Agenda

1. **Interacting variables in regression**
2. **Causal analysis in regression**
3. **Mediation, moderation, confounding, and collision**
4. **Building indicator (dummy) variables in R**

Interacting dummies

$$\log(\text{Inc}_i) \sim \text{Norm}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 W_i + \beta_2 A_i$$
$$\alpha, \beta_1, \beta_2 \sim \text{Norm}(0, 30)$$
$$\sigma \sim \text{Unif}(0, 50)$$

W_i
Indicator variable for women

A_i
Indicator variable for respondents over 35 years old

	Mean	Std. Dev.	5%	95%
α	9.87	0.04	9.81	9.94
β_1	–0.48	0.04	–0.55	–0.42
β_2	0.70	0.04	0.62	0.77
σ	1.16	0.01	1.14	1.18

β_1 : exp(–0.48) ≈ 0.62

(women make about 62% as much as men, on average)

β_2 : exp(0.70) ≈ 2.01

(people over 35 years old make about twice as much as people 35 and under)

Interacting dummies

$$\log(\text{Inc}_i) \sim \text{Norm}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 W_i + \beta_2 A_i + \beta_3 W_i A_i$$
$$\alpha, \beta_1, \beta_2, \beta_3 \sim \text{Norm}(0, 30)$$
$$\sigma \sim \text{Unif}(0, 50)$$

$W_i A_i$
Interaction between both indicators

	Mean	Std. Dev.	5%	95%
α	9.82	0.05	9.74	9.91
β_1	–0.38	0.07	–0.50	–0.26
β_2	0.77	0.06	0.67	0.87
β_3	–0.15	0.09	-1.29	-0.01
σ	1.16	0.01	1.14	1.18

Interacting dummies

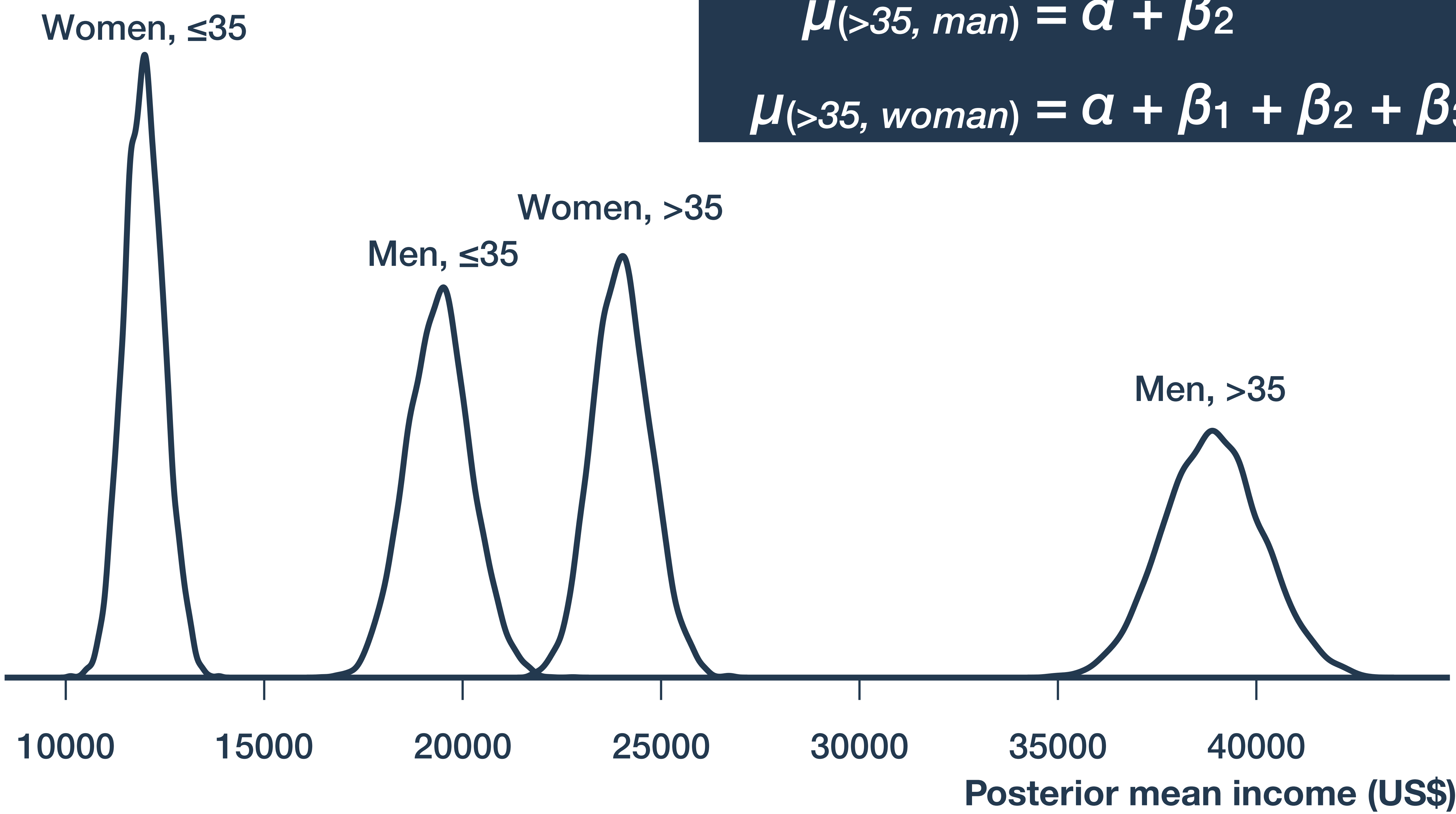
$$\mu_i = \alpha + \beta_1 W_i + \beta_2 A_i + \beta_3 W_i A_i$$

