One proxy variable for multiple unobserved confounders, what kind of bias can arise?

Yizhou FAN　Hiroshima University
Ran NAKAO　Hiroshima City University

Constructs and Proxies in Social science studies

Construct

An abstract idea or concept that is not directly observable.

Theoretical Variable

GAP

Operational definition

Error Term

Proxy

An indicator as a practical substitute which stands in for some aspects of construct that is not easily observed or measured.

Empirical Variable

☒Validity of measurement
☑Validity of inference

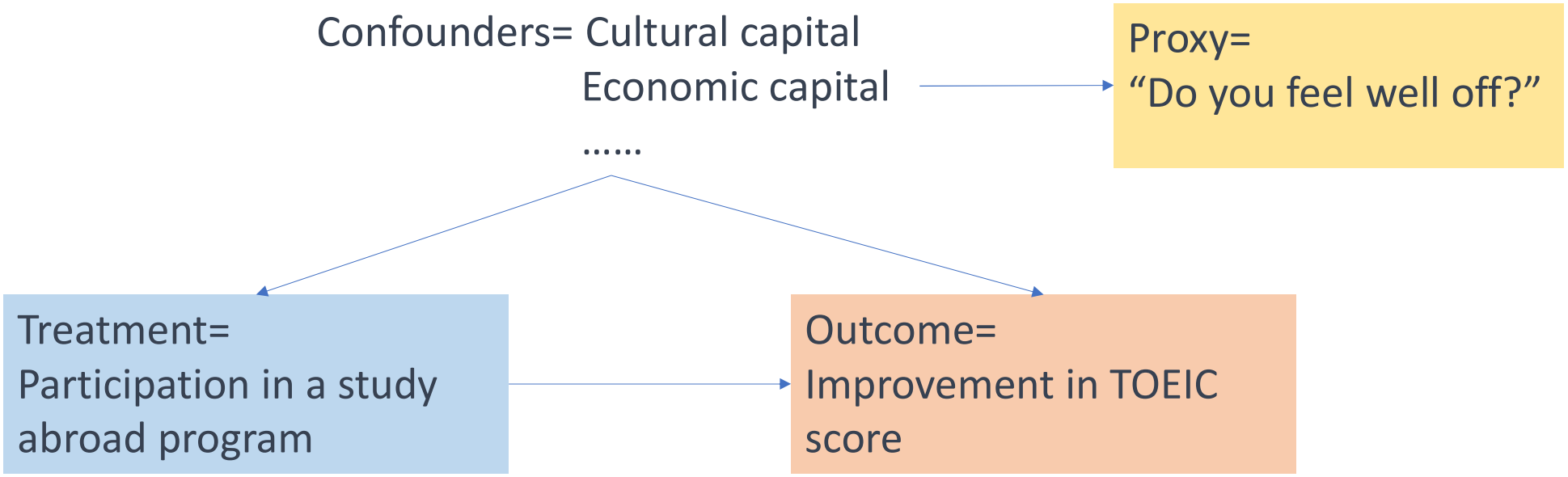
When focusing on the validity of the measurement, **an accurate proxy is one with fewer error terms.**
=evaluate the proxy by its correlation with the construct

When focusing on the validity of the inference, **a preferable proxy is one that results in a more unbiased causal inference.**
=evaluate the proxy by a framework for sensitivity analysis.

A **pragmatism** approach~
A proxy, although not the most accurate, may still produce unbiased results in causal inference.
On the other hand, in some cases, suppressing the error terms could actually amplify the bias in causal inference.

