Identifying Peer Effects In Dyadic Longitudinal Data Using Genes as Instrumental Variables

Felix Elwert

University of Wisconsin-Madison

HCEO, Chicago, April 18-19, 2014

O'Malley J, Elwert F, Rosenquist N, Zaslavsky A, Christakis N. (Forthcoming) "Estimating Peer Effects in Longitudinal Dyadic Data Using Instrumental Variables." *Biometrics*.

Genes and Social Processes

1. Investigating the role of genes per se

2. Use genes as tools (IV) to identify social mechanisms in which they're otherwise uninvolved

Mendelian Randomization

- Prior work: Intra-individual
  - My genes, my treatment, my outcome
  - Criticism: exclusion challenges—pleiotropic threat; "identification off ignorance"

- Our take: Inter-individual
  - Instrument interpersonal effects (peer effects): My genes, my treatment, your outcome
  - Advantage: exclusion more defensible since my genes are invisible to social others?
  - Still making strong assumptions, but perhaps more palatable ones

Association or Causation?

The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL ARTICLE

The Spread of Obesity in a Large Social Network Over 32 Years

Nicholas A. Christakis, M.D., Ph.D., M.P.H., and James H. Fowler, Ph.D.

Overview

- Grant two problems
  1. Latent homophily
  2. Latent confounding
- IV strategies in various models for causal peer effects
  1. Alleles per se not promising
  2. Time-varying gene expression more promising
  3. Grant several additional complications
  4. Limitations

Notation & Question

Question: Identify the total causal effect of alter BMI at time (t-1) on ego BMI at time t, $Y_{j(t-1)} \rightarrow Y_{it}$, from observational data?

Dyad: i (ego), j (alter)

Time: q

$Y_{k(q)}$ BMI

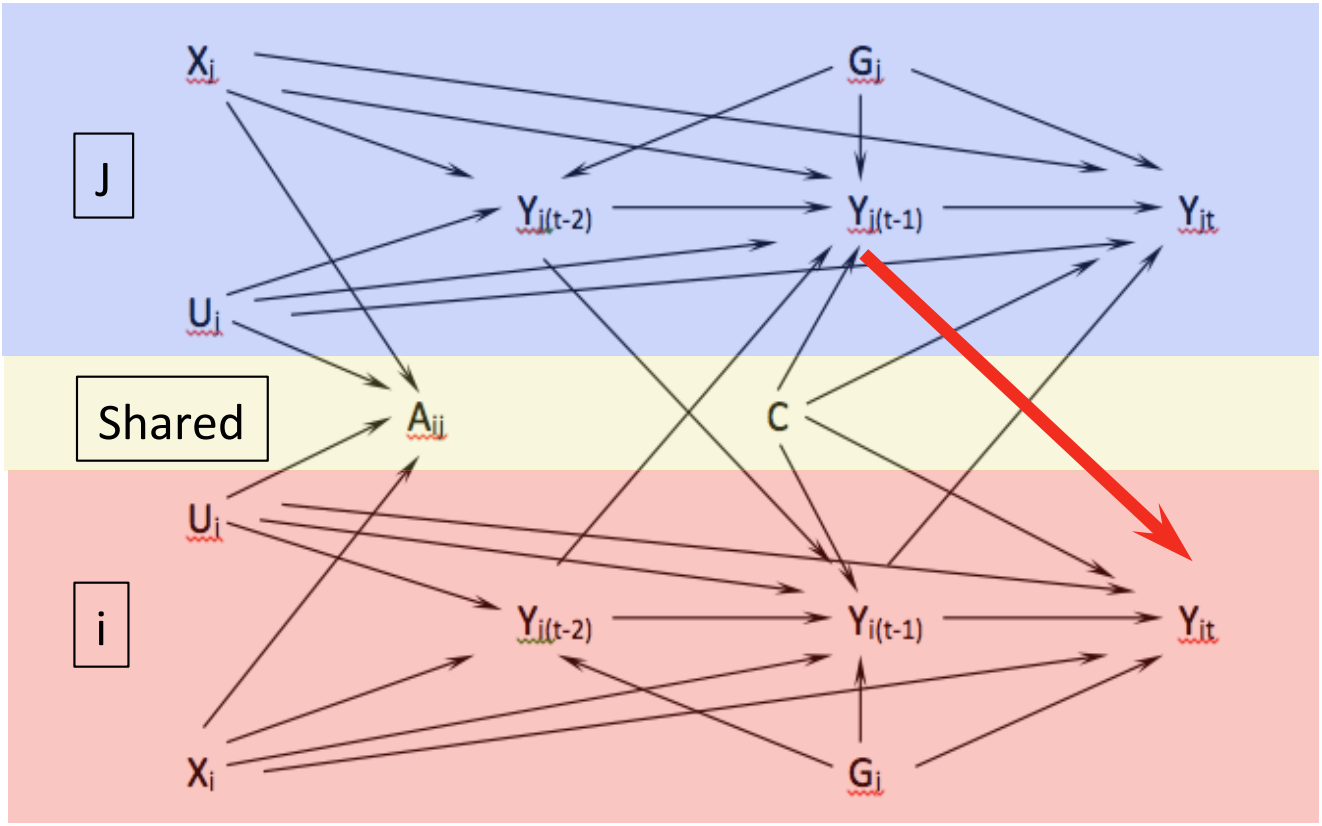
$X_{k(q)}$ Observed predictors

$U_{k(q)}$ Unobserved predictors

A_{ij} Social tie between i,j

C Latent shared context

G_k "Fat genes" alleles (FTO, MC4R, 6 states)