$$\mu_{(\leq 35,\ man)} = \alpha$$

$$\mu_{(\leq 35,\ woman)} = \alpha + \beta_1$$

$$\mu_{(>35,\ man)} = \alpha + \beta_2$$

$$\mu_{(>35,\ woman)} = \alpha + \beta_1 + \beta_2 + \beta_3$$



Interacting dummies

$$\mu_i = \alpha + \beta_1 W_i + \beta_2 A_i + \beta_3 W_i A_i$$

$$\mu_{(\leq 35,\ man)} = \alpha$$

$$\mu_{(\leq 35,\ woman)} = \alpha + \beta_1$$

$$\mu_{(>35,\ man)} = \alpha + \beta_2$$

$$\mu_{(>35,\ woman)} = \alpha + \beta_1 + \beta_2 + \beta_3$$

	Mean	exp(Mean)
α	9.82	**18398.051**
β_1	-0.38	**0.684**
β_2	0.77	**2.16**
β_3	-0.15	**0.861**

Interpreting the interaction coefficient β_3

The pay benefit of being over 35 (β_2) is diminished by about 14% for women (β_3).

OR

The pay gap for women (β_1) is exacerbated by about 14% for those over 35 (β_3).

Interacting continuous variables

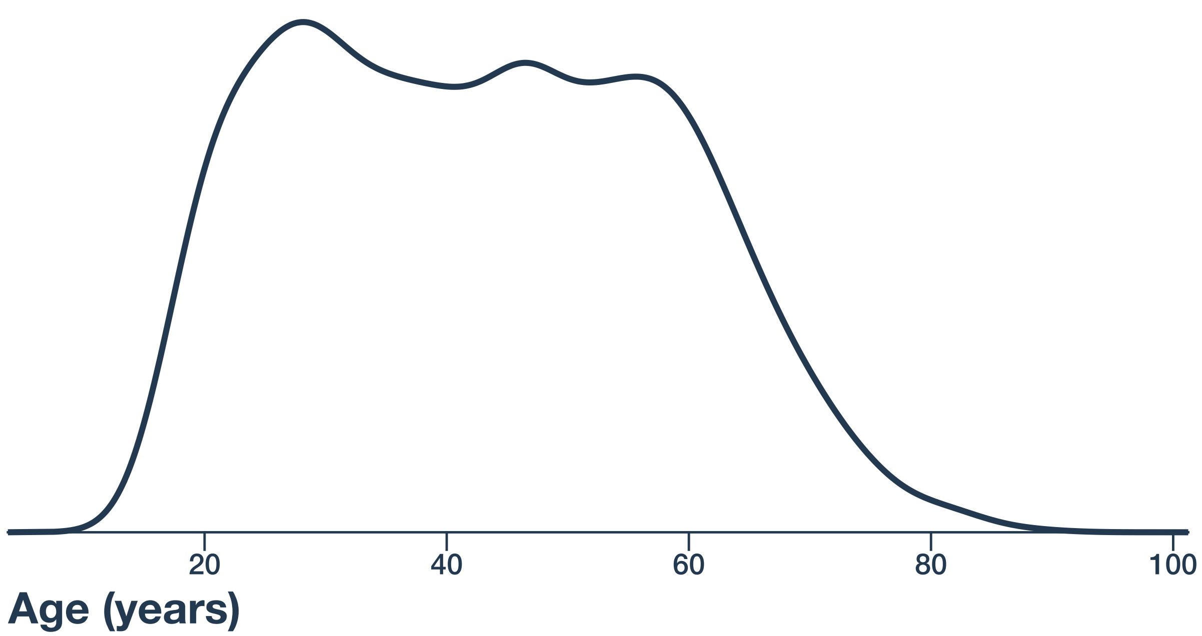
$$\log(\mathrm{Inc}_i) \sim \mathrm{Norm}(\mu_i, \sigma)$$