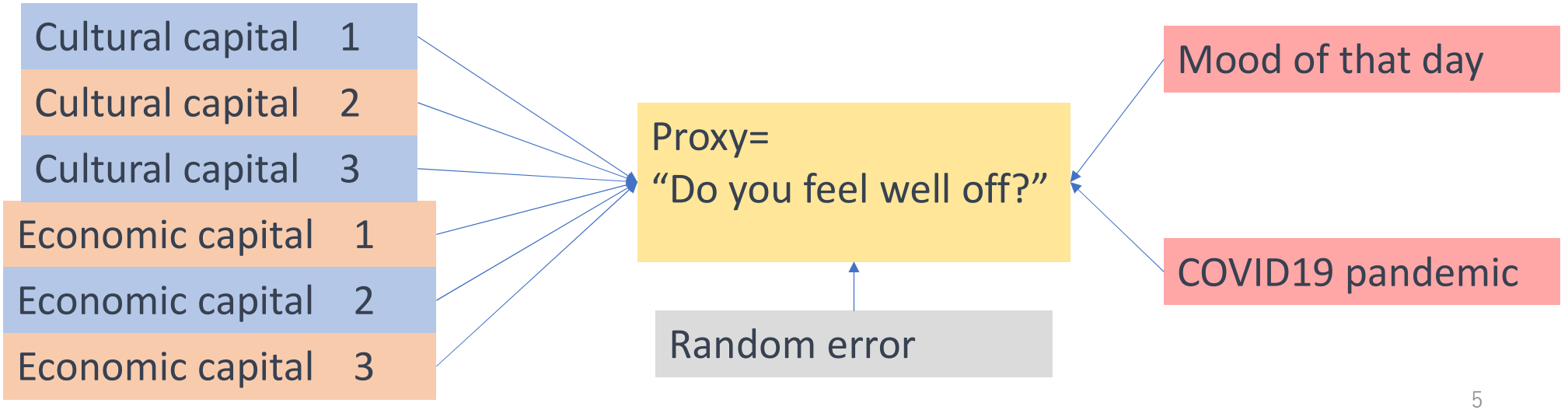
Proxy for Confounder(s)

- Causal inference in social science frequently encounters confounding variables that are abstract constructs, which aren't directly measurable.
- By using operational definitions, it is possible to control confounders partially.

