Dati multi-fonte e analisi territoriali

Marco Tosi
a.a. 2025/2026

1- Disegno della ricerca con dati sociali

Finalità

1- Scopo Descrittivo

- Come i tassi di fecondità evolvono nel tempo; Quanti persone sono senza parenti prossimi (figli, fratelli) in vari paesi Europei.

2- Scopo Esplicativo

- L'espansione del sistema scolastico influisce sui bassi tassi di fecondità? Non avere parenti prossimi ha conseguenze negative sulla salute mentale?
- Causalità è il fine desiderabile (non sempre realizzabile).

1- Scopo descrittivo

Descrizione di un fenomeno

- Validità esterna: rappresentatività del campione rispetto alla popolazione. I risultati dipendono fortemente dal tipo di campionamento (probabilistico o non-probabilistico) e valori mancanti.
- Misurazione del fenomeno: con quale precisione le variabili catturano il concetto.

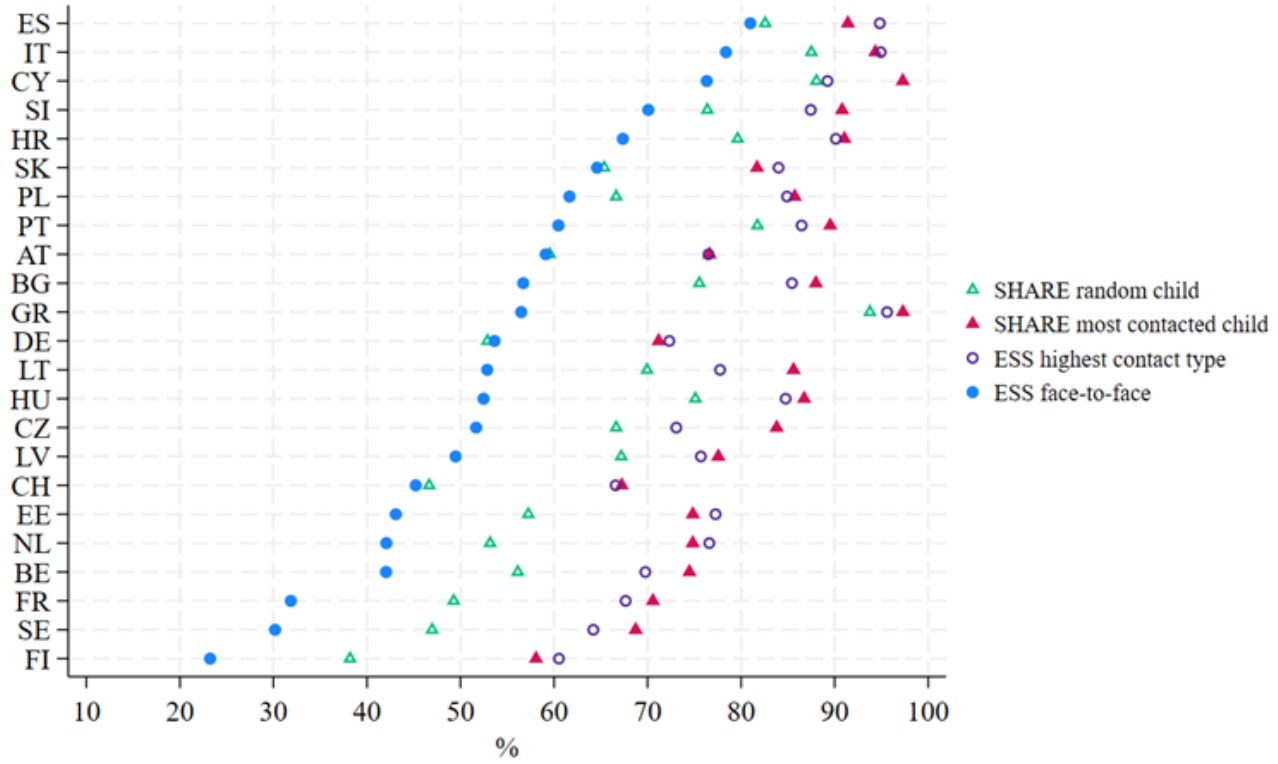
1- Un esempio di analisi descrittiva

Percentuale di genitori che contattano i figli più di una volta a settimana a seconda del tipo di misurazione e fonte dati.

ESS (European Social Survey): dati su un figlio estratto casualmente e diversi tipi di contatto (di persone, telefonico, digitale). Campione di 16+

SHARE (Survey of Health Ageing and Retirement in Europe): dati su 7 figli e contatti più frequenti. Campione di 50+

Armonizzare i dati…



1- Distorsione nell'analisi descrittiva

- Una possibile distorsione è introdotta dai **dati mancanti:**
  - Per il meccanismo generativo dei missing: Ad esempio avere opinioni individualistiche che privilegiano la realizzazione personale rispetto alla formazione di una famiglia può essere associato ad una bassa volontà a rispondere a quesiti sui figli.
    - Stima osservata è **sovrastimata** perché tendiamo ad osservare quelli che hanno opinioni e valori pro-famiglia.
  - Per la struttura del campionamento probabilistico: Chi non risponde ha caratteristiche non rappresentate nei dati.

