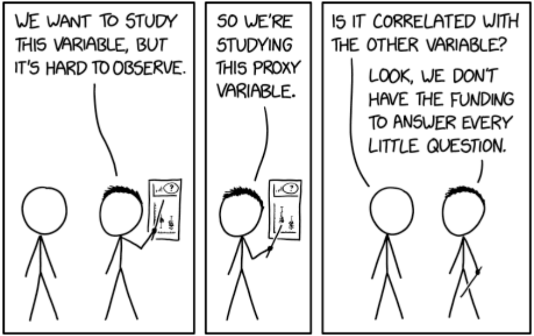
What are proxy variables?



Source: https://xkcd.com/2652/

Suppose we have a simple linear regression model

$$y = X\beta + \varepsilon$$

Suppose we have a simple linear regression model

$$y = X\beta + \varepsilon$$

- In observational data, X may be correlated with unobserved factors inside ε

Suppose we have a simple linear regression model

$$y = X\beta + \varepsilon$$

- In observational data, X may be correlated with unobserved factors inside ε
- Then OLS is biased because $\mathrm{Cov}(X, \varepsilon) \neq 0$ (omitted variable bias)

Suppose we have a simple linear regression model

$$y = X\beta + \varepsilon$$

- In observational data, X may be correlated with unobserved factors inside ε
- Then OLS is biased because $\mathrm{Cov}(X, \varepsilon) \neq 0$ (omitted variable bias)
- One way to reduce this bias is to include a proxy variable

Suppose we have a simple linear regression model

$$y = X\beta + \varepsilon$$

- In observational data, X may be correlated with unobserved factors inside ε
- Then OLS is biased because $\mathrm{Cov}(X, \varepsilon) \neq 0$ (omitted variable bias)
- One way to reduce this bias is to include a proxy variable
- A proxy is an observed variable correlated with an unobserved confounder

Suppose we have a simple linear regression model

$$y = X\beta + \varepsilon$$

- In observational data, X may be correlated with unobserved factors inside ε
- Then OLS is biased because $\mathrm{Cov}(X, \varepsilon) \neq 0$ (omitted variable bias)
- One way to reduce this bias is to include a proxy variable
- A proxy is an observed variable correlated with an unobserved confounder
- Under certain conditions, a proxy can reduce or even eliminate the bias

Suppose we have a simple linear regression model

$$y = X\beta + \varepsilon$$

- In observational data, X may be correlated with unobserved factors inside ε
- Then OLS is biased because $\text{Cov}(X, \varepsilon) \neq 0$ (omitted variable bias)
- One way to reduce this bias is to include a proxy variable
- A proxy is an observed variable correlated with an unobserved confounder
- Under certain conditions, a proxy can reduce or even eliminate the bias
- e.g. IQ a proxy for unobserved ability that biases returns to schooling

- Proxies are often imperfect: we rarely measure the confounder exactly

- Proxies are often imperfect: we rarely measure the confounder exactly

- Suppose our proxy P measures the true confounder Z with error:

$$P = Z + \nu, \quad \mathrm{Cov}(Z, \nu) = \mathrm{Cov}(X, \nu) = 0$$

- Proxies are often imperfect: we rarely measure the confounder exactly

- Suppose our proxy P measures the true confounder Z with error:

$$P = Z + \nu, \quad \text{Cov}(Z, \nu) = \text{Cov}(X, \nu) = 0$$

- This is the "classical measurement error" (CEV) case

- Proxies are often imperfect: we rarely measure the confounder exactly
- Suppose our proxy P measures the true confounder Z with error:

$$P = Z + \nu, \quad \text{Cov}(Z, \nu) = \text{Cov}(X, \nu) = 0$$

- This is the "classical measurement error" (CEV) case
- OLS estimates are attenuated toward zero ("attenuation bias")

- Proxies are often imperfect: we rarely measure the confounder exactly
- Suppose our proxy P measures the true confounder Z with error:

$$P = Z + \nu, \quad \mathrm{Cov}(Z, \nu) = \mathrm{Cov}(X, \nu) = 0$$

- This is the "classical measurement error" (CEV) case
- OLS estimates are attenuated toward zero ("attenuation bias")
- Intuition: noise in the proxy dilutes the true signal