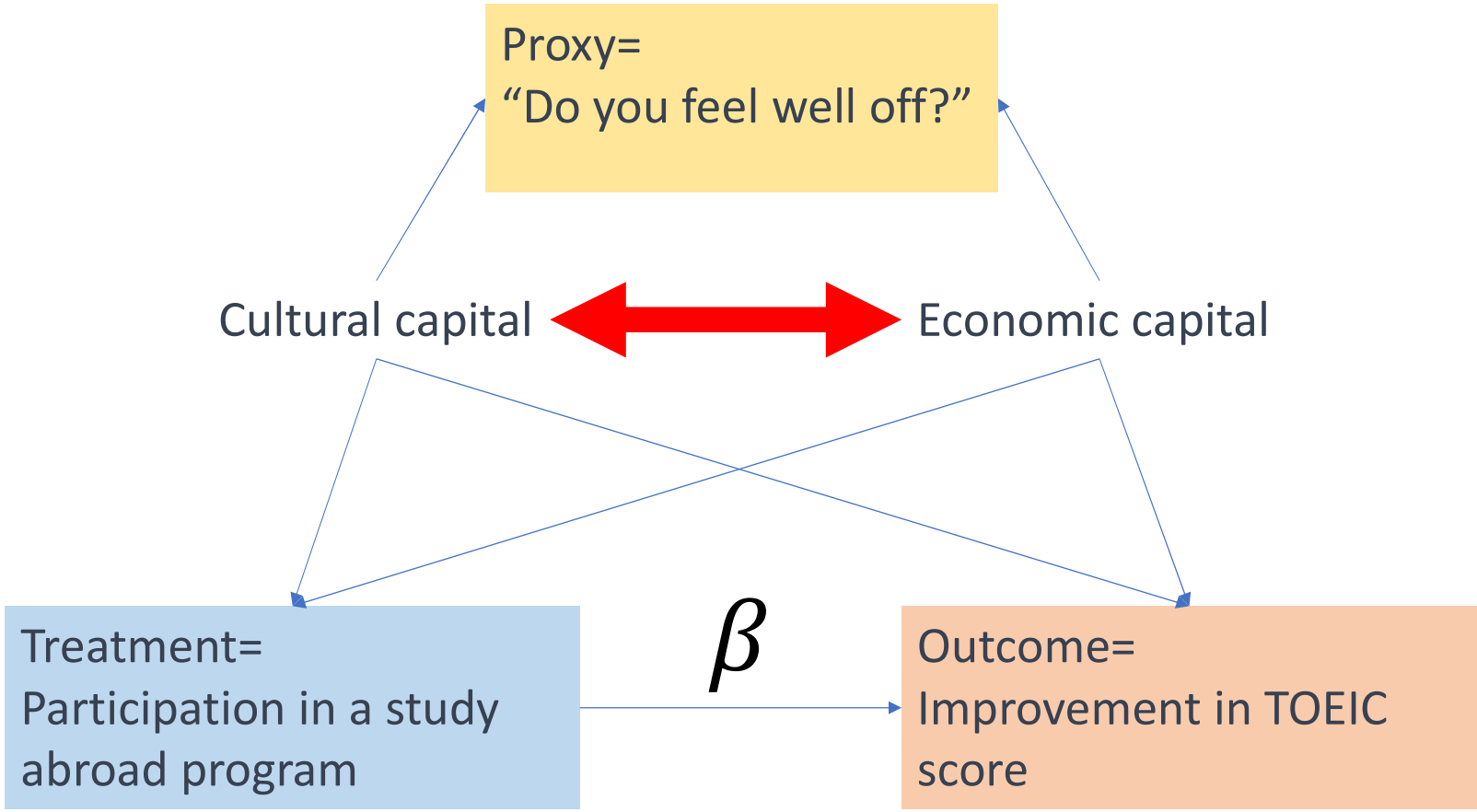


What happens if we control only one proxy for multiple unobserved confounders?

- A construct like “Cultural capital” or “Economic capital” may encompass various meanings.
- Controlling for just one proxy means attempting to encapsulate multiple confounders within a single dimension.