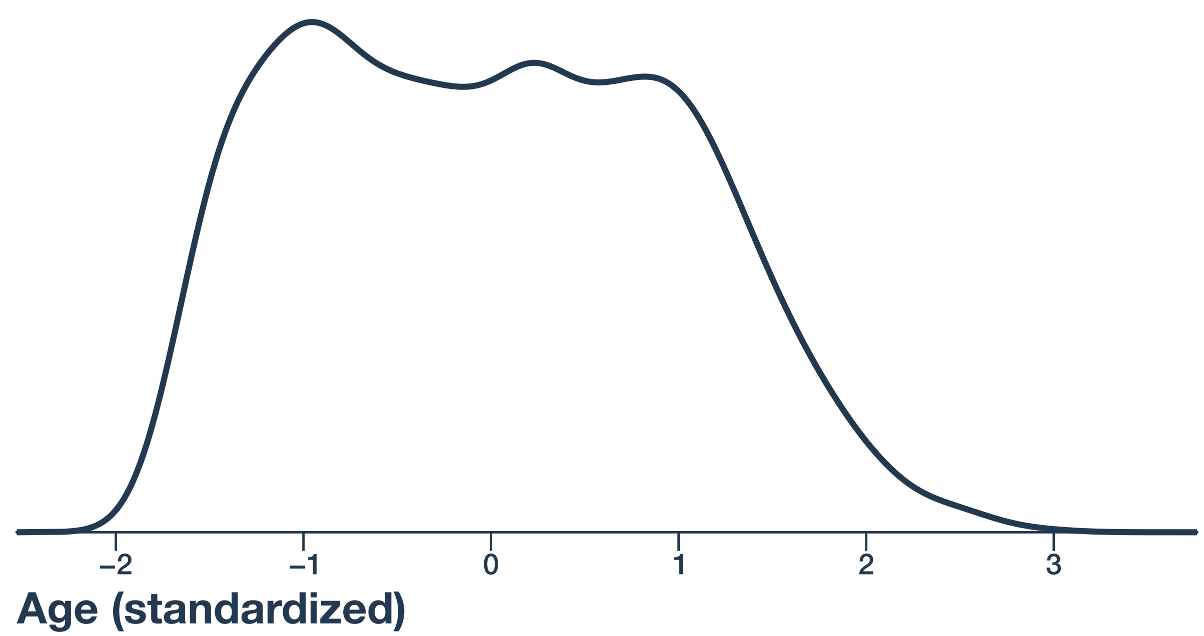
$$\mu_i = \alpha + \beta_1 \mathrm{Occ}_i + \beta_2 \mathrm{Age}_i + \beta_3 \mathrm{Occ}_i \mathrm{Age}_i$$

Occupational income index (standardized)

Age (standardized)

Standardization: Transforming a variable X to so that **mean(X)=0** and **sd(X)=1**





Interacting continuous variables

$$\mu_i = \alpha + \beta_1 \text{Occ}_i + \beta_2 \text{Age}_i + \beta_3 \text{Occ}_i \text{Age}_i$$

	Mean	exp(Mean)
α	10.25	**28282.542**
β_1	0.48	**1.616**
β_2	0.35	**1.419**
β_3	-0.05	**0.951**

Interpreting the interaction coefficient β_3

The pay benefit of being in a high-prestige job (β_1) is diminished by about 5% for each one standard deviation increase in age (β_3).