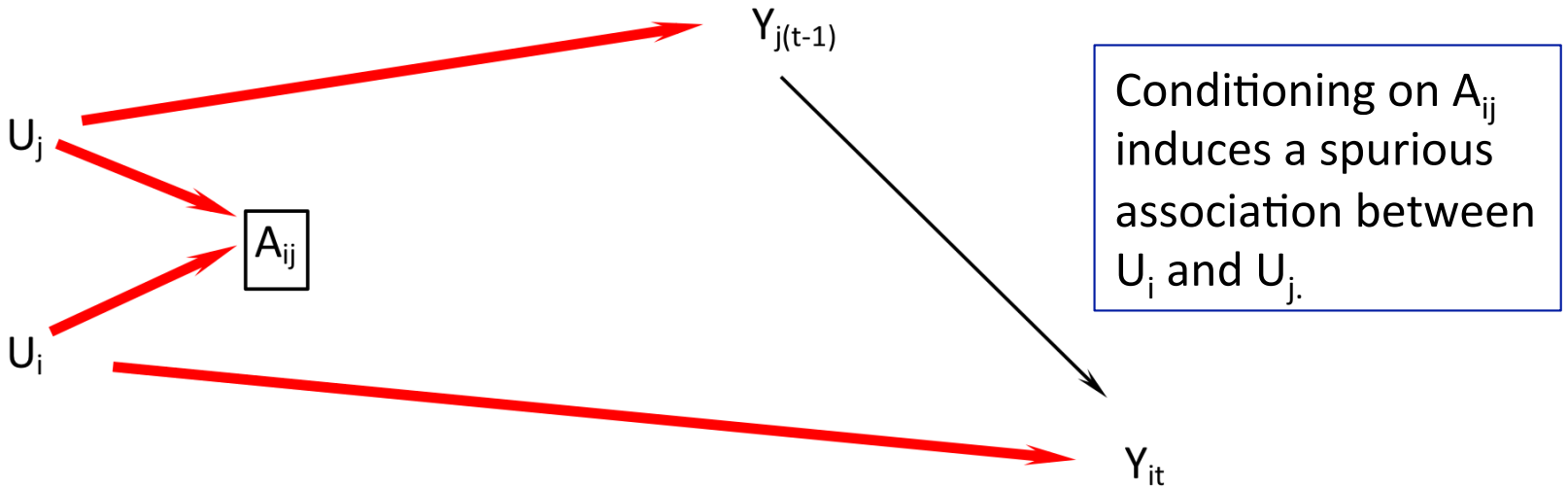


Assume linear & homogenous model for now (no normality assumption).

Read implied associations off causal model: d-separation (~Wright's path rules)

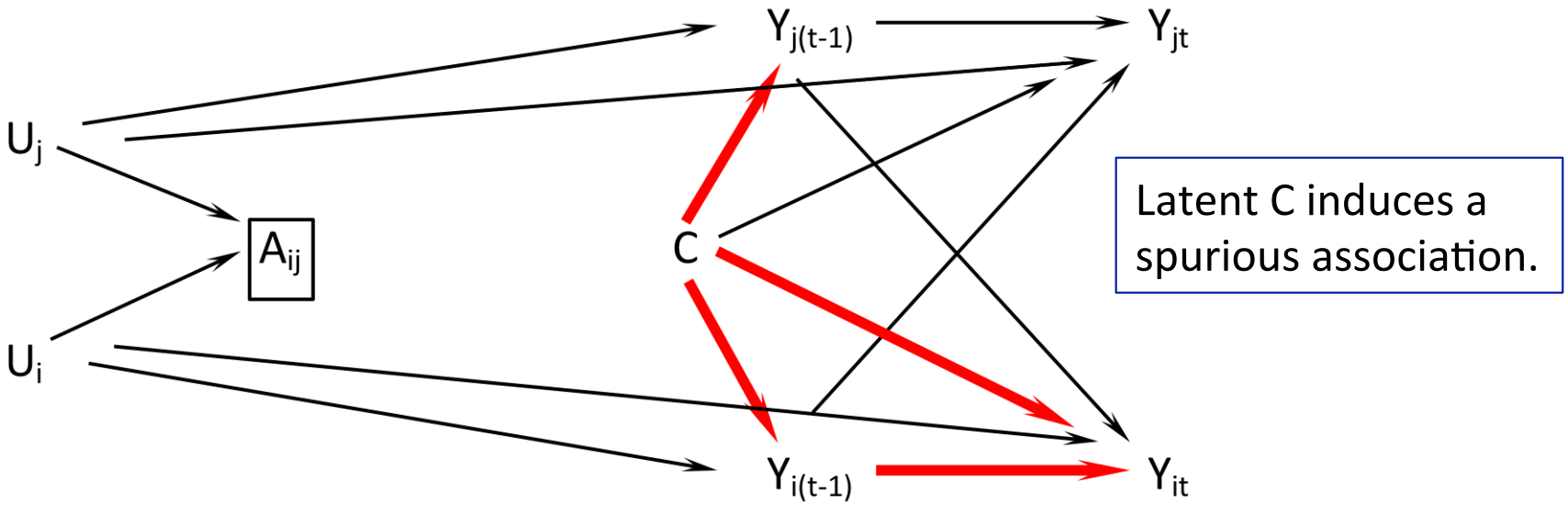
First Problem: Latent Homophily Bias

- Friendship ties A_{ij} are not formed at random (homophily)
- Latent homophily : Unobserved individual-level factors U_k affect tie formation and BMI (e.g., tastes in food, hobbies, exercise habits)
- Computing associations in a sample of dyads amounts to conditioning on the friendship tie, A_{ij} (Shalizi and Thomas 2011)
- Conditioning on the collider A_{ij} induces an association between alter BMI and ego BMI