2- Scopo esplicativo

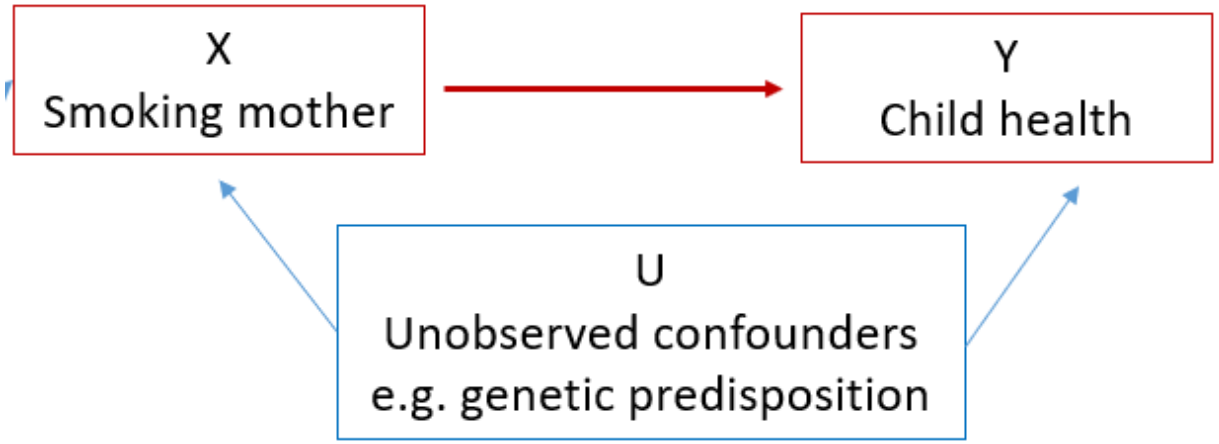
Ipotizziamo teoreticamente che X -> Y

- Validità interna: quanto la relazione stimate di X -> Y rappresenta l'effetto «vero» di X -> Y (valori mancanti possono inficiare la relazione quando M -> X e M -> Y).
- Esperimenti: manipoliamo X (ricevere trattamento) e analizziamo le differenze in Y tra trattati (X=1) e controlli (X=0). Assegnazione casuale di X rende i due gruppi identici per caratteristiche.
- Studi Osservazionali: Non possiamo manipolare X (es. numero di figli non è distribuito casualmente nella popolazione ma dipende da una serie di fattori confondenti che possono inficiare X -> Y).

2a- Studi osservazionali

Il problema:

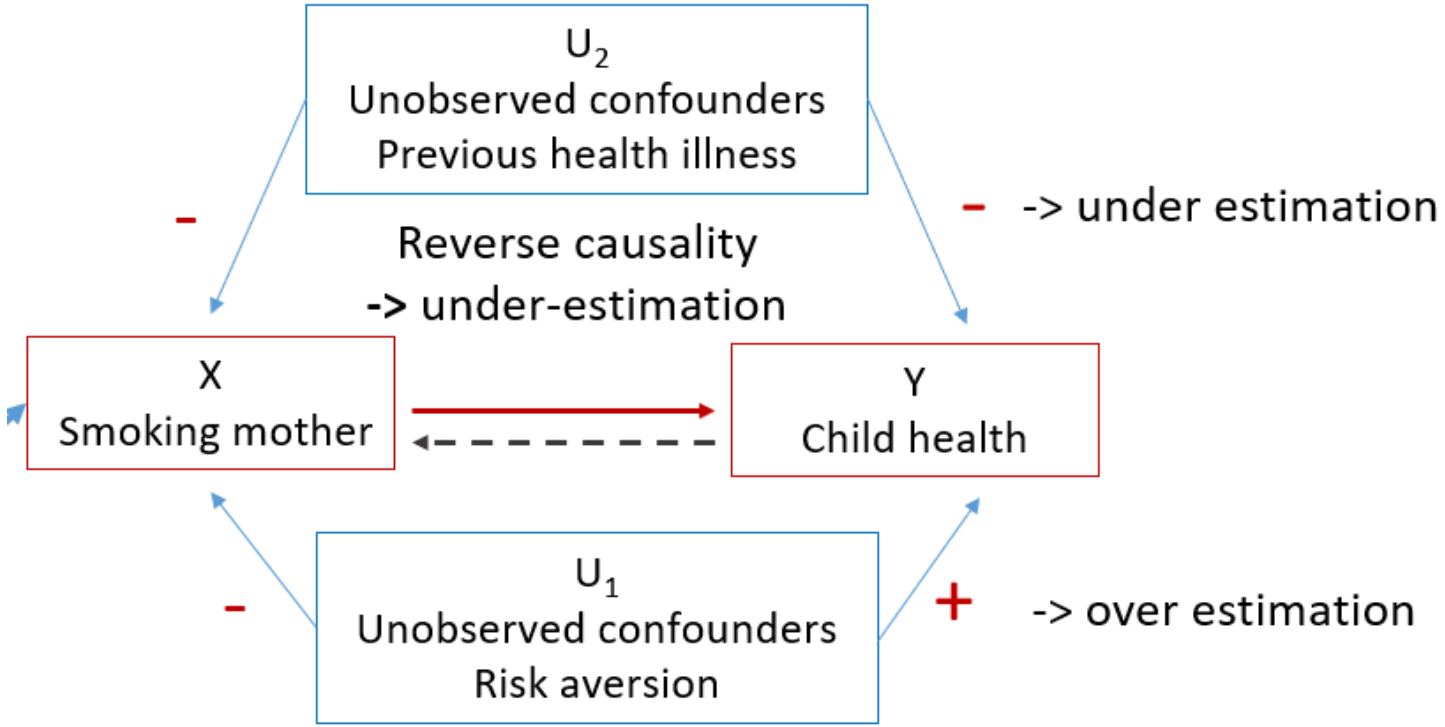
- La probabilità di osservare X non è omogeneamente distribuita nella popolazione: avere una madre fumatrice dipende da fattori osservati (es., classe sociale) e non-osservati (es., predisposizioni genetiche). I due gruppi (X=1 e X=0 avere o meno una madre fumatrice) non sono comparabili.