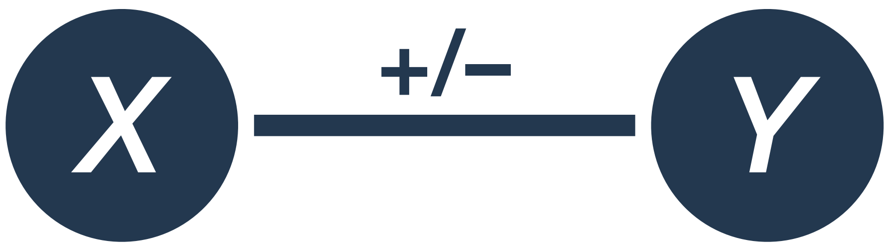
OR

The pay benefit of being older (β_2) is diminished by about 5% for each one standard deviation increase in occupational prestige (β_3).

Causal analysis

Causal question: Does a change in one variable (*X*) *cause* a change in another (*Y*)?

Regression only identifies statistical relationships, not causal relationships



To draw a "causal arrow" you need theory

Causal analysis

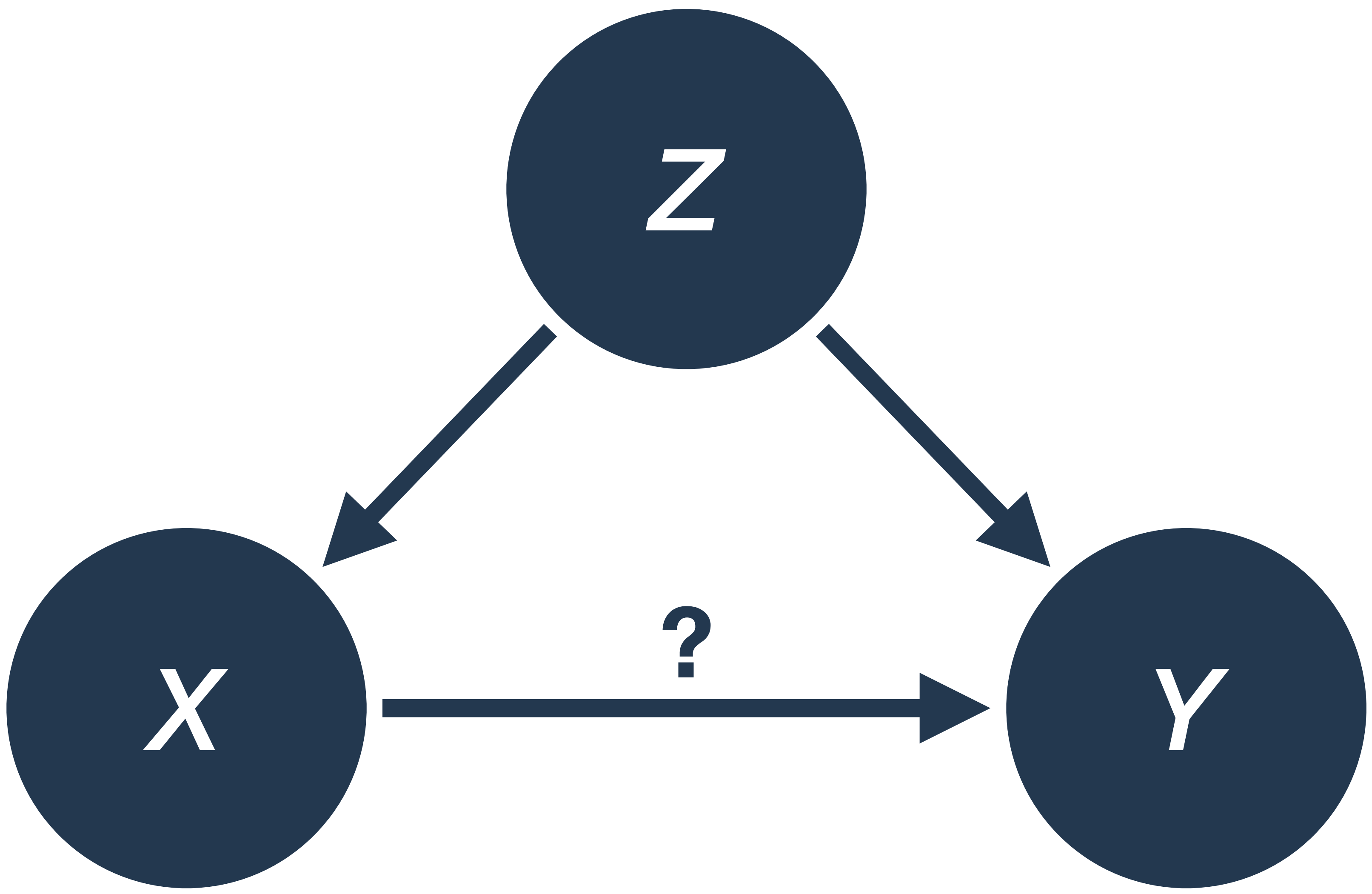
To establish a causal relationship you (usually) need