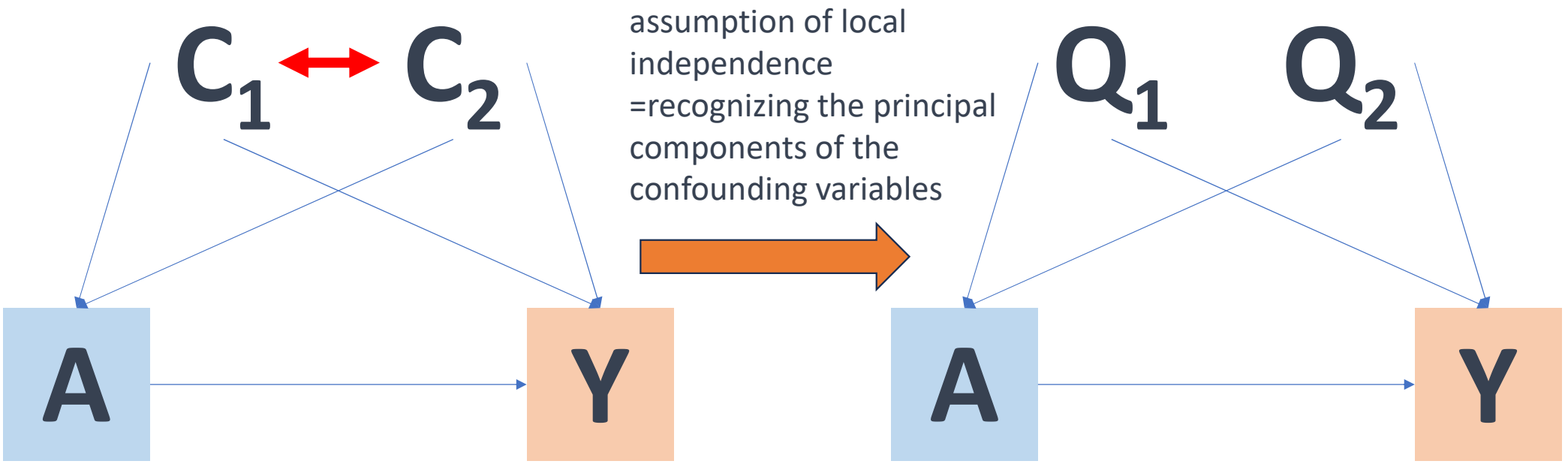


Model


Proxy=
“Do you feel well off?”
Cultural capital
Economic capital
Treatment=
Participation in a study abroad program
β
Outcome=
Improvement in TOEIC score

Assumption of local independence

To calculate the bias, one of the challenges is that the causality among the confounders can be also considered. This makes calculation quite complex.



The generation process of proxy

$$P_i = \sum_j Q_{ij} T_j + U_i$$

$$P = QT + U$$

The proxy P is determined by the sum of the impacts given by each dimension of the confounders Q_j, each with its own weight T_j. U is the term of random errors.

$$\mathrm{Y} = b_0 + b_1 \mathrm{A} + b_2 \mathrm{P}$$

We are going to calculate the deviation of the estimated treatment effect (b_1) from the true value (β) when such proxy is controlled for.

The proxy caused bias = "p-bias"

P-bias written with part correlation coefficients

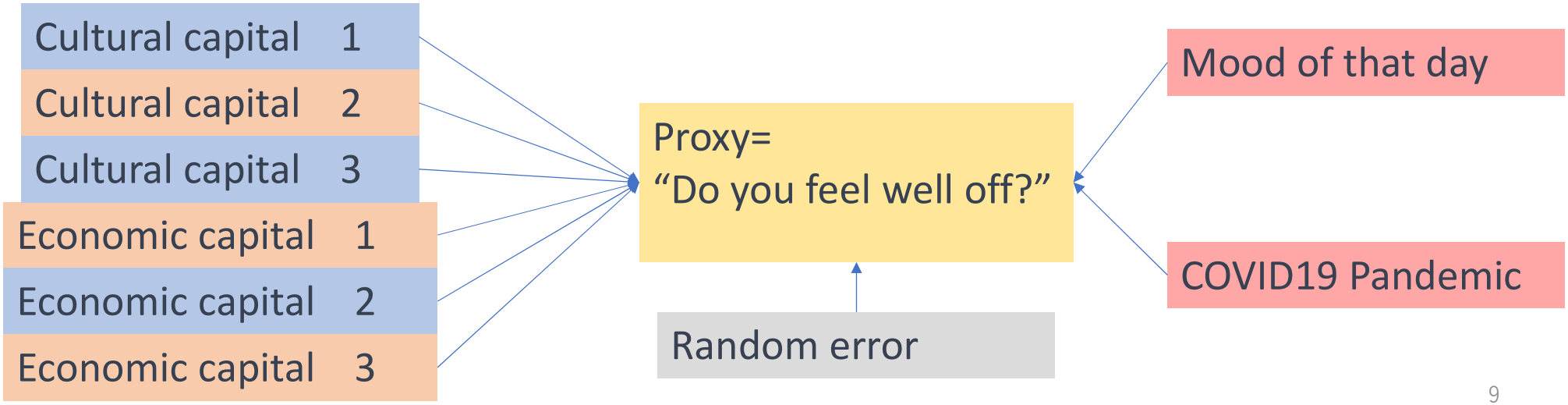
The omitted variable bias due to confounders Q

