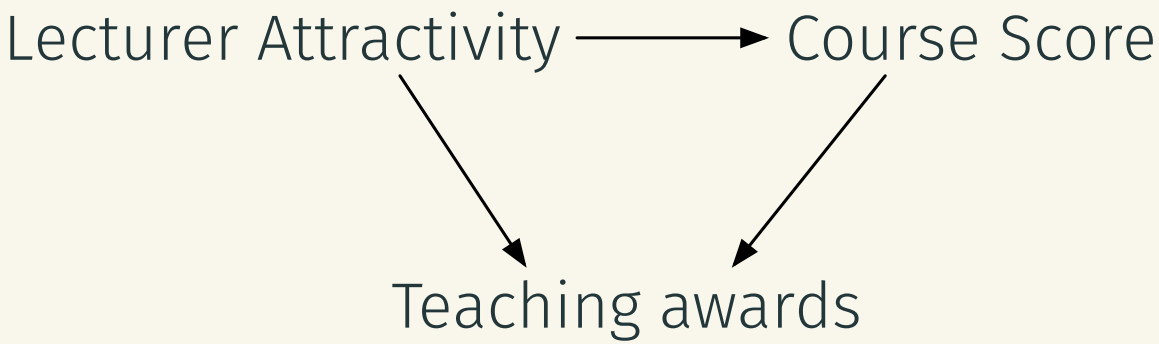
Colliders

To get a teaching award, lecturer needs to have great course score, be very attractive or both → Negative correlation when looking at awards winners.



Colliders

We **never to control for colliders**. It creates correlation that doesn't represent causal effects.



Questions?

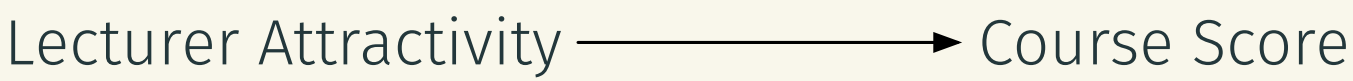
Mediators

Mediators

Mediator are **middle steps** on the causal path.

Mediators

We can either estimate the **total effect**:

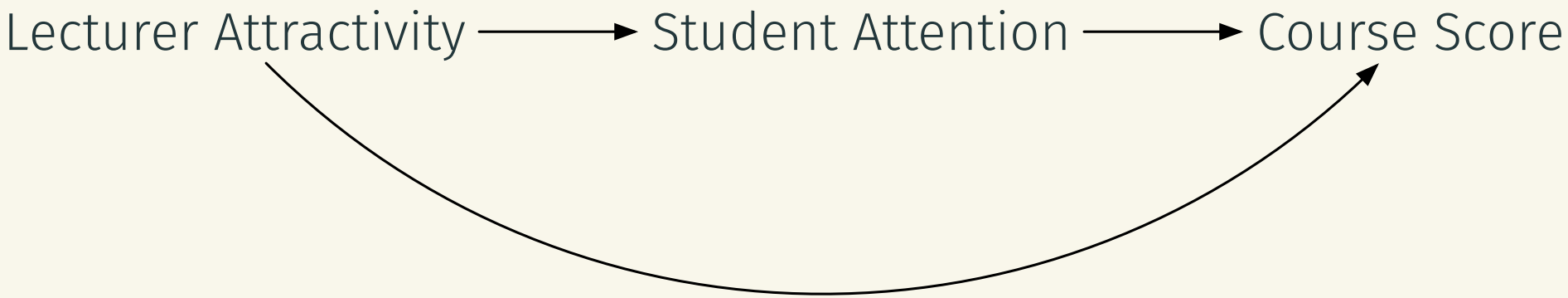


Or the **direct (partial) effect**:

Mediators

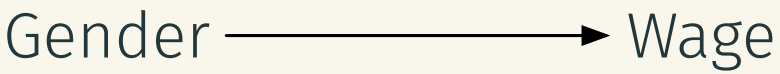
Adding mediator to our model isolate its effect.

It helps us answer „To what extend attractive lecturers have better course scores because students pay them more attention?“