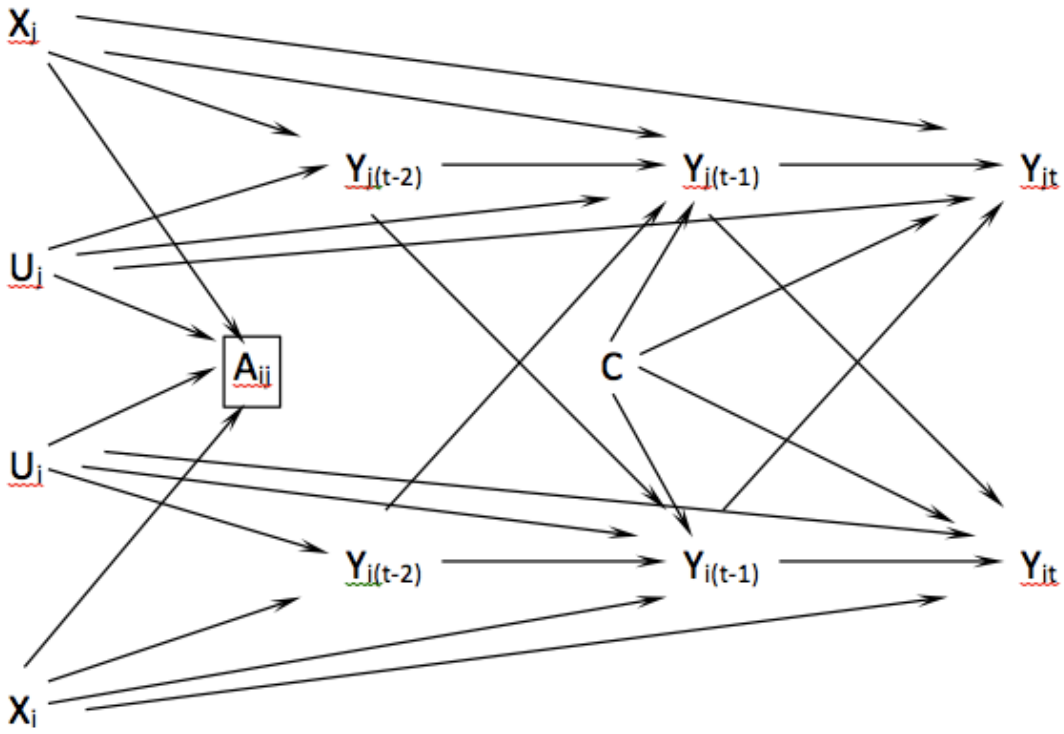
Second Problem: Confounding Bias

- Dyad members share latent contextual exposures, C, which may affect BMI (e.g., food sources, environmental toxins, local norms)
- Non-causal association between alter and ego BMI—confounding bias in addition to homophily.



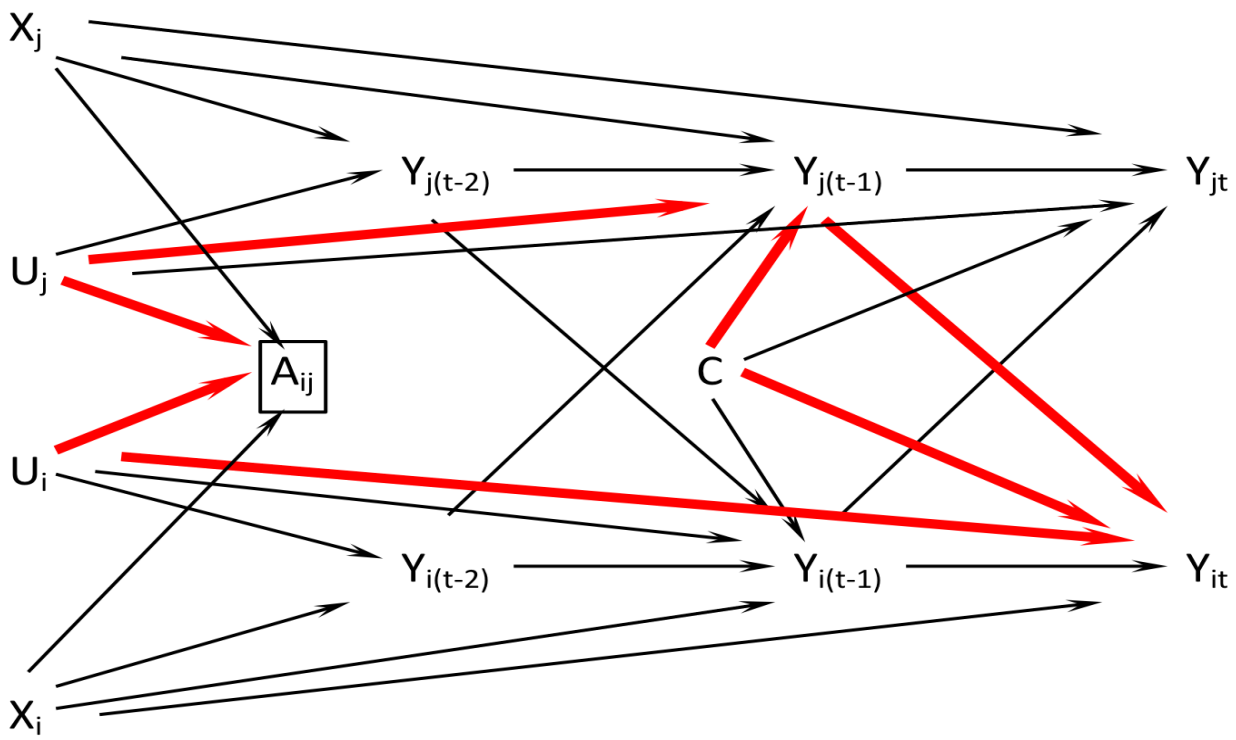
More structure

1. Assume 1-lagged peer effects (= no simultaneity)
2. Multiple-periods (here, three periods)
3. Homophily bias in all periods (U_k affects A and $Y_{k(q)}$)
4. Unobserved confounding in all periods (C affects $Y_{i(q)}$ and $Y_{j(q)}$)
5. Add observed individual level confounders, X_k, e.g., race, age, smoking., which may vary over time.



So far: Nonparametrically Unidentified

- The causal effect $Y_{j(t-1)} \rightarrow Y_{it}$ is not identified
- The association between $Y_{j(t-1)}$ and Y_{it} is a mixture of:
    1. The causal effect: $Y_{j(t-1)} \rightarrow Y_{it}$
    2. Homophily bias: $Y_{j(t-1)} \leftarrow U_j \rightarrow [A_{ij}] \leftarrow U_i \rightarrow Y_{it}$
    3. Confounding bias: $Y_{j(ti1)} \leftarrow C \rightarrow Y_{it}$
    4. Other paths (depending on what observables are conditioned on)



IV Criterion (nothing new here)

<u>Graphical IV Criterion (Brito and Pearl 2002):</u>

G is an IV for the total causal effect of treatment on outcome (conditional on variables V) if:

1. Relevance: Under the null, there is an open path (association) between G and treatment given V
2. Exclusion: Under the null, there are no open path (association) between G and outcome given V