The accuracy of proxy

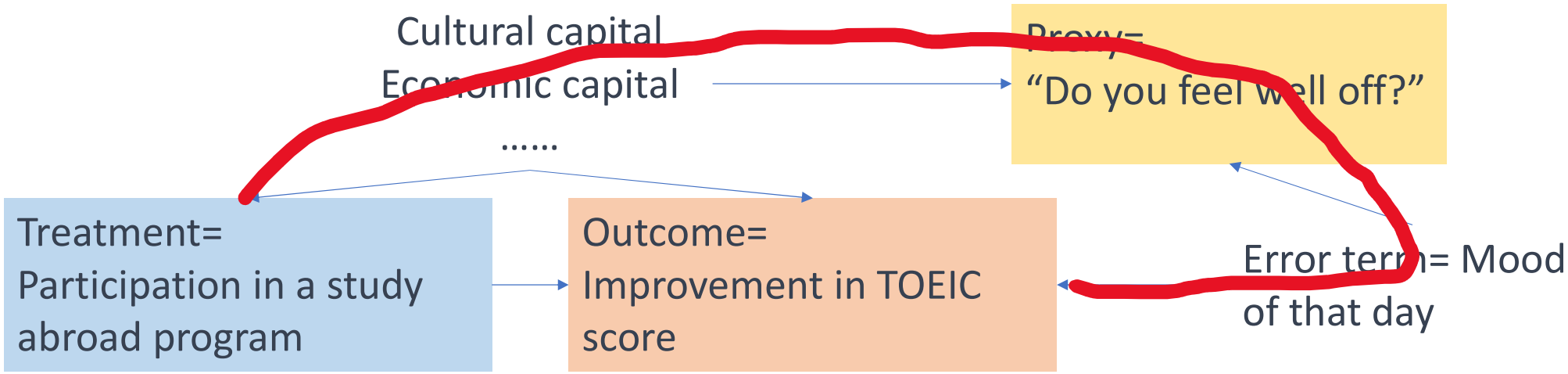
The amount of virtual confounder "A←P→Y"

$$p-bias=\frac{\sum_j\sqrt{\rho_{Aj}\rho_{Yj}}-\rho_{PQ}\sum_j\sqrt{\rho_{Aj}\rho_{Pj}}\cdot\sum_j\sqrt{\rho_{Yj}\rho_{Pj}}}{1-\rho_{PQ}\left(\sum_j\sqrt{\rho_{Aj}\rho_{Pj}}\right)^2}$$

VIF (Variance Inflation Factor) of A when controlling for P



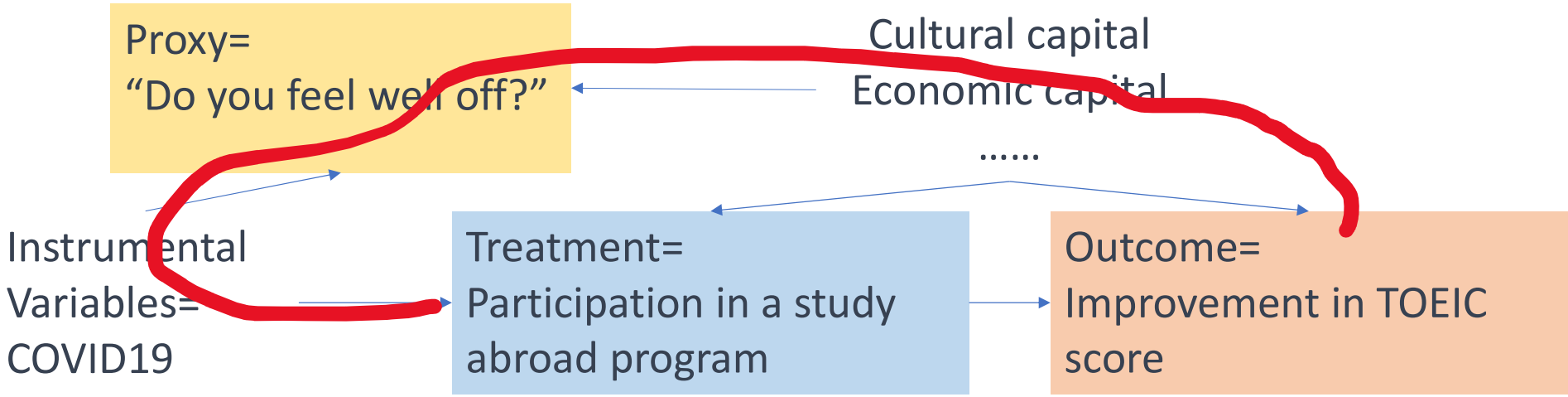
P influenced by the error term of outcomes