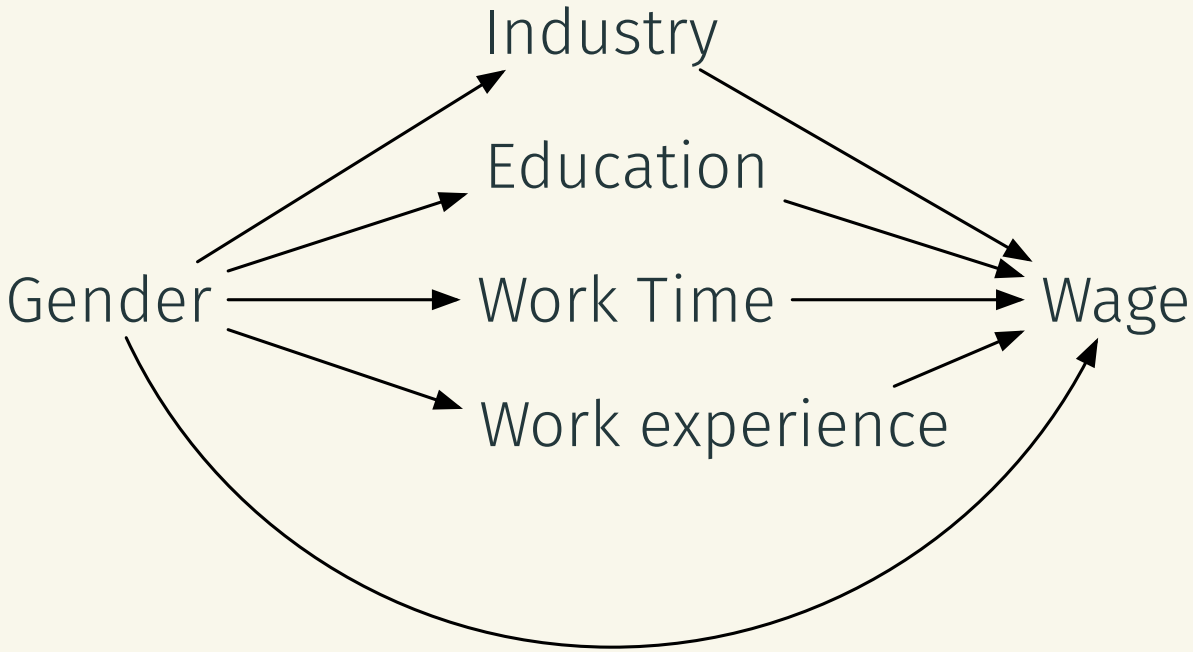


Mediators

Total Gender Gap
(without mediators)



Adjusted Gender Pay Gap
(with mediators)



Mediators

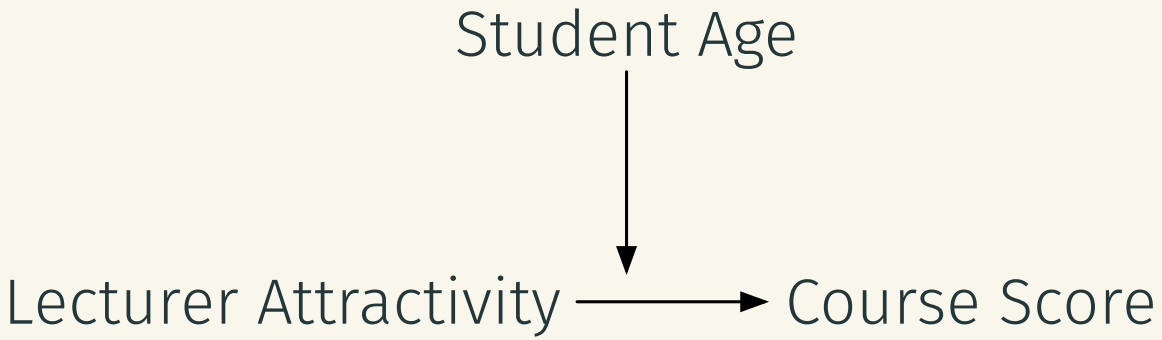
We control for mediators if we want to isolate a specific causal pathway.

Questions?

Moderators

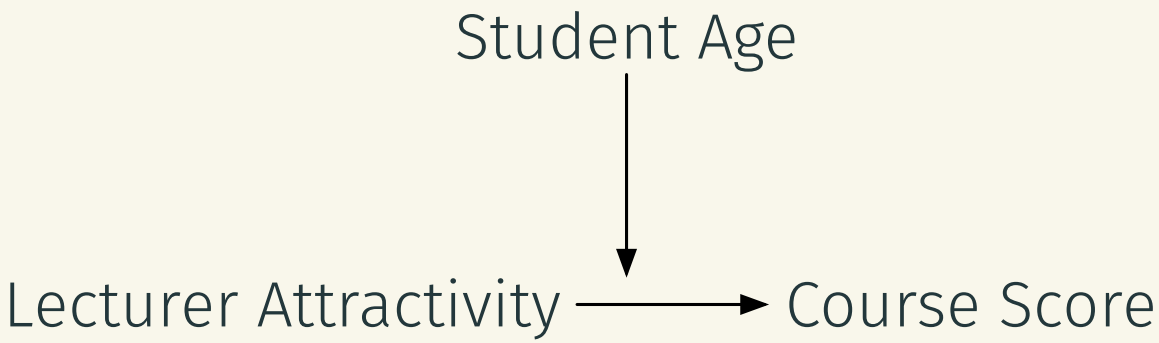
Moderators

Moderators are essentially interactions.