As always, worry about

1. Existence
2. Strength

} Relevance

3. Unconfoundedness of the instrument
4. No other effects of IV on outcome

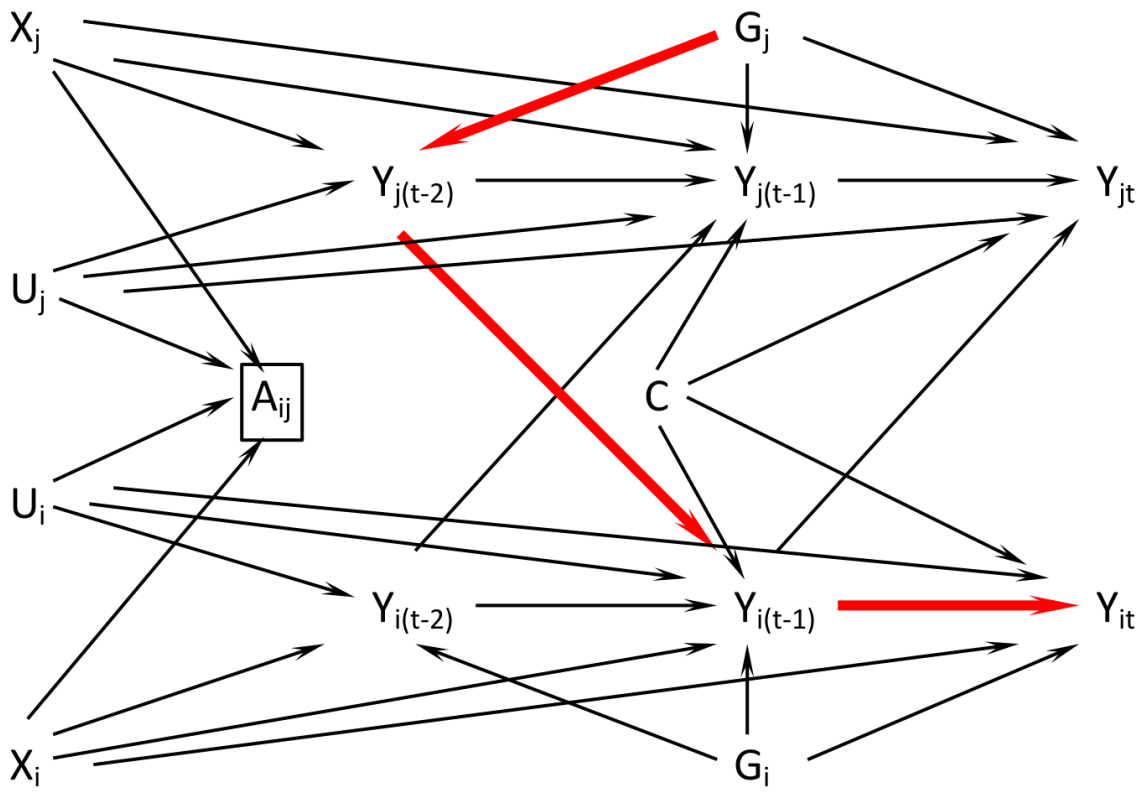
} Exclusion

IV fails if Genes Affect Past Treatments

Fact: “Fat genes” likely affect BMI at all times.

Problem: Exclusion violation via $G_j \rightarrow Y_{j(t-2)} \rightarrow Y_{i(t-1)} \rightarrow Y_{it}$ cannot be healed through conditioning

Note: M_{test} Encodes null of no effect



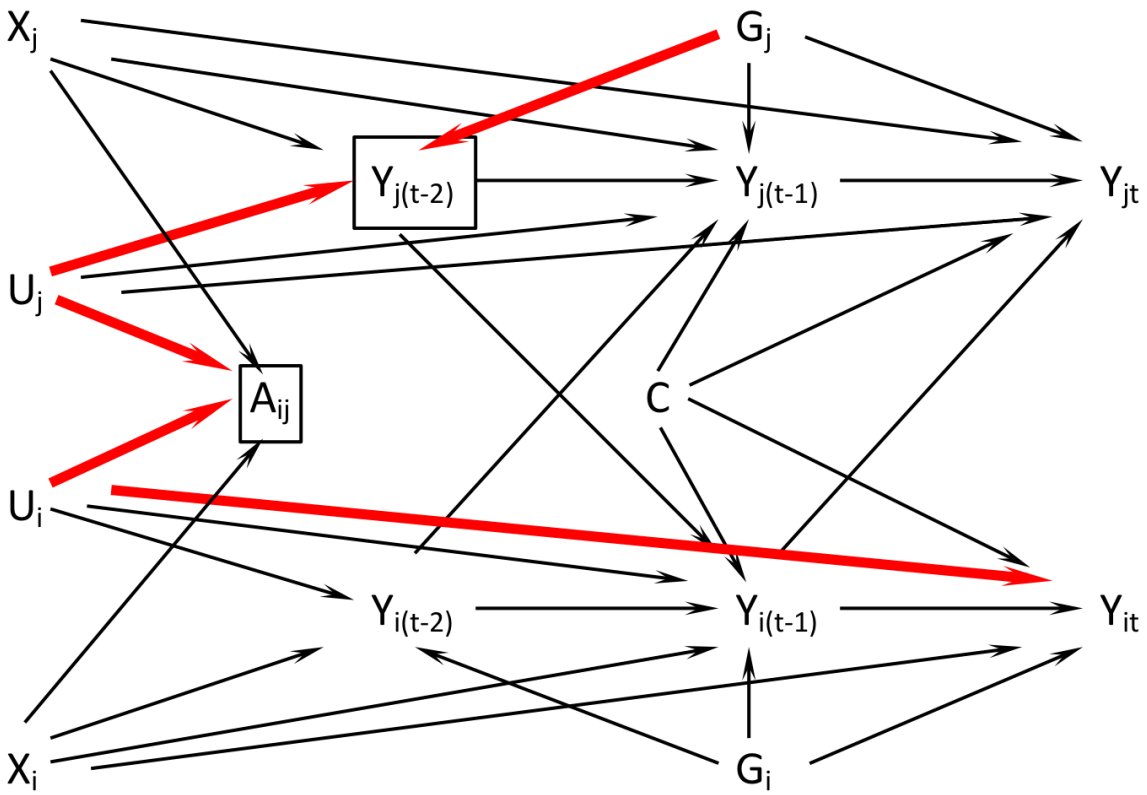
IV fails if Genes Affect Past Treatments

Fact: "Fat genes" likely affect BMI at all times.