$$p-bias=\frac{\sum_j\sqrt{\rho_{Aj}\rho_{Yj}}-\rho_{PQ}\sum_j\sqrt{\rho_{Yj}\rho_{Pj}}\sum_j\sqrt{\rho_{Aj}\rho_{Pj}}-\rho_{PQ}\sum_j\sqrt{\rho_{Aj}\rho_{Pj}}\sum_r\sqrt{\rho_{Yr}\rho_{Pr}}}{1-\rho_{PQ}\left(\sum_j\sqrt{\rho_{Aj}\rho_{Pj}}\right)^2}$$

*** Back-door path "A←Q→P←Y" has been opened due to controlling for P.**

P influenced by instrumental variables



$$p-bias=\frac{\sum_j\sqrt{\rho_{Aj}\rho_{Yj}}-\rho_{PQ}\sum_j\sqrt{\rho_{Aj}\rho_{Pj}}\sum_j\sqrt{\rho_{Yj}\rho_{Pj}}-\rho_{PQ}\sum_r\sqrt{\rho_{Ar}\rho_{Pr}}\sum_j\sqrt{\rho_{Yj}\rho_{Pj}}}{1-\rho_{PQ}\left(\sum_j\sqrt{\rho_{Aj}\rho_{Pj}}+\sum_r\sqrt{\rho_{Ar}\rho_{Pr}}\right)^2}$$

***Back-door path "A←IV→P←Q←Y" has been opened due to controlling for P.**

*The denominator became smaller. In other words, the VIF (Variance Inflation Factor) increased. This suggests an increase in multicollinearity between A and P, leading to an expansion of estimation bias.

Let's ignore the random error term in the proxy. Is it always safe to control for such proxy?

- No random error = an extremely accurate proxy

$$P = QT$$

- What happens if we control for such proxy?

Intuitively, one might think that while it's aggregating multiple confounders into a single proxy, it might not cover all the information, however, it would likely be better than not controlling for anything.

- **No! In some case, controlling for the proxy might actually increase the estimation bias, even there is no random error term.**
- From the theory of causal inference and the perspective of DAG, this can be explained by the collider bias and the VIF exaggeration due to controlling P.

From a geometrical perspective

Vectors concern to the generating processes of A, Y and P

Dot product of ϱ_A and ϱ_Y

Dot product of ϱ_A and ϱ_P

$$p - bias = \frac{\sum_j \sqrt{\rho_{Aj}\rho_{Yj}} - \sum_j \sqrt{\rho_{Aj}\rho_{Pj}} \cdot \sum_j \sqrt{\rho_{Yj}\rho_{Pj}}}{1 - \left(\sum_j \sqrt{\rho_{Aj}\rho_{Pj}}\right)^2}$$

Dot product of ϱ_Y and ϱ_P

Dot product of ϱ_A and ϱ_Y

- ρ_{Aj} = The portion of variable A determined by Q_j The **part correlation coefficient** of the j-th confounder Q_j towards treatment A.
- ρ_{Yj}= The portion of variable Y determined by Q_j The **part correlation coefficient** of the j-th confounder Q_j towards outcome Y.
- ρ_{Pj}= The portion of variable P determined by Q_j The **part correlation coefficient** of the j-th confounder Q_j towards proxy P.