2a- Studi osservazionali

Il problema:

- Come i confondenti possono distorcere le stime? Teoricamente, la stima «vera» può essere sia > che < di quella osservata).



2a- (Quasi-)Esperimenti con dati sociali: un esempio

Applicazioni sperimentali: es. studi con vignette

- Quesito: le obbligazioni familiari (*sentirsi in dovere in aiutare un parente in difficoltà*) variano a seconda del tipo di parente?
- Campione per quote (non-probabilistico) per genere, età, livello di istruzione, regione (*validità interna vs. esterna*) raccolti tramite un'indagine online. ≈ 5000 di età 45-65.
- Vignette: (1) ridurre desiderabilità sociale; (2) ridurre fattori confondenti (X =0 ≈ X=1); (3) controfattuale è dato dall'assegnazione casuale delle vignette (non da situazione reale del rispondente).
- X = genitore, figlio, fratello, nipote, zii, cugino.

2a- Vignetta

Quasi-esperimenti in campioni non-probabilistici

Maria/Marco ha 50 anni, è sposato/a e lavora a tempo pieno come impiegato/a in un ufficio postale della sua città. *Maria/Marco* ha *un/a figlia adulta/ figlio adulto/ madre/ padre/ fratello/ sorella/ zio/ zia/ cugino/ cugina/ nipote (figlio del fratello) / nipote (figlia del fratello), Anna/Fabio*, che abita nella stessa città e che recentemente ha avuto un incidente in macchina, subendo la frattura del ginocchio. La frattura si è rivelata più grave del previsto e i medici dicono che non potrà tornare a camminare per i prossimi sei mesi.

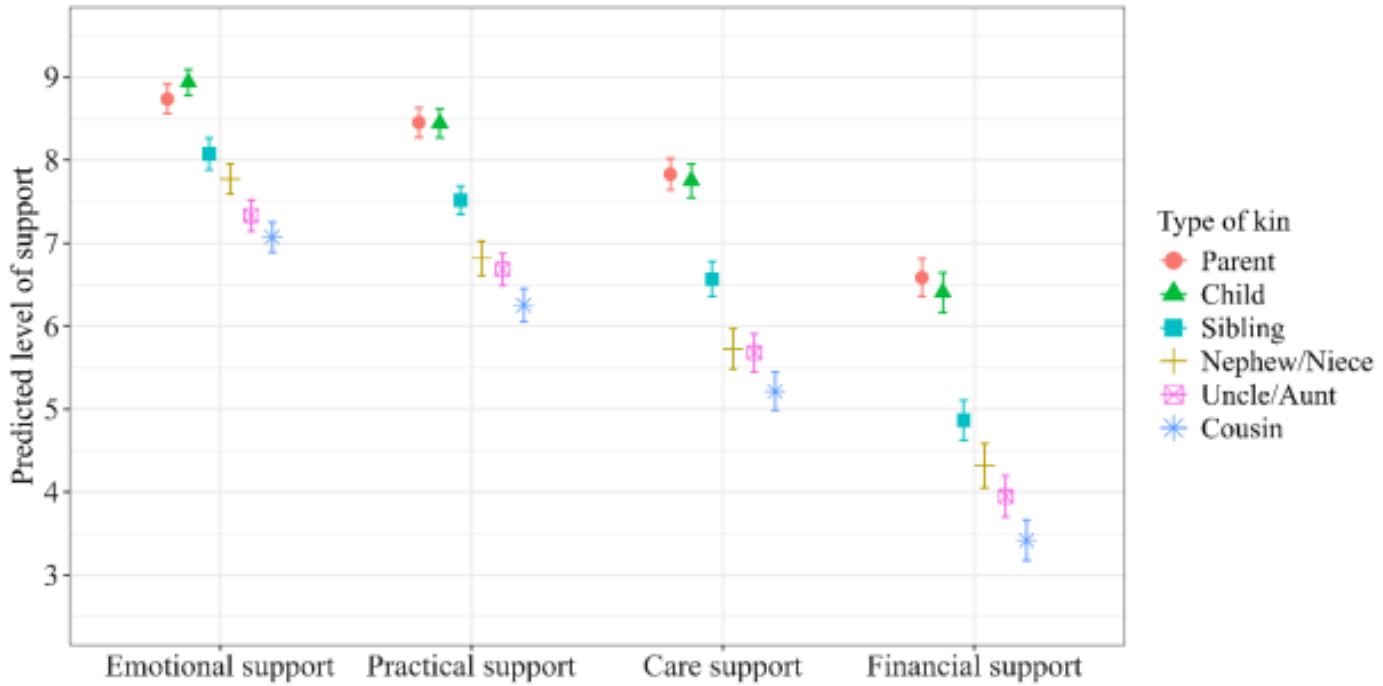
- 2^2 x6^1=24 combinazioni -> 208 casi in ogni cella (208.3 = 5,000/24)
- Una vignetta per ogni rispondente assegnata di modo casuale.
- Un quesito per controllare l'attenzione dei rispondenti (dati online).
- Outcome: divesi tipi di supporto

1.1b In che misura *Maria/Marco* dovrebbe aiutare *Anna/Fabio* nelle attività domestiche (fare la spesa e pulire la casa) o accompagnarlo/la dal medico?