Problem: Exclusion violation via $G_j \rightarrow Y_{j(t-2)} \rightarrow Y_{i(t-1)} \rightarrow Y_{it}$ cannot be healed through conditioning

1. Conditioning on the collider $Y_{j(t-2)}$ opens red path: exclusion violation

Note: D_{test} encodes Null of no effect



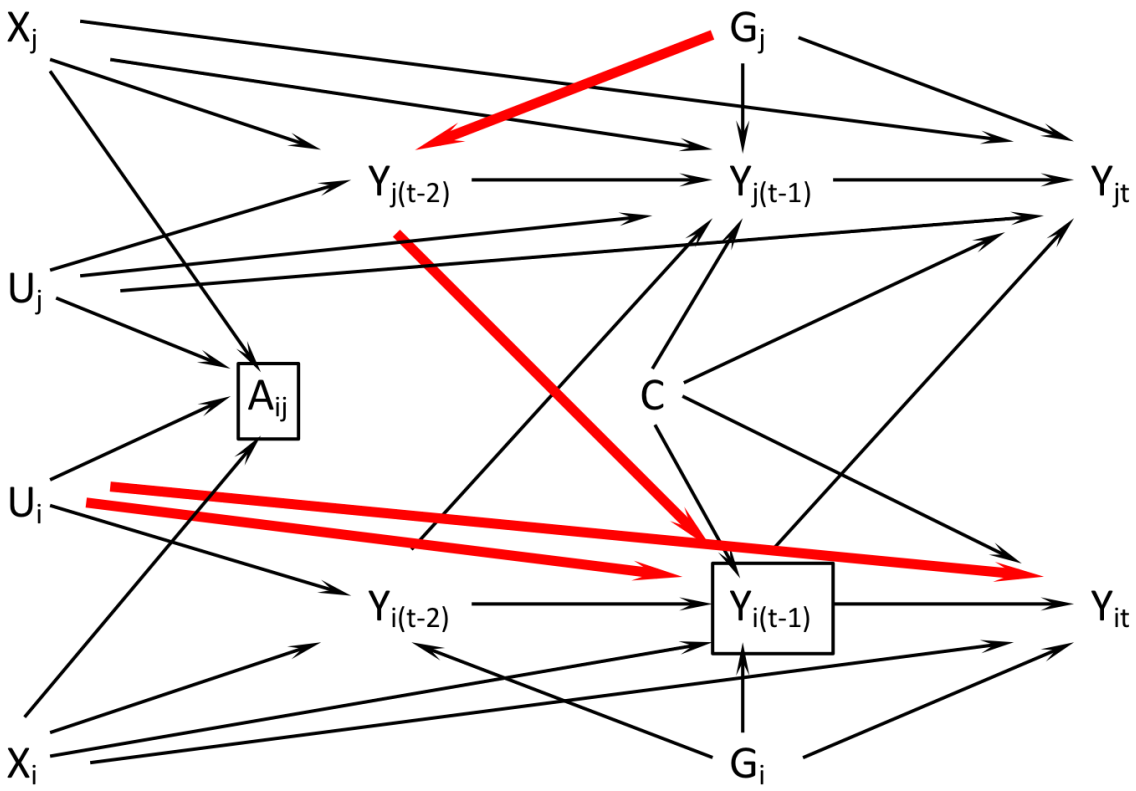
IV fails if Genes Affect Past Treatments

Fact: "Fat genes" likely affect BMI at all times.

Problem: Exclusion violation via $G_j \rightarrow Y_{j(t-2)} \rightarrow Y_{i(t-1)} \rightarrow Y_{it}$ cannot be healed through conditioning

1. Conditioning on the collider $Y_{j(t-2)}$ opens red path: exclusion violation
2. Conditioning on the collider $Y_{i(t-1)}$ opens red paths: exclusion violation