Let vector ϱ_A be the vector formed by arranging $\sqrt{\rho_{Aj}}$
Let vector ϱ_Y be the vector formed by arranging $\sqrt{\rho_{Yj}}$
Let vector ϱ_P be the vector formed by arranging $\sqrt{\rho_{Pj}}$

Part correlation coefficient

$$\sum_j \rho_{Aj} = \sum_j \rho_{Pj} = 1$$

$$\sum_j \rho_{Yj} = 1 - \beta^2$$

$$\|\varrho_A\| = \|\varrho_P\| = 1$$
$$\|\varrho_Y\| = \sqrt{1 - \beta^2}$$

So that···

The omitted variable bias due to confounders Q

The amount of virtual confounder "A←P→Y"

$$p - bias = \sqrt{1-\beta^2}\,\frac{cosAY - cosAP \cdot cosYP}{1-(cosAP)^2}$$

VIF (Variance Inflation Factor) of A when controlling for P

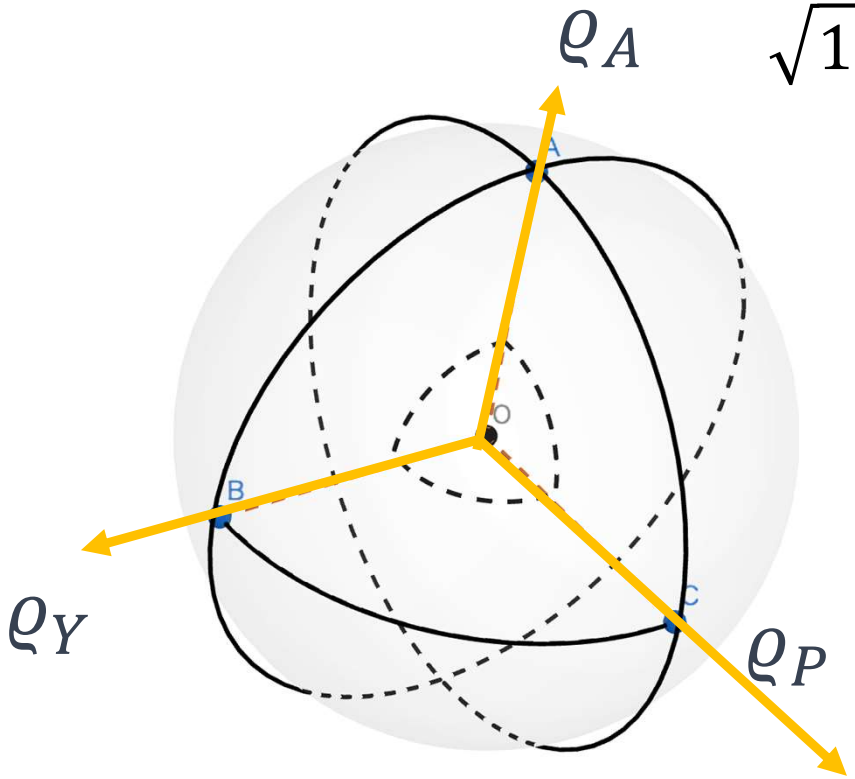
- $\sum_j \sqrt{\rho_{Aj}\rho_{Yj}}$ is the $\sqrt{1-\beta^2}$ times cosine of the angle between ϱ_A and ϱ_Y ~cosAY
- $\sum_j \sqrt{\rho_{Aj}\rho_{Pj}}$ is the cosine of the angle between ϱ_A and ϱ_P ~cosAP
- $\sum_j \sqrt{\rho_{Yj}\rho_{Pj}}$ is the $\sqrt{1-\beta^2}$ times cosine of the angle between ϱ_Y and ϱ_P ~cosYP

The value of p-bias is determined by the relative positional relationship among the three vectors ---- $\varrho_A, \varrho_Y, \varrho_P$

=How do the generating process of P , A and Y, from Q, are alike to each other.

→ With what positional relationship, p-bias will be exactly zero?