0	1	2	3	4	5	6	7	8	9	10
Marco/Maria non è tenuto farlo										Marco/Maria è tenuto a farlo

2a- Risultati

L’effetto del tipo di parente sulla forza degli obblighi familiari.



Note: estimates from OLS models are adjusted with post-stratification weights. The models (see Table S2 in Appendix) include giver gender and respondents’ gender, age, area of residence and education.

2b- Spiegazione e domanda di ricerca

Coerenza

- La scelta delle variabili, le scelte metodologiche, le tecniche statistiche e i dati devono essere ***coerenti*** con la domanda di ricerca.
- Selezione di variabili per aumentare il potere predittivo ***VS.*** Coerenza del disegno di ricerca (es., descrizione può includere poche o a volte anche solo 2 variabili, e metodi semplici tipo un istogramma).

Associazione vs. Causazione

- Una relazione o associazione tra X e Y non implica una causazione: https://www.tylervigen.com/spurious-correlations

2b- Associazione *vs.* Causazione

Associazione

- E’ una relazione tra variabili (X <-> Y) che ***non*** implica una direzione o un meccanismo (mediazione), e può essere influenzata da variabili confondenti o di selezione.
- Nella maggior parte dei casi ci interessa capire che una associazione *esista* nella popolazione di riferimento.
- Problema delle relazioni spurie.

Causazione

- Una relazione di causazione implica una direzione, un meccanismo generativo, una «depurazione» dai fattori confondenti o di selezione. X produce un effetto su Y (X -> Y).
- La causa del momento iniziale di cambiamento (causa efficiente)

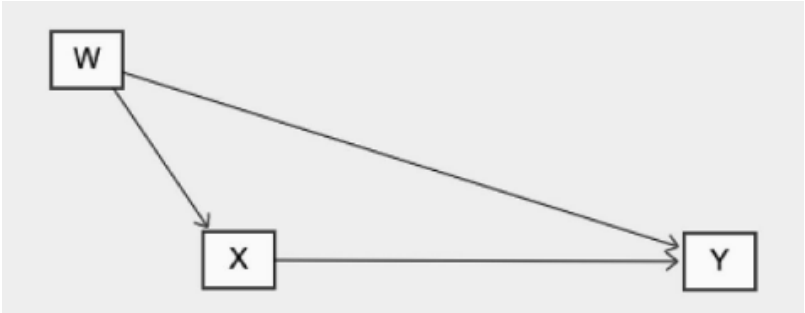
2c- Causazione

- Il link causale rimane un ideale nella ricerca sociale che può trovare evidenza empirica nei dati (in termini probabilistici) e una spiegazione nella teoria. Non è possibile identificare deterministicamente la causa X che produce il cambiamento di Y.
- Tuttavia rimane centrale l'obbiettivo della Scienza di stimare l'*effetto* di una causa. Dobbiamo quindi stimare:
  - L'associazione X -> Y, dove X è antecedente a Y.
  - L'associazione X-> Y, al netto dei fattori confondenti W.
  - Spiegare teoricamente l'effetto di X -> Y.

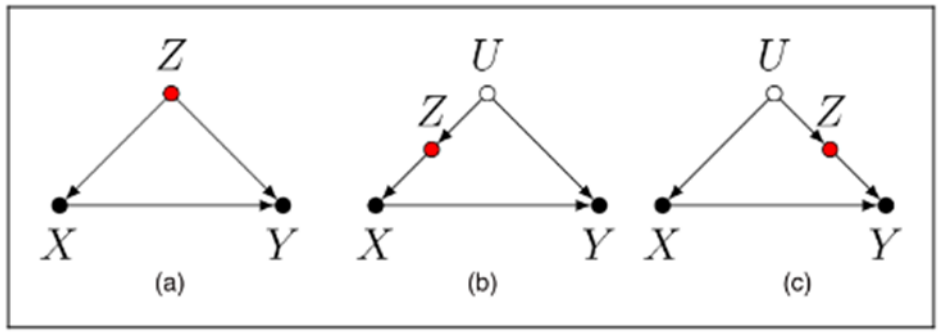
2c- Causalità inversa

- Un problema comune è quello della causalità inversa (Y -> X), ossia l'associazione stimata tra X e Y è prodotta (in parte o del tutto) dall'effetto di Y su X.
  - Ad esempio, il numero di figli (X) può influire sul grado di solitudine (Y), ma anche le persone più sole possono essere meno inclini ad avere figli. Quindi dati in un punto nel tempo:
    - Numero figli -> + Solitudine
    - Solitudine -> - Numero di figli
    - Una possibile soluzione è utilizzare tecniche tipiche dell'inferenza causale (Variabili strumentali, es. il genere dei primi due figli (assegnato casualmente per natura -> propensione ad avere un terzo figlio).