- Proxies are often imperfect: we rarely measure the confounder exactly
- Suppose our proxy P measures the true confounder Z with error:

$$P = Z + \nu, \quad \mathrm{Cov}(Z, \nu) = \mathrm{Cov}(X, \nu) = 0$$

- This is the "classical measurement error" (CEV) case
- OLS estimates are attenuated toward zero ("attenuation bias")
- Intuition: noise in the proxy dilutes the true signal
- In non-linear models or if CEV fails, the bias can go in any direction

- Proxies are often imperfect: we rarely measure the confounder exactly
- Suppose our proxy P measures the true confounder Z with error:

$$P = Z + \nu, \quad \text{Cov}(Z, \nu) = \text{Cov}(X, \nu) = 0$$

- This is the "classical measurement error" (CEV) case
- OLS estimates are attenuated toward zero ("attenuation bias")
- Intuition: noise in the proxy dilutes the true signal
- In non-linear models or if CEV fails, the bias can go in any direction
    - "non-classical measurement error"

- Unless they happen to resolve the endogeneity problem, proxy variables won't work

- Unless they happen to resolve the endogeneity problem, proxy variables won't work
- (We'd need $\mathbb{E}(\varepsilon \mid X, P) = \mathbb{E}(\varepsilon \mid P)$)

- Unless they happen to resolve the endogeneity problem, proxy variables won't work

- (We'd need $\mathbb{E}(\varepsilon \mid X, P) = \mathbb{E}(\varepsilon \mid P)$)

- And usually proxies don't satisfy this requirement

- Unless they happen to resolve the endogeneity problem, proxy variables won't work
- (We'd need $\mathbb{E}(\varepsilon \mid X, P) = \mathbb{E}(\varepsilon \mid P)$)
- And usually proxies don't satisfy this requirement
- You can use instrumental variables to solve the ME problem

- Unless they happen to resolve the endogeneity problem, proxy variables won't work
- (We'd need $\mathbb{E}(\varepsilon \mid X, P) = \mathbb{E}(\varepsilon \mid P)$)
- And usually proxies don't satisfy this requirement
- You can use instrumental variables to solve the ME problem
- But mainly in linear models

- Unless they happen to resolve the endogeneity problem, proxy variables won't work
- (We'd need $\mathbb{E}(\varepsilon \mid X, P) = \mathbb{E}(\varepsilon \mid P)$)
- And usually proxies don't satisfy this requirement
- You can use instrumental variables to solve the ME problem
- But mainly in linear models
- And, of course, instrument validity is almost always in question

- Unless they happen to resolve the endogeneity problem, proxy variables won't work
- (We'd need $\mathbb{E}(\varepsilon \mid X, P) = \mathbb{E}(\varepsilon \mid P)$)
- And usually proxies don't satisfy this requirement
- You can use instrumental variables to solve the ME problem
- But mainly in linear models
- And, of course, instrument validity is almost always in question
- So it seems we face a trade-off between OVB and attenuation bias

- What if we have many correlated proxies?

- What if we have many correlated proxies?

- For the unobserved ability question, we might have many different proxies

- What if we have many correlated proxies?
- For the unobserved ability question, we might have many different proxies
- e.g. individuals might take multiple standardized tests

- What if we have many correlated proxies?
- For the unobserved ability question, we might have many different proxies
- e.g. individuals might take multiple standardized tests
- How do we know which test scores to attempt to use as proxies?

- What if we have many correlated proxies?
- For the unobserved ability question, we might have many different proxies
- e.g. individuals might take multiple standardized tests
- How do we know which test scores to attempt to use as proxies?
- What if each test itself suffers from measurement error?

- What if we have many correlated proxies?
- For the unobserved ability question, we might have many different proxies
- e.g. individuals might take multiple standardized tests
- How do we know which test scores to attempt to use as proxies?
- What if each test itself suffers from measurement error?
- What if the test scores are highly correlated with each other?