Moderators

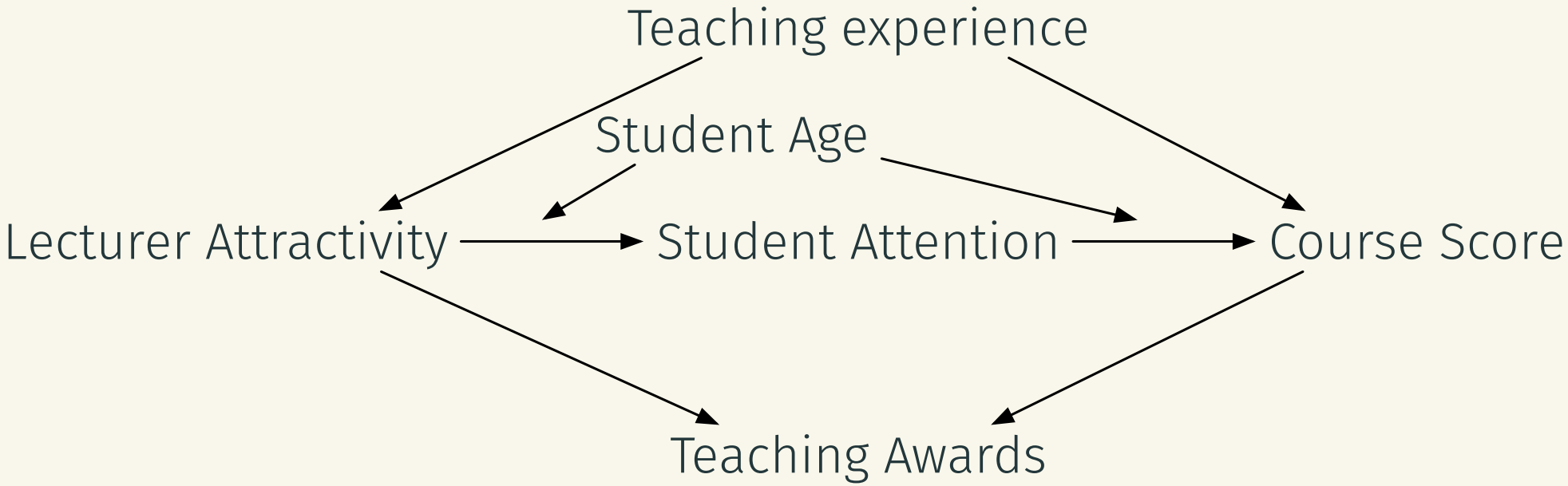
We add moderators if we want to estimate how does the effect of interest changes across subpopulations.



Questions?