2c- Confondenti

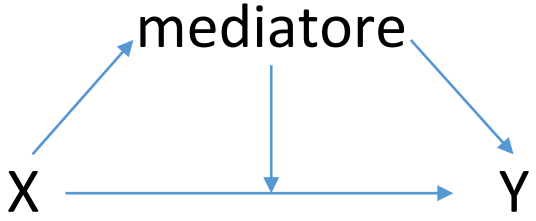


- Un altro problema comune è quello della selezione delle variabili di controllo o confondenti (W) **devono essere antecedenti** a X ed essere associate a Y.
  - W è un vettore di fattori osservati (età) e non-osservati (intelligenza) nei dati. Quando omettiamo uno di questi fattori abbiamo una distorsione («omitted variable bias»).
  - Ad esempio la relazione tra l'altezza e la capacità di linguaggio è distorta dall'età, visto che i bambini tipicamente sono più bassi e devono ancora sviluppare le capacità che ha un adulto.
  - Se W non è antecedente temporalmente a X e quindi X -> W introduciamo una distorsione in X -> Y.



2c- Over-control

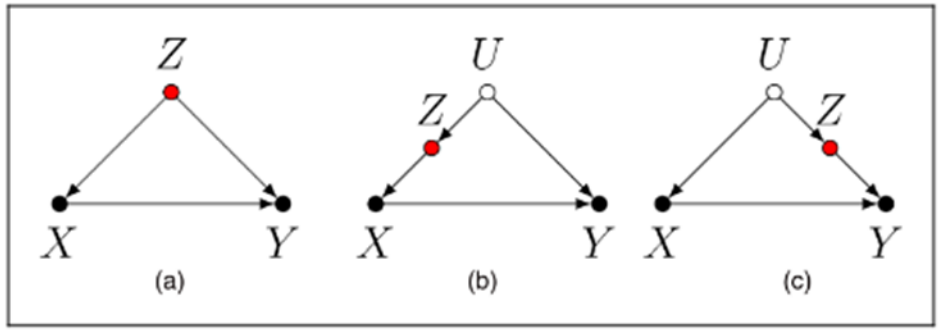
- Il problema della scelta delle variabili di controllo si estende a quelle che vengono definite come mediatori e
  - Se W non è antecedente a X e quindi X -> W introduciamo una distorsione, in quanto W è considerato come un mediatore ossia una variabile che «spezza» il percorso causale tra X e Y.

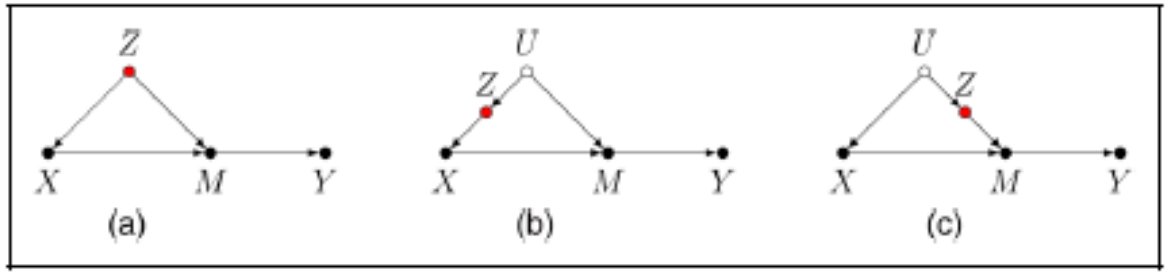


  - Ci sono casi in cui W è associato a X ma non a Y, oppure a Y ma non a X: la distorsione è dovuta alla riduzione della variabilità di una delle due variabili riducendo la precisione della stima X -> Y.