1. **Causal precedence**
   A theoretical reason to believe changes in *X* could affect *Y* (e.g. *X* precedes *Y* in time)

2. **Statistical association**
   An established statistical association between *X* and *Y* (e.g. a convincing coefficient estimate)

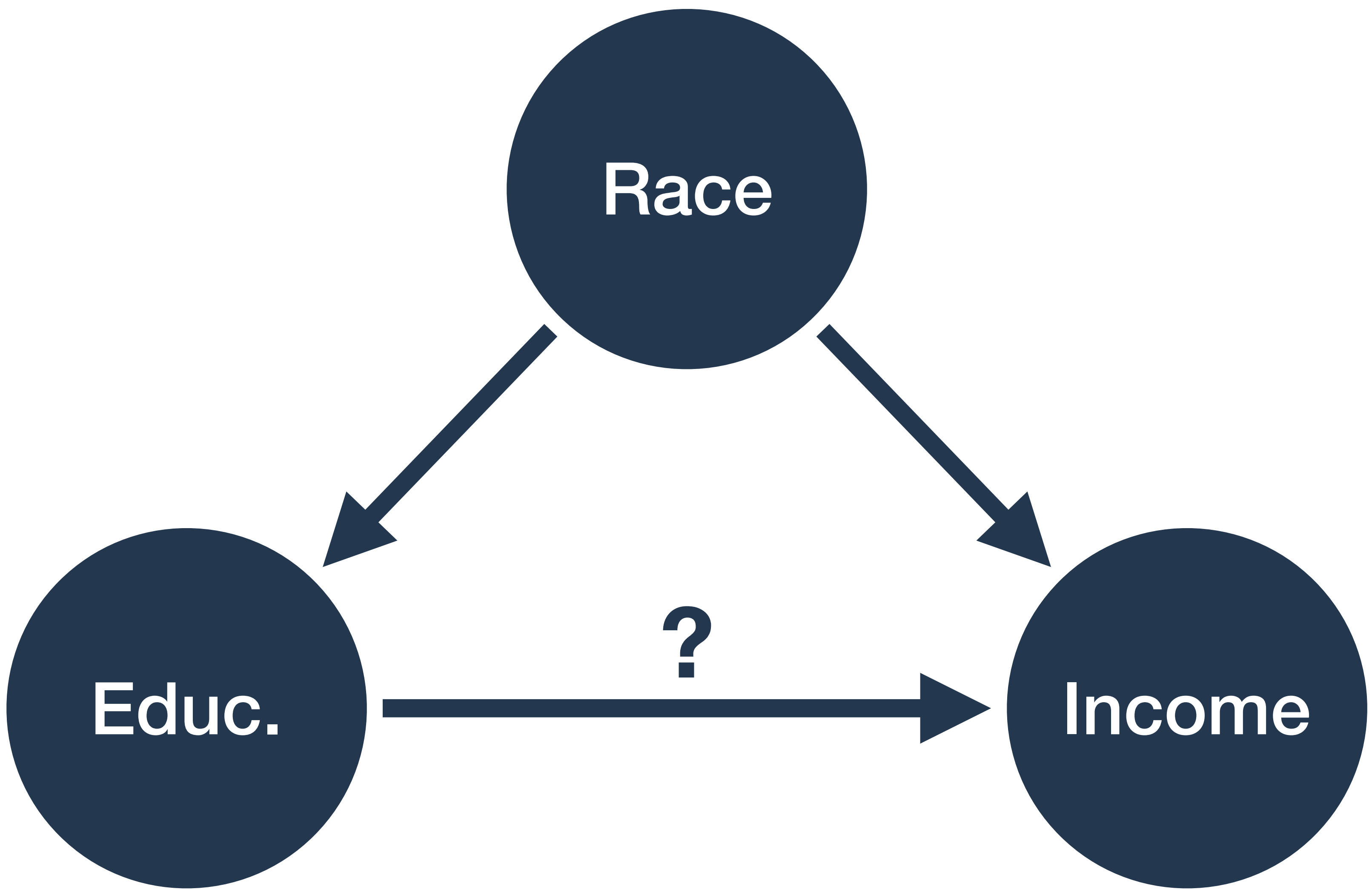
3. **No unaccounted-for confounders**
   No other variables, observed or otherwise, that *confound* the association between *X* and *Y*