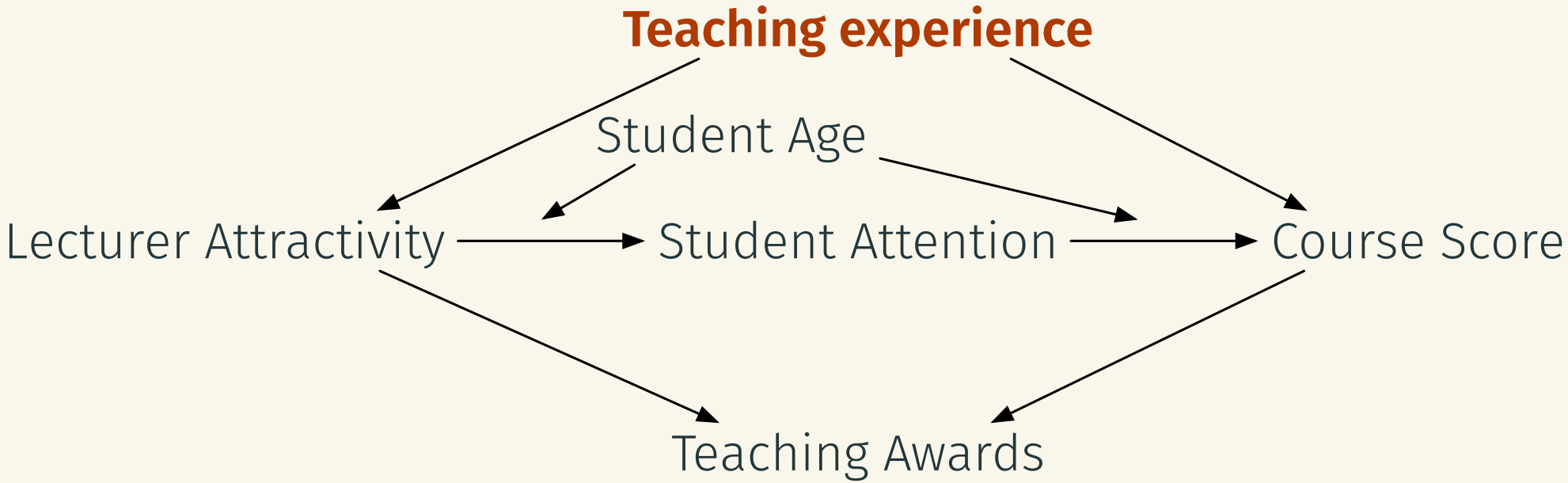
The Whole Game

The Whole Game



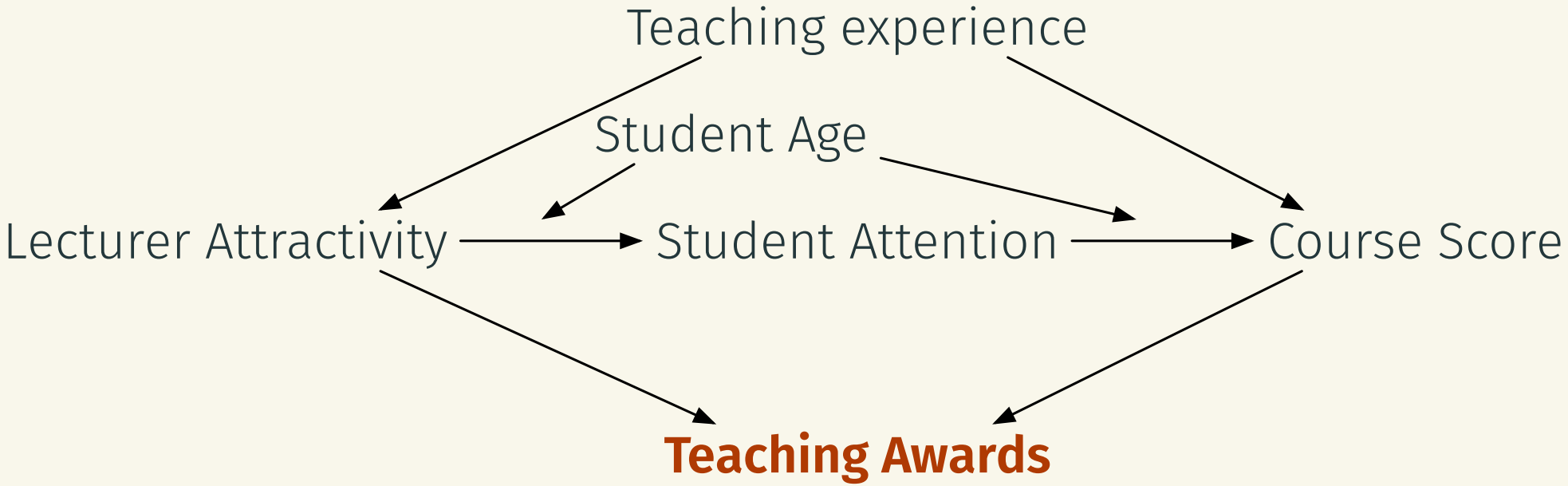
The Whole Game

Always control for **confounders** (common parents).



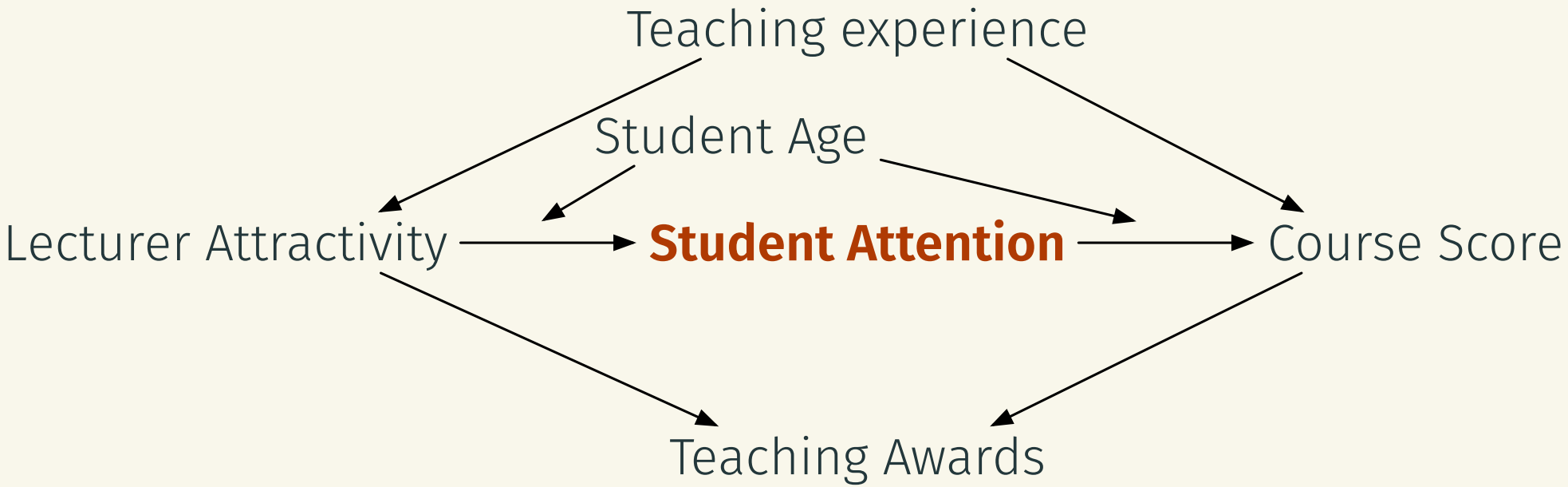
The Whole Game

Never control for **colliders**.