2c- Scelta dei confondenti

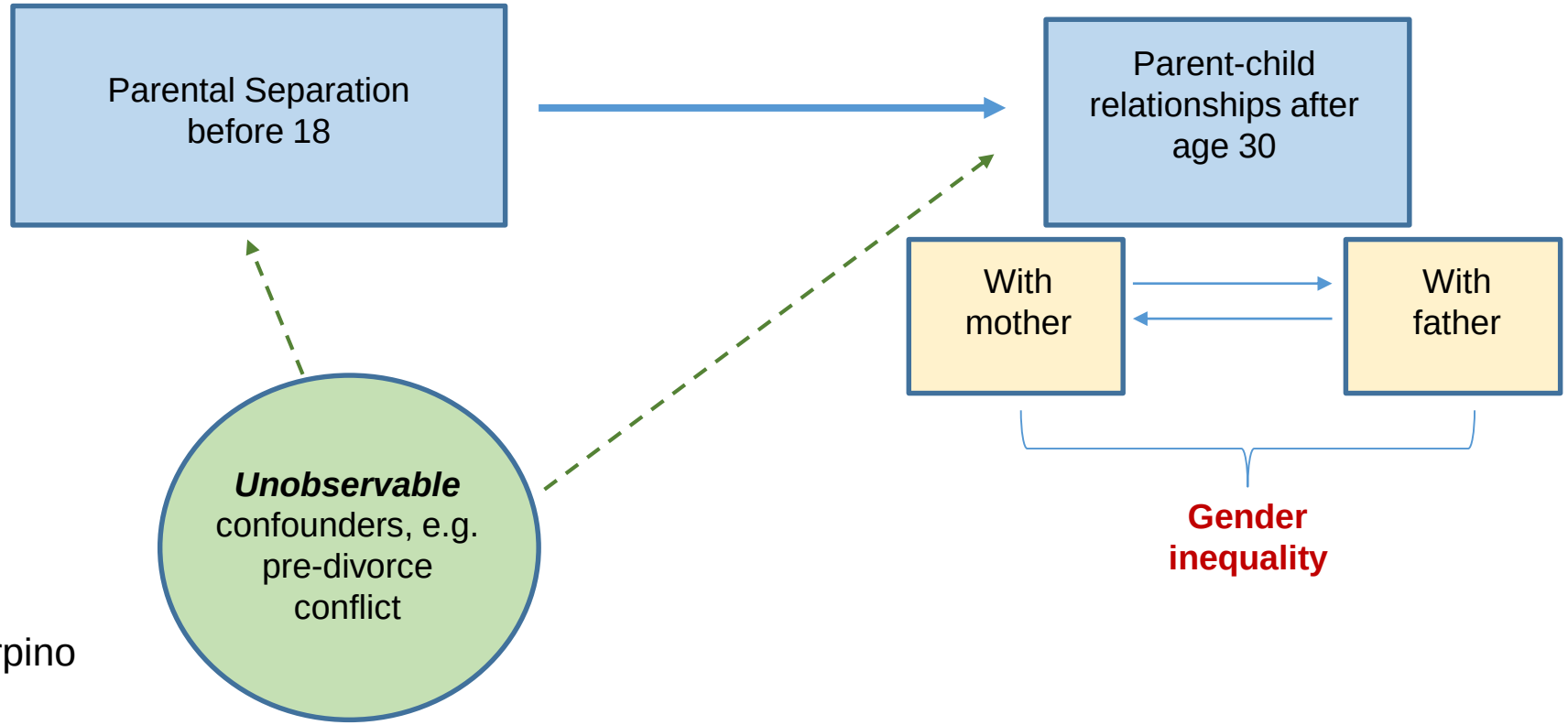
- Alcune regole da seguire nella scelta dei fattori confondenti:
  - Bloccano i percorsi non causali tra X e Y
  - Lasciano spazio ad ogni fattore di mediazione per la stima dell'effetto totale
  - Non aprono lo spazio a nuovi effetti spuri tra X e Y.





2c- Confondenti: Considerare l'inosservabile

- Parental separation *affects* gender (in)equality in later parent-child relationships?



FE models to compare the relation that each respondent has with his/her mother and father

Time = 1 mother
2 father

Multilevel structure.

See Tosi & Arpino 2025

Likely constant across parent-child dyads

2c- Un Esempio

- La separazione dei genitori tende ad indebolire le relazioni genitori-figli adulti nel lungo periodo, in particolare per i padri che solitamente non hanno la custodia dei figli. Quindi le differenze di genere sono più grandi nelle famiglie di separati.
  - Figli devono vedere i genitori in occasioni distinte.
  - Figli tendono a prendere la parte della madre nei conflitti.
  - Genitori risposati possono distaccarsi dai figli nati dalla relazione precedente.
  - Esperienze di convivenza padre- figli è più breve per rinsaldare legami.
- Il problema è che i confondenti inosservabili, es., il clima familiare prima della separazione, può influire sulle associazioni osservate.

Dati: Famiglie e Soggetti sociali 2016

- Informazione su un rispondente per famiglia e i suoi genitori: caratteristiche chieste separatamente per madri e padri (che vivono assieme oppure separati): età, istruzione…
- Outcome: frequenza dei contatti personali, telefonici, e digitali (dicotomica: più di una volta a settimana).
- Modello di regressione *within o modelli ad effetti fissi.*

$$Y_{it} - \overline{Y_i} = \beta_1(X_{it} - \overline{X_i}) + \beta_2(Z_i - \overline{Z_i}) + (e_{it} - \overline{e_i}) + (u_i - \overline{u_i})$$

Data

In this context the outcome is binary, distinguishing between more than weekly contact or less (meet, phone, and digit).

Data are re-shaped in long format:
Observations (dyads) clustered in units (respondents)

	id	pparent	eta2	meet RECODE of ved_	phone RECODE of tel_	pdivorce19	sex	digit
1	43851	2	54	1	0	1	2	0
2	43854	1	46	1	1	2	1	1
3	43854	2	46	1	1	2	1	0
4	43855	1	52	1	1	1	2	0
5	43857	1	38	1	0	1	1	0
6	43858	1	48	1	1	1	1	0
7	43858	2	48	1	1	1	1	0
8	43859	1	35	1	1	1	2	0
9	43859	2	35	1	1	1	2	0
10	43871	1	55	0	0	1	1	0
11	43871	2	55	0	0	1	1	0

Some units have only one observation, given that one of the two parents might be dead -> their contribution to the estimates is limited to β_0

Tosi, M., & Arpino, B. (2025). Gender inequality in intergenerational contact after parental separation in the digital era. Journal of Marriage and Family, 87(2), 824-839.