Confounding variables



A variable Z is a **confounder** of the relationship between X and Y if Z is a causal influence on both X and Y

Confounding variables

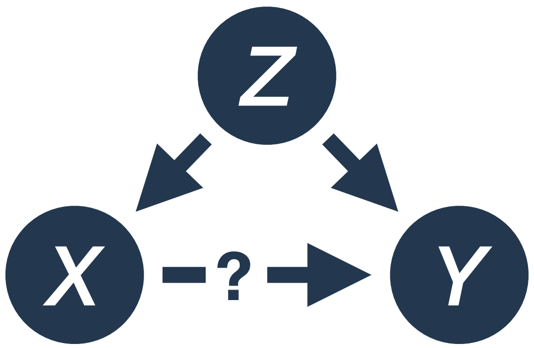


A variable *Z* is a **confounder** of the relationship between *X* and *Y* if *Z* is a causal influence on both *X* and *Y*

For example:
To establish a causal relationship between education and income, you need to account for race, which could affect both education and income

Types of covariates

Confounder

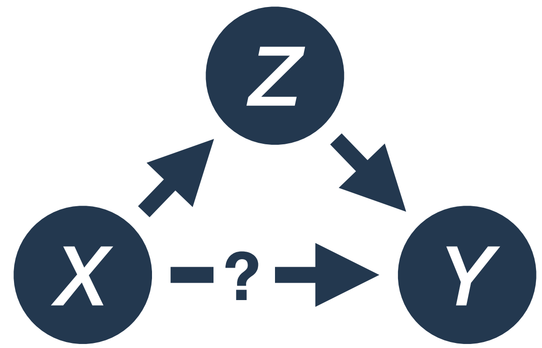


Z is a causal factor on both *X* and *Y*.

Must be "controlled for" to establish non-spurious relationship between *X* and *Y*.

E.g.:
Race confounds the relationship between education and income.

Mediator

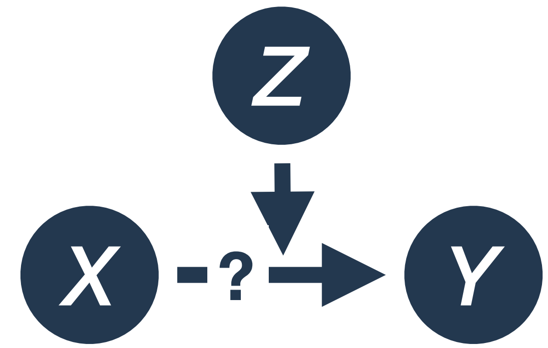


Z is influenced by *X* and influences *Y*.

Including as covariate elaborates on relationship between *X* and *Y*.

E.g.:
Occupation mediates the relationship between gender and income.

Moderator

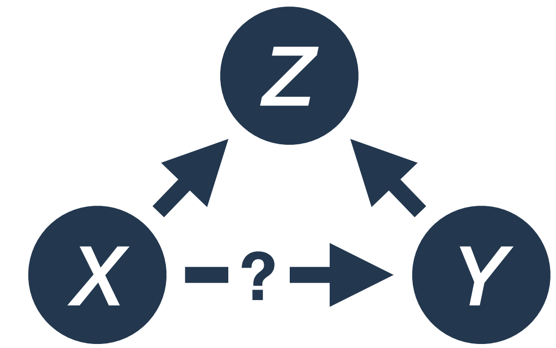


Z alters the relationship between *X* and *Y*.

Can be included as interaction variable to better describe the relationship between *X* and *Y*.

E.g.:
Marital status moderates the relationship between gender and income.

Collider



Z causally influenced by both *X* and *Y*.

Must *not* be "controlled for" when establishing relationship between *X* and *Y*.

E.g.:
Income is a collider for the relationship between gender and occupation.