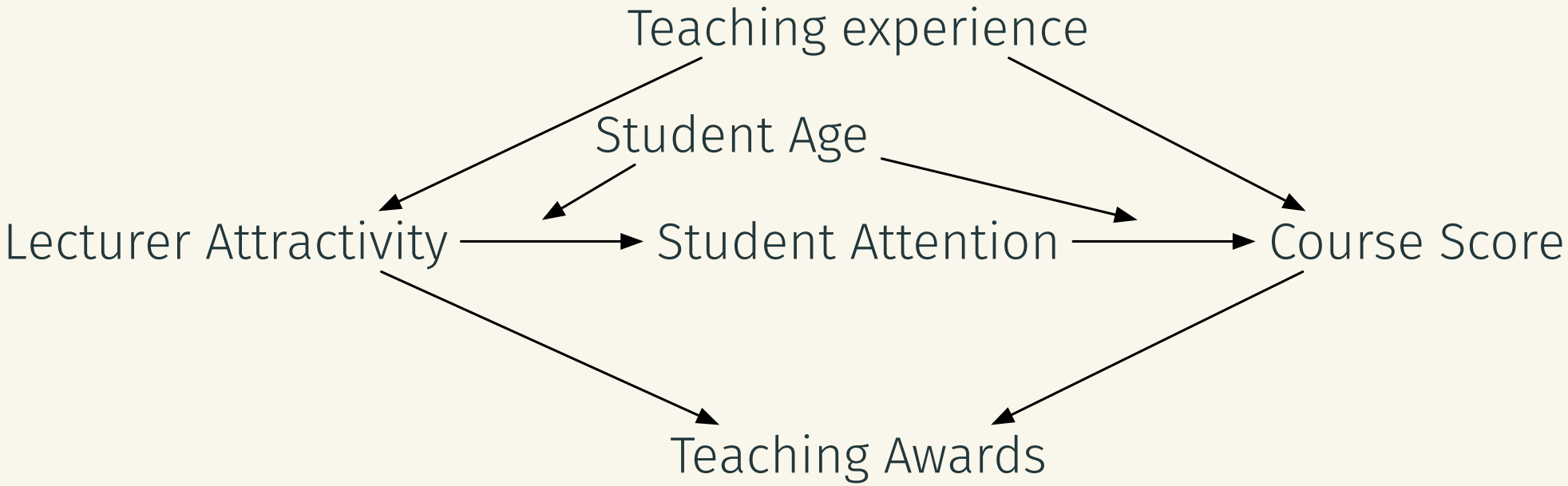


The Whole Game

Add **mediators** if you are interested in specific pathway/mechanism.

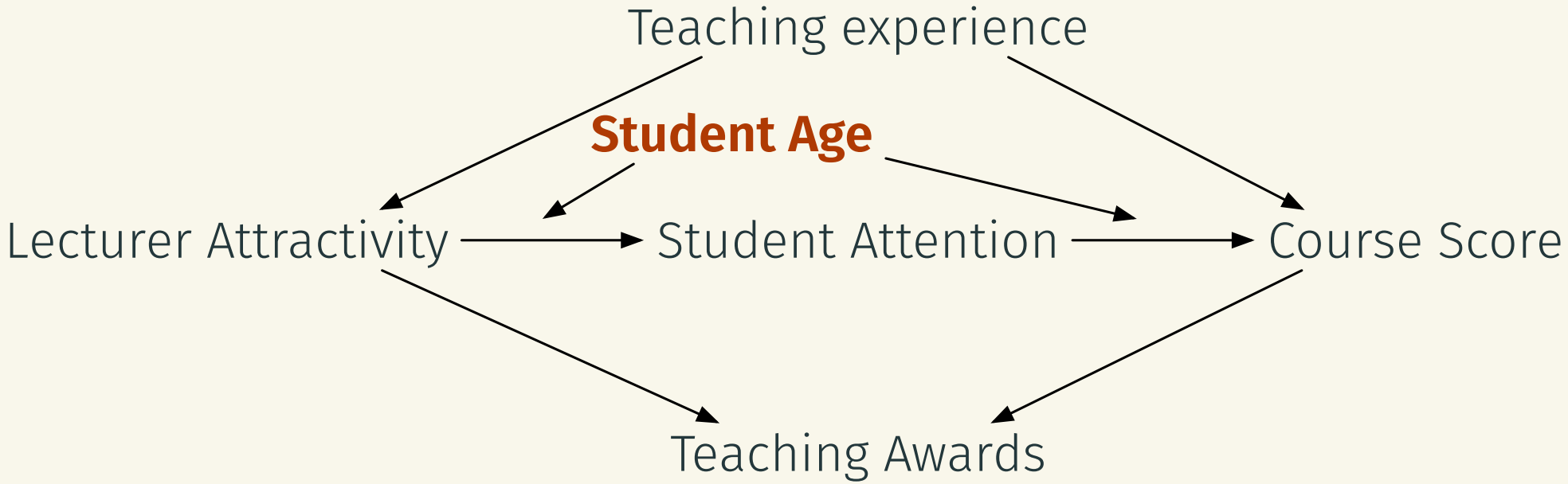


The Whole Game



The Whole Game

Add **moderators** (as interactions) if you want to check whether the effect changes across subpopulations.



Questions?

Let’s Try It!

Caveats

DAGs Are Not Magic (But They Are Useful)

DAGs Are Not Magic (But They Are Useful)

Just because you can draw a DAG, it doesn't mean it's correct.

But, it helps us think about and discuss our research.

They help prevent akward discussion („Why haven't you control for...")

The Need For Theory

The Need For Theory

You can't tell if a variable is confounder, collider, mediator or moderator based on correlations alone.

You need theory.

The Need For Theory

Research is to an extent a rhetoric excercise - you need to argue your world model is plausible.

If you convince your audience your assumptions make sense, they'll accept your results.