R- Code: panel linear regression (plm)

```
library(haven)
> JMF <- read_dta("C:/Users/tosi/Desktop/JMFsample.dta")
> library(plm) > # Ensure your variables are factors
> JMF$pparent <- as.factor(JMF$pparent)
> JMF$pdivorce19 <- as.factor(JMF$pdivorce19)
> JMF$sex <- as.factor(JMF$sex)
> # Estimate fixed-effects model
> model <- plm(meet ~ pparent + pdivorce19, data = JMF, + index = c("id", "pparent"), model = "within")
> summary(model)
Oneway (individual) effect Within Model Call:
plm(formula = meet ~ pparent + pdivorce19, data = JMF, model = "within", index = c("id", "pparent"))
Unbalanced Panel: n = 6770, T = 1-2, N = 11041
Residuals:
     Min.    1st Qu. Median   3rd Qu.  Max.
-0.513346 -0.013346 0.000000 0.013346 0.513346

Coefficients: Estimate Std. Error t-value Pr(>|t|)
Pparent2     -0.0266916 0.0030617 -8.7179 < 2.2e-16 ***
--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 87
Residual Sum of Squares: 85.479
R-Squared: 0.017488
Adj. R-Squared: -1.5403
F-statistic: 76.0012 on 1 and 4270 DF, p-value: < 2.22e-16
```

Not estimated because collinear with individual-specific intercepts

Fathers' probability to have frequent meetings with children is 2.67 p.p. lower than mothers' one

Most families in the data are intact -> father and mothers can meet their children on the same occasion

Three-way interactions

```
fe_model <- plm( + meet ~ pparent * pdivorce19 * sex, + data = JMF,
 + index = c("id", "pparent"), + model = "within")
summary(fe_model)
Oneway (individual) effect Within Model Call:
plm(formula = meet ~ pparent * pdivorce19 * sex, data = JMF, model = "within",
index = c("id", "pparent"))

Unbalanced Panel: n = 6770, T = 1-2, N = 11041
Residuals:
     Min.      1st Qu.       Median      3rd Qu.       Max.
-0.6858407 -0.0072993  0.0000000  0.0072993  0.6858407
Coefficients:
                        Estimate Std. Error t-value Pr(>|t|)
             pparent2  -0.0145985  0.0043714 -3.3396 0.0008463 ***
pparent2:pdivorce192  -0.0906646  0.0201223 -4.5057 6.793e-06 ***
pparent2:sex2         -0.0012523  0.0060163 -0.2081 0.8351232
pparent2:pdiv:sex2    -0.2651660  0.0273192 -9.7062 < 2.2e-16 ***
```

Estimates:

Sons:

-Diff in intact fam= -0.01

-Diff in divorced fam= -0.01-0.09 = -10p.p.

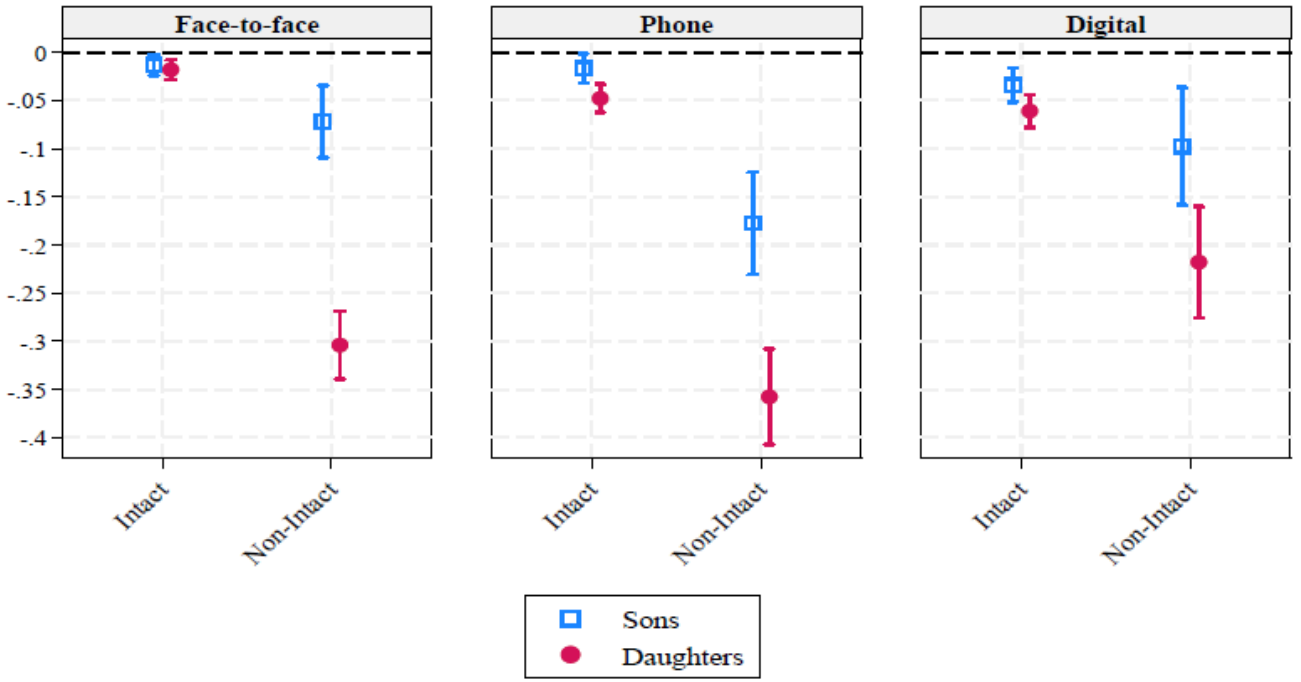
Daughters:

-Diff. in intact fam= -0.01-0.001

-Diff. in divorced fam= -0.01-0.001-0.265 = -27.6 p.p.