Hint from Spherical Trigonometry



The omitted variable bias due to confounders Q

The amount of virtual confounder "A←P→Y"

$$\sqrt{1-\beta^2}\frac{cosAY - cosAP \cdot cosYP}{1-(cosAP)^2} = 0$$

VIF (Variance Inflation Factor) of A when controlling for P

$$cosAY - cosAP \cdot cosYP = 0$$

When and only when **the plane formed by ϱ_A and ϱ_P, and the plane formed by ϱ_Y and ϱ_P are just vertical to each other perfectly.**

An example of a bad proxy

$$p - bias = \frac{cosAY - cosAP \cdot cosYP}{1 - (cosAP)^2}$$

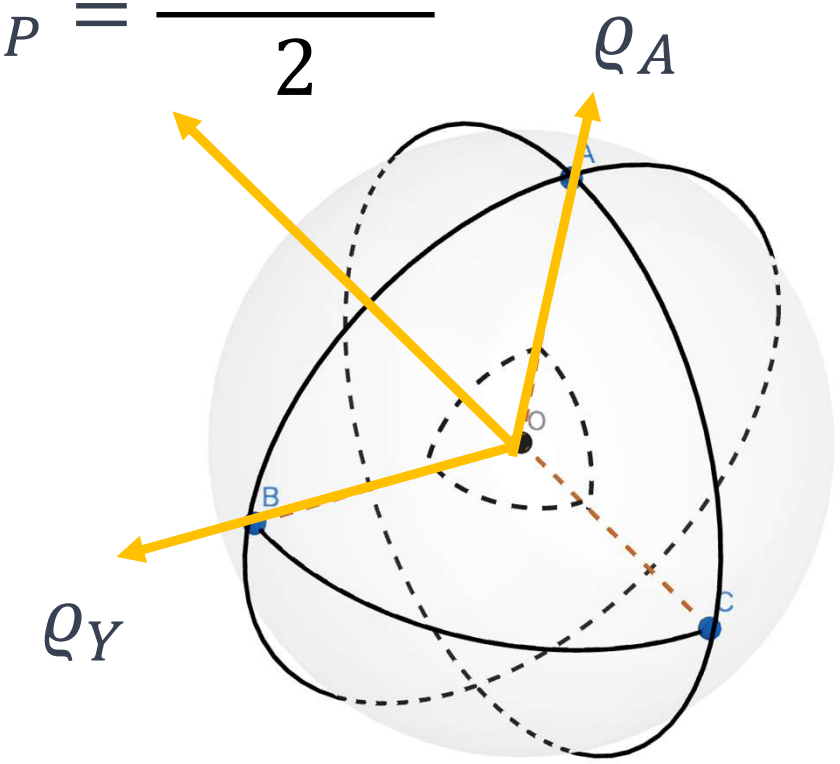
$$= \frac{cos2AP - cosAP \cdot cosAP}{1 - (cosAP)^2} = \frac{-(sinAP)^2}{1 - (cosAP)^2}$$

$$= -1$$

$$< \boldsymbol{-cosAY}$$

In this case, p-bias will be bigger than the omitted variable bias due to confounders Q.

$$\varrho_P = \frac{\varrho_A + \varrho_Y}{2}$$



Conclusions

- We explored the formula for estimation bias, termed “p-bias,” when incorporating a single proxy in a model with multiple unobservable confounders.
- Introducing a proxy influenced by error term of outcomes or instrumental variables will cause a dilemma between partial controlment of unobserved confounders and collider bias. It is necessary to consider case by case whether a proxy is worthy to be controlled for.
- In certain situations, adjusting for such proxy can exaggerate the estimation bias, even without any error term present.
- The value of p-bias is determined by the relative positional relationship among the three vectors ϱ_A, ϱ_Y, ϱ_P. These vectors illustrate the similarities in how P, A, and Y are generated from Q.
- From a geometrical perspective, p-bias will be 0 only if the plane formed by $\boldsymbol{\varrho_A}$ and $\boldsymbol{\varrho_P}$, and the plane formed by $\boldsymbol{\varrho_Y}$ and $\boldsymbol{\varrho_P}$ are perfectly perpendicular.