Controling By Filtering

Controling By Filtering

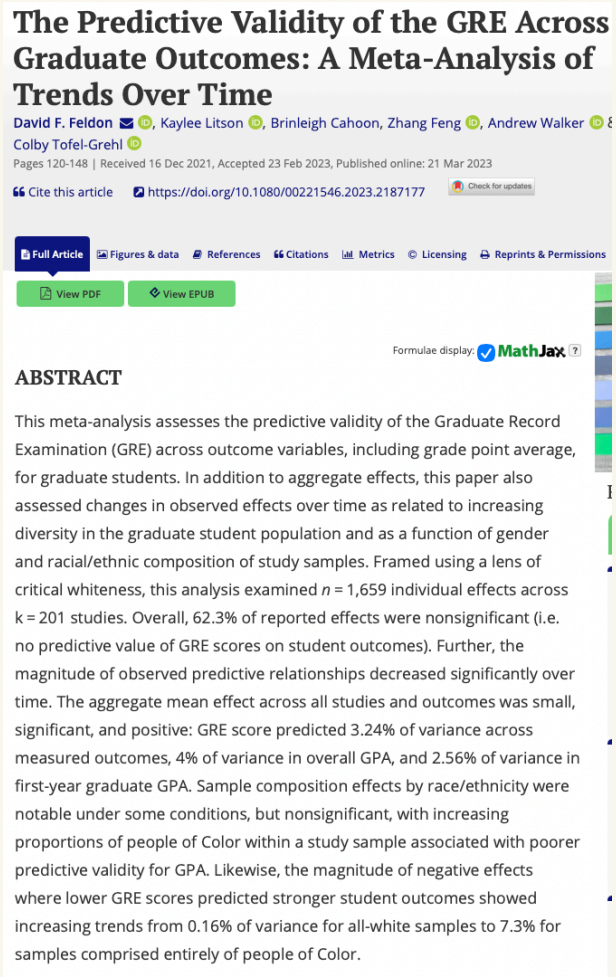
Controling for a variable means fixing it a specific value.

You can control for a variable by filtering your data. This isn't always a good thing!

Controling By Filtering

Many studies show there is no correlation between entrance exam scores and student perfomance.

Does it mean entrance exams are useless?



The Predictive Validity of the GRE Across Graduate Outcomes: A Meta-Analysis of Trends Over Time

David F. Feldon, Kaylee Litson, Brinleigh Cahoon, Zhang Feng, Andrew Walker, Colby Tofel-Grehl

Pages 120-148 | Received 16 Dec 2021, Accepted 23 Feb 2023, Published online: 21 Mar 2023

Cite this article https://doi.org/10.1080/00221546.2023.2187177

Full Article | Figures & data | References | Citations | Metrics | Licensing | Reprints & Permissions

View PDF | View EPUB

Formulae display: MathJax

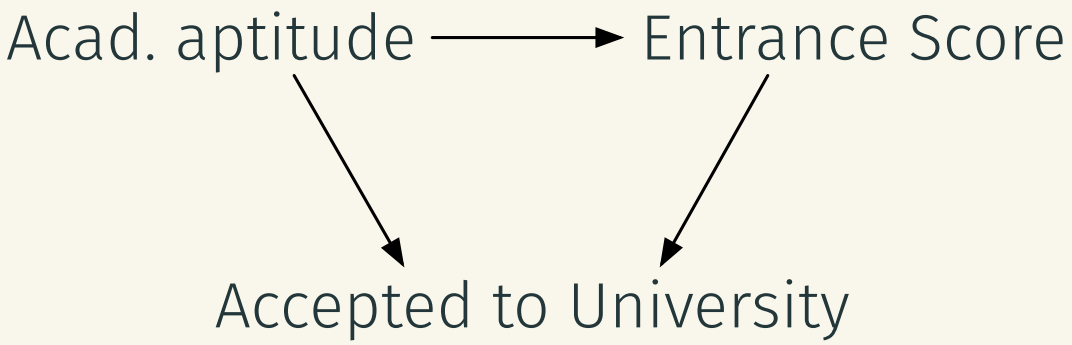
ABSTRACT

This meta-analysis assesses the predictive validity of the Graduate Record Examination (GRE) across outcome variables, including grade point average, for graduate students. In addition to aggregate effects, this paper also assessed changes in observed effects over time as related to increasing diversity in the graduate student population and as a function of gender and racial/ethnic composition of study samples. Framed using a lens of critical whiteness, this analysis examined $n = 1,659$ individual effects across $k = 201$ studies. Overall, 62.3% of reported effects were nonsignificant (i.e. no predictive value of GRE scores on student outcomes). Further, the magnitude of observed predictive relationships decreased significantly over time. The aggregate mean effect across all studies and outcomes was small, significant, and positive: GRE score predicted 3.24% of variance across measured outcomes, 4% of variance in overall GPA, and 2.56% of variance in first-year graduate GPA. Sample composition effects by race/ethnicity were notable under some conditions, but nonsignificant, with increasing proportions of people of Color within a study sample associated with poorer predictive validity for GPA. Likewise, the magnitude of negative effects where lower GRE scores predicted stronger student outcomes showed increasing trends from 0.16% of variance for all-white samples to 7.3% for samples comprised entirely of people of Color.

Controling By Filtering

Probably not. These studies only include accepted students.

But if a student was accepted, they are either good at standardized tests or just good at studying in general

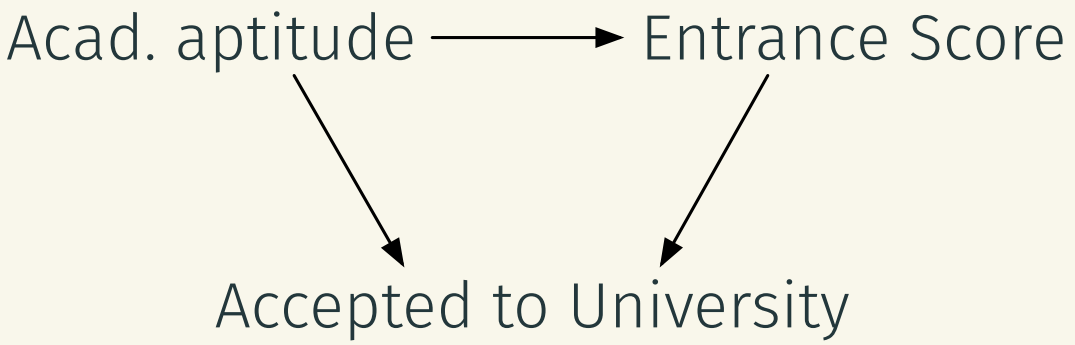


Questions?