Note: M_{test} encodes Null of no effect

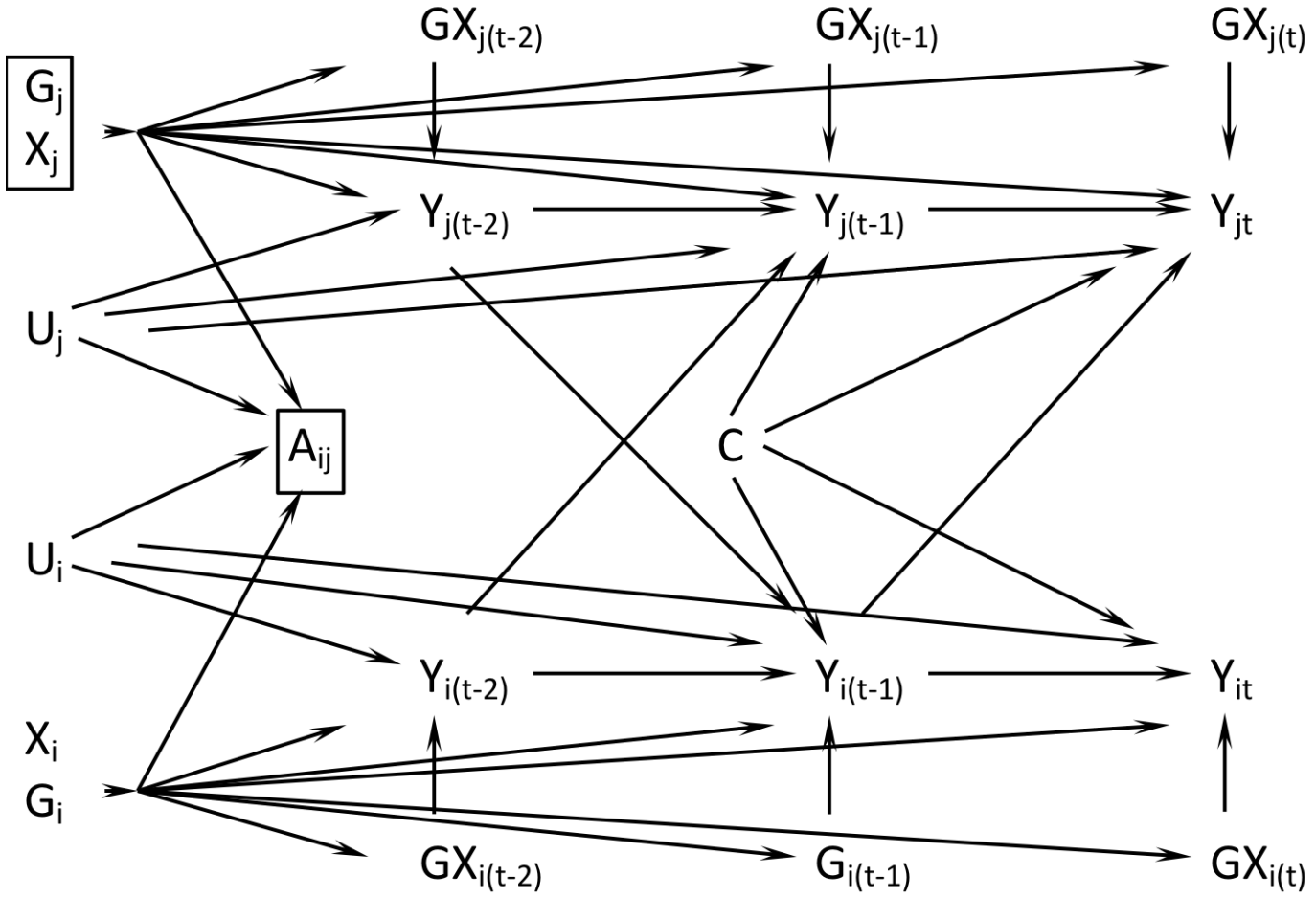


Time-Varying Gene Expression as IV

Trick: "Fat genes" affect BMI differently over age

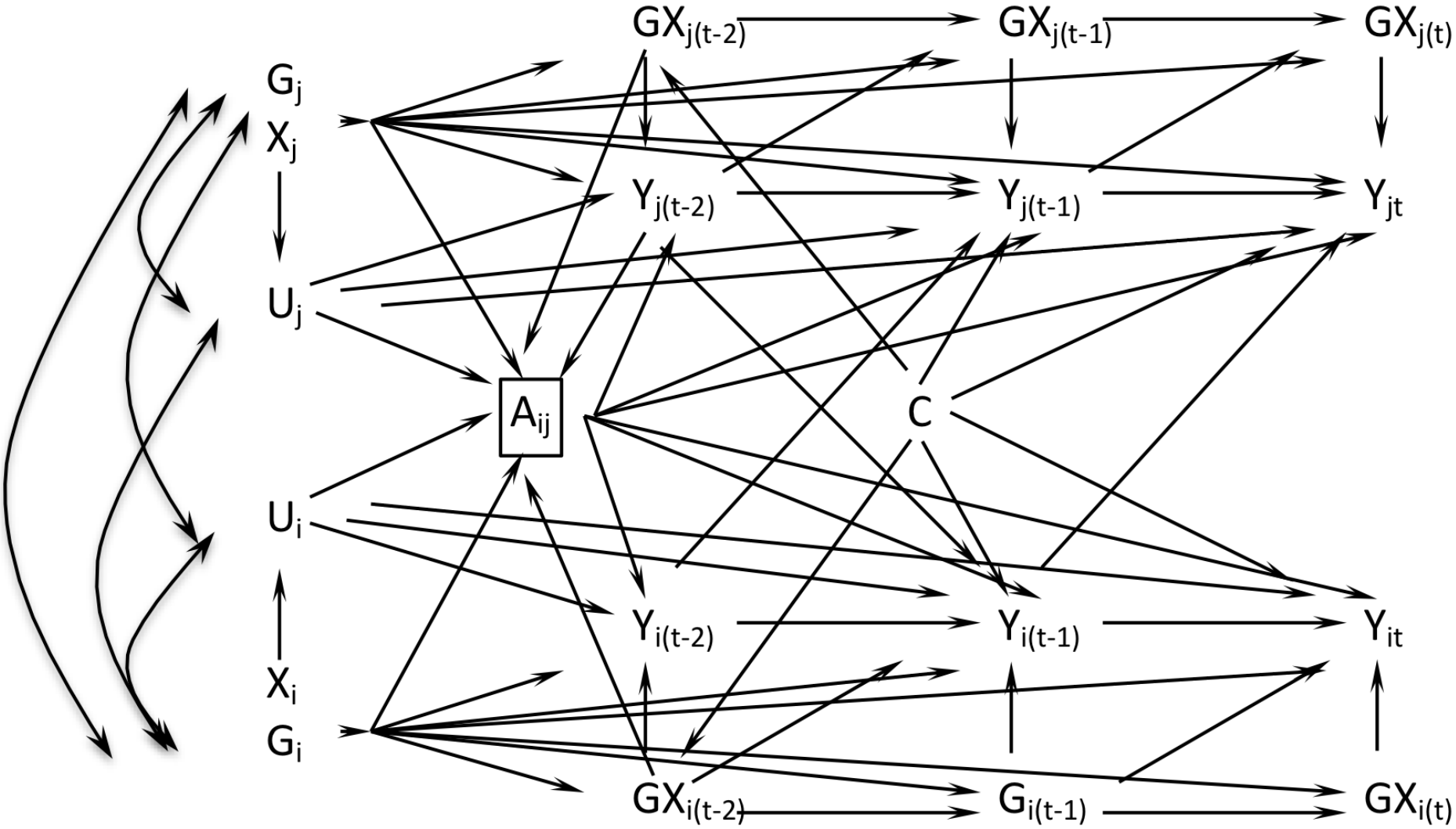
GX: Interaction between gene G and age X (time-varying "gene expression")

$GX_{j(t-1)}$ is a valid IV for peer effect $Y_{j(t-1)} \rightarrow Y_{i(t)}$ conditional on G_j, X_j, A_{ij}.



Accommodating Complications

$GX_{j(t-1)}$ remains a valid IV for the peer effect conditional on observables.



Acceptable Complications Include:

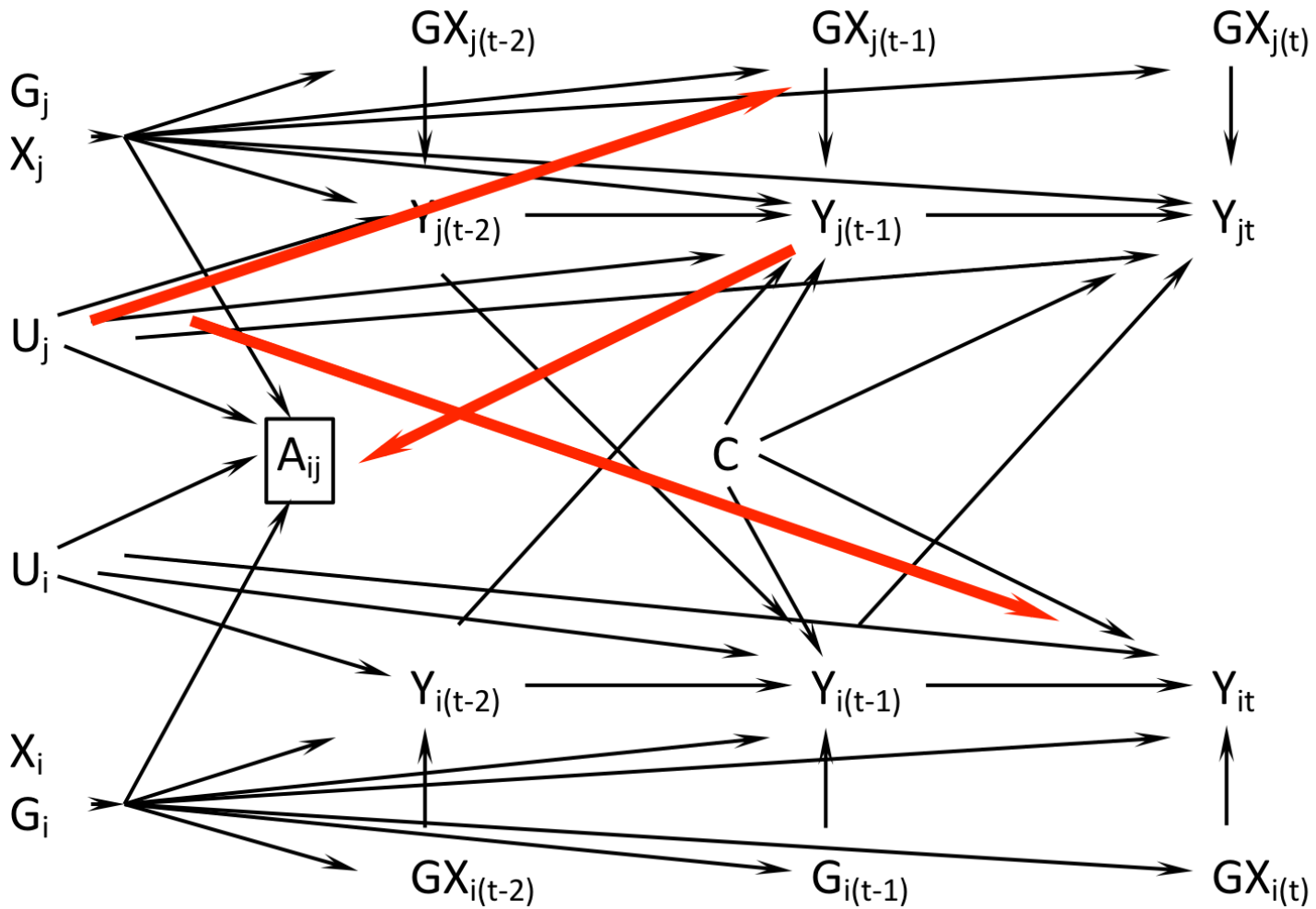
- Homophily on past phenotype $Y_{t-q} \rightarrow A_{ij}$
- Homophily on genotype $G \rightarrow A_{ij}$
- Pleiotropy on observables $G \rightarrow X$
- Pleiotropy on certain unobservables $G_k \rightarrow U_k$
- Population stratification $G_i \leftarrow U_s \rightarrow G_j$ (add dyad fixed effect)
- Inter-phenotype peer effects $\{X_j, U_j\} \rightarrow Y_i$
- Environmental confounding of
  past gene expression $C \rightarrow GX_{(t-2)}$
- " " present gene expression $C \rightarrow GX_{(t-1)}$ (add dyad fixed effect)
- Epigenetic effects $Y_{t-q} \rightarrow GX_{t-q+1}$
- Serial dependent gene expr. $GX_{t-q} \rightarrow GX_{t-q+1}$
- Tie effects $A_{ij} \rightarrow Y$

Pleiotropy

- Recognized problem for intra-individual IV identification:
  - j's gene likely affects j's outcome via lots of pathways other than the treatment phenotype—exclusion violation.
- Peer-effect inference likely more robust to pleiotropy. In our model:
  1. G_j <u>can</u> affect **everything** observed or unobserved; just condition on G_j
  2. G_j <u>can</u> even affect ego directly (though that's unlikely); just condition on G.
  3. GX_j <u>can</u> pleitropically affect observables X_j; just condition on X_j
  4. GX_j <u>can</u> pleiotropically affect unobservables U on the pathway to j's phenotype. This is in fact desirable—strengthens the first stage.
  5. GX_j <u>can</u> pleiotropically affect unobservables as long as those don't exert peer effects in i's outcome, as many unobservables won't.
  6. GX_j may <u>not</u> pleiotropically affect unobservables that exert peer effects in i's outcome. Has FTO been implicated strongly in other (non-BMI) phenotypes that may exert peer effects? If so, bounds, sensitivity analysis.

Some Irreparable Violations

- Contemporary homophily on phenotype $Y_{t-1} \rightarrow A$ (unlikely)
- No peer effects from latent homophilic traits $U_j \rightarrow Y_{it}$
- No role of latent homophilic traits in present gene expression $U_j \rightarrow GX_{j(t-1)}$



Conclusions

- Can we use genetic information as IV? Perhaps—depends.

- Exercise in how far we can roll back assumptions and still use genes as IV to identify peer effects.
  - Trick: use not alleles per se, but time-varying gene expression as IV
  - Accommodate confounding, latent homophily, various pleiotropic effects, population stratification, etc.

- Sticking points: relevance and exclusions.
  - Focused on defending exclusion
  - Price: permitting serially correlated gene expression and latent population stratification requires conditioning on predictors of the IV and adding dyad fixed effects, resulting in a very weak first stage.

What next?

- Other applications?
  - Problem of finding other specific genes with stronger first stage than FTO for phenotypes of interest to social scientists. Psychopathology?
  - Cannot switch from specific genes to polygenic scores because they inflate the threat of pleitropic exclusion violation.
- Current application
  - Results consistent with positive peer effects, but largely statistically inconclusive.
  - Need larger N.
  - In progress: ditch estimation, switch to nonparametric testing.