Average marginal effects

FIGURE 2. AVERAGE MARGINAL EFFECTS OF BEING FATHER VS. MOTHER ON THE PROBABILITY OF HAVING FREQUENT FACE-TO-FACE, PHONE AND DIGITAL CONTACT WITH PARENTS (FIXED EFFECTS MODELS).



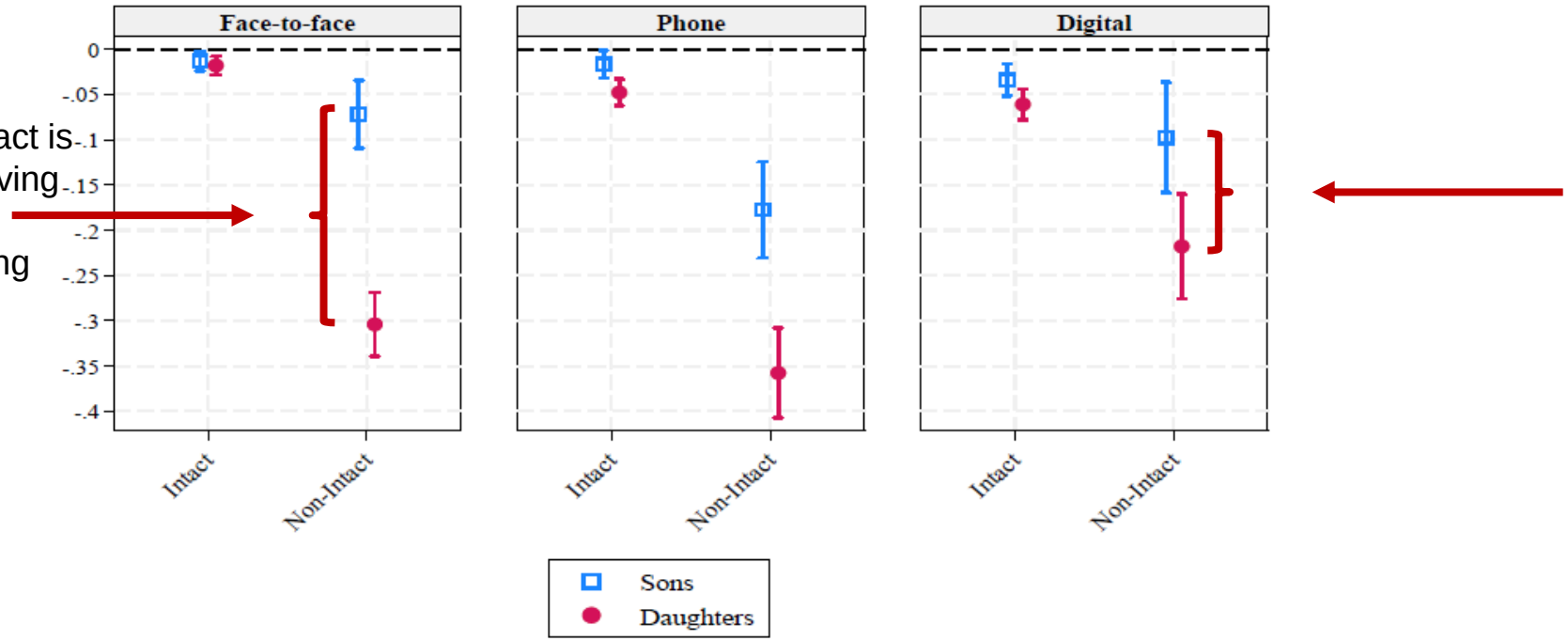
Here, other control variables are included in the model (estimates change slightly)

The effect of gender estimated within different groups: intact and non-intact families, and within sons and daughters.

Average marginal effects

FIGURE 2. AVERAGE MARGINAL EFFECTS OF BEING FATHER VS. MOTHER ON THE PROBABILITY OF HAVING FREQUENT FACE-TO-FACE, PHONE AND DIGITAL CONTACT WITH PARENTS (FIXED EFFECTS MODELS).

In FE models we ***do not estimate average probabilities***, which can be problematic for the interpretation:
Having frequent digital contact is much less common than having face-to-face contact, which can be a reason for observing ***larger effects in terms of percentage points.***



3- Un esercizio senza dati

- *Domanda di ricerca:* vogliamo studiare come avere i figli che vivono vicino possa proteggere dai rischi di solitudine. Nei dati SHARE abbiamo informazione sia su 7 figli sia sugli stati psicologici (scala di solitudine).
  - Come definiamo il campione (età, stato civile)?
  - Come misuriamo X (tra i 7 figli)?
  - Possibili problemi che potremmo incorrere: causalità inversa e fattori confondenti?

Questi aspetti sono fondamentali per capire la **logica e il disegno della ricerca con dati sociali**.