Controling By Filtering

Being accepted is a **collider** on path between academic aptitude and entrance score.

By looking only at the accepted students, we are biasing the results.

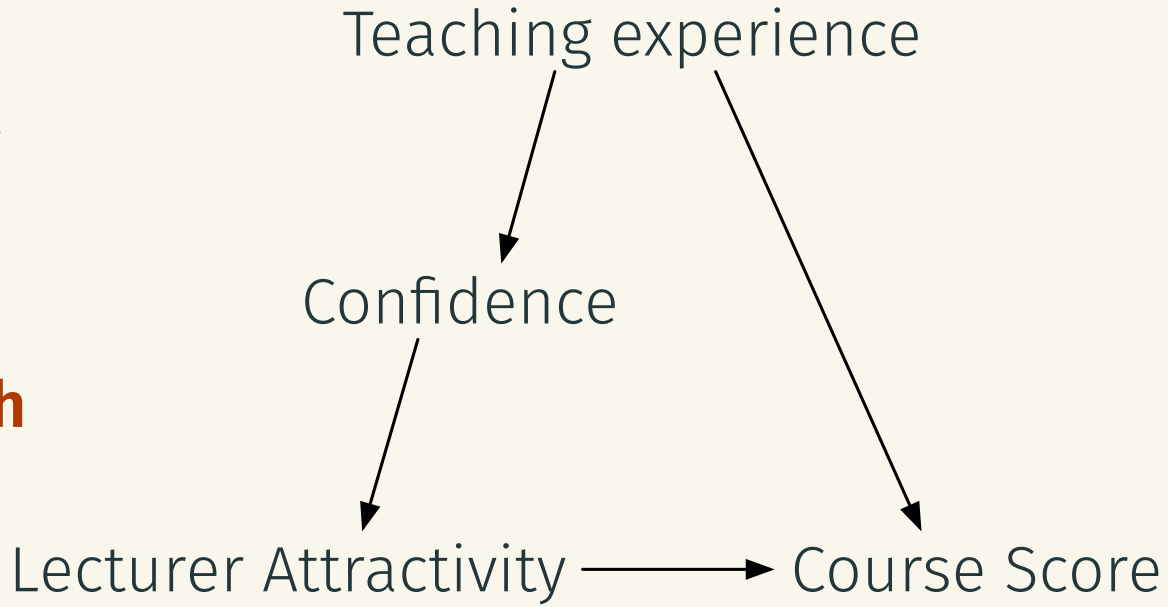


Indirect Relationships

Indirect Relationships

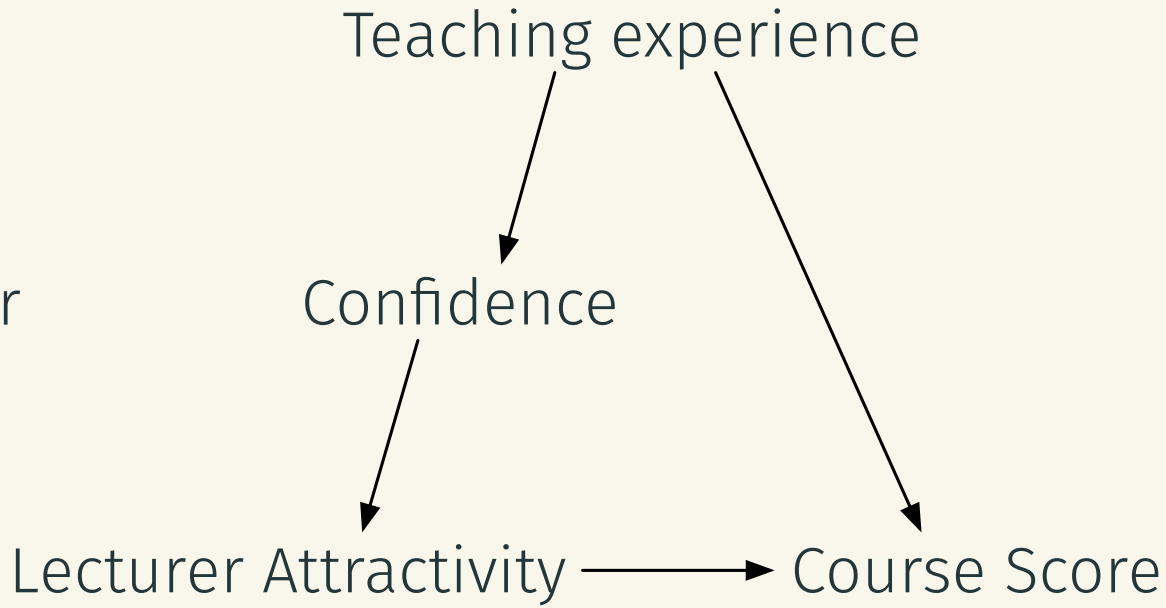
Bad news - You need to **account for indirect relationships**.

Good news - you only need to account for **one variable on each path**.



Indirect Relationships

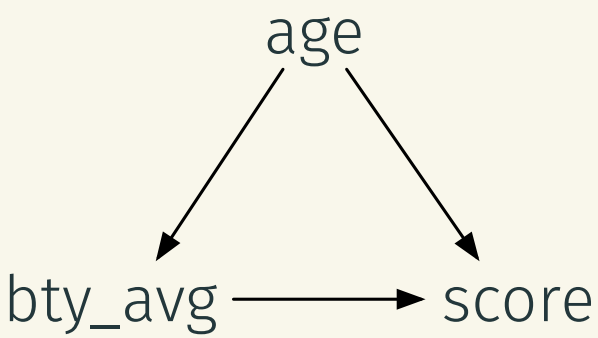
In this case, we need to control for **either** teaching experience or confidence.



Avoid Table 2 Fallacy

Avoid Table 2 Fallacy

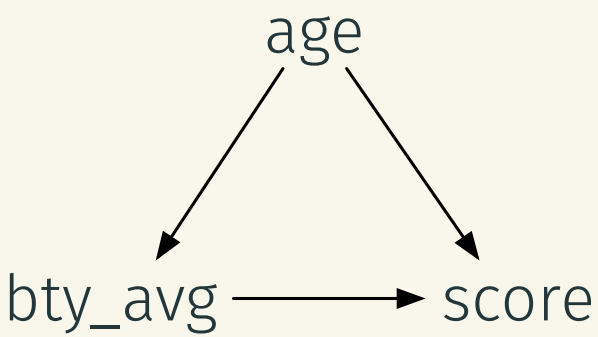
We are interested in the effect of `bty_avg` on `score`. We assume this model:



Parameter	Coefficient
(Intercept)	4.05473
bty_avg	0.06066
age	-0.00306

Avoid Table 2 Fallacy

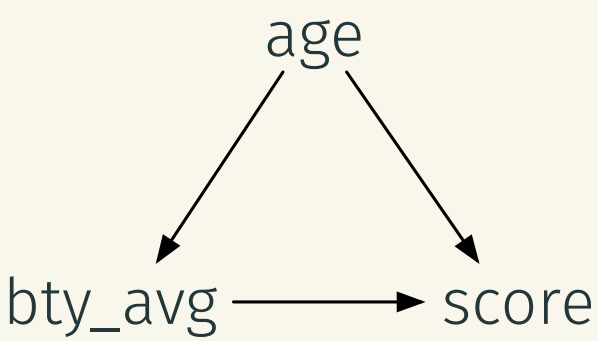
The `bty_avg` coefficient represent the total effect. `age` is a collider.



Parameter	Coefficient
(Intercept)	4.05473
bty_avg	0.06066
age	-0.00306

Avoid Table 2 Fallacy

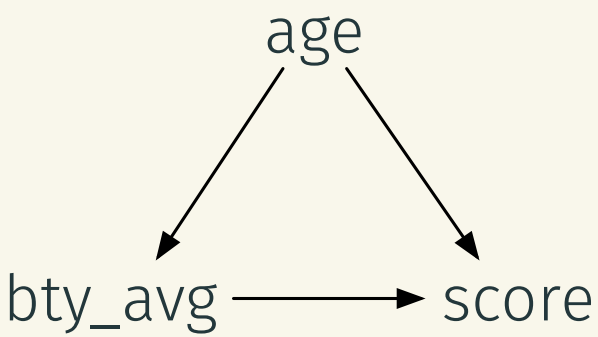
Now we are interested in the effect of `age` on `score`. Does its coefficient represent the same kind of effect?



Parameter	Coefficient
(Intercept)	4.05473
bty_avg	0.06066
age	-0.00306

Avoid Table 2 Fallacy

No! `bty_avg` is a mediator on the path from `age` to `score`.



Parameter	Coefficient
(Intercept)	4.05473
bty_avg	0.06066
age	-0.00306

Avoid Table 2 Fallacy

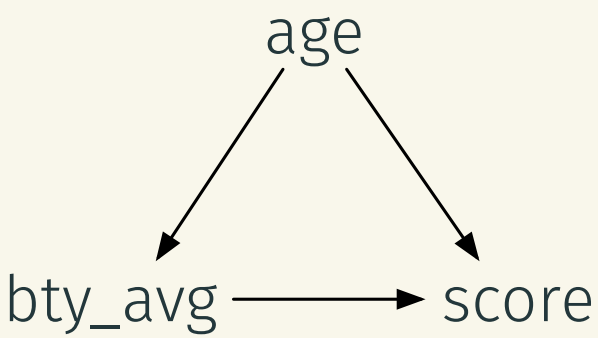
(Wrongly) **assuming all regression coefficients represent total effects** is known as **Table 2 fallacy**.

(Because in many papers, second table is the one with regression coefficients.)

In reality, we often need to fit separate models for each predictors we are interested in.

Avoid Table 2 Fallacy

Question: How should we change our model to estimate total effect of `age`?



Parameter	Coefficient
(Intercept)	4.05473
bty_avg	0.06066
age	-0.00306

Questions?