Supplementary slides

Data & Estimation

- Framingham Heart Study
  - Offspring cohort, seven waves, 1971-2003
  - 3,462individuals in N=9,270 dyads
  - Analyze two types of ties separately: friends & spouses
- Genes: FTO & MC4R
  - 4 dummies for 6 alleles
  - BMI association confirmed in replication (Speliotes et al. 2010)
- Covariates (demographics, smoking, location)
- 2SLS

First Stage: G, GX→BMI

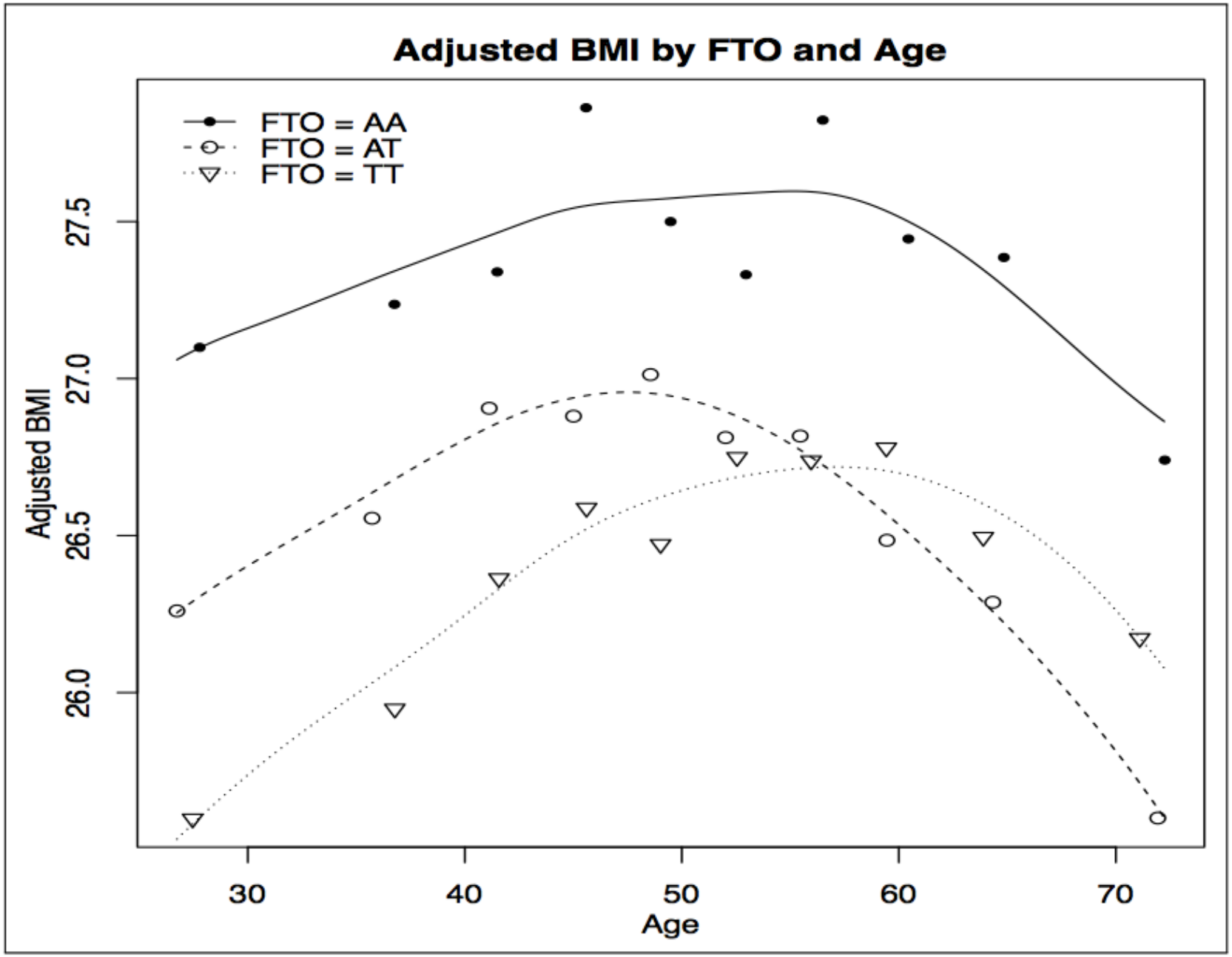


Table 2

Dyadic Peer Effect Analysis of Lag Alter BMI using Time-varying gene-age expression as an instrument

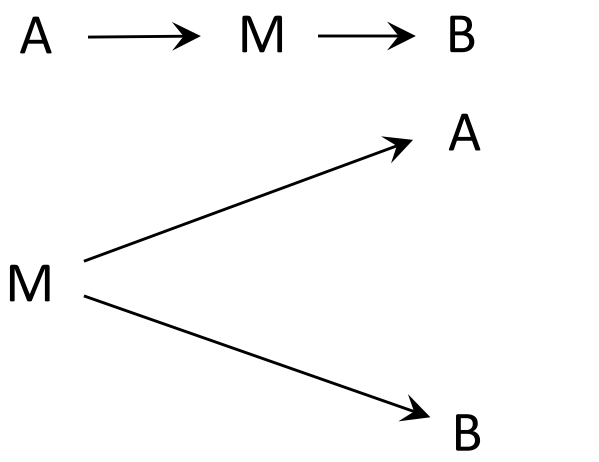
Discretionary $\boldsymbol{Z}_{(t)}$ Terms			IV Regression (2SLS)[a]			Regression (OLS)		
$\boldsymbol{GX}_{2(t-2)}$	$Y_{1(t-1)}$	F_5[b]	Estimate	95% CI		Estimate	95% CI	
			Nominated friend					
Exclude	Exclude	2.150	0.888	0.063	1.713	-0.011	-0.121	0.100
Exclude	Covariate	1.731	0.874	-0.031	1.779	0.009	-0.071	0.089
Covariate	Exclude	1.181	0.133	-0.796	1.062	-0.086	-0.193	0.021
Covariate	Covariate	1.144	-0.003	-0.911	0.906	-0.077	-0.181	0.028
			Spouse					
Exclude	Exclude	4.064	0.099	-0.324	0.522	0.066	0.039	0.094
Exclude	Covariate	4.351	0.101	-0.287	0.488	0.032	0.008	0.055
Covariate	Exclude	0.268	-0.102	-1.855	1.652	0.050	0.017	0.082
Covariate	Covariate	0.181	0.906	-1.832	3.643	0.023	-0.006	0.051

[a] $\boldsymbol{Z}_{(t)} = (\boldsymbol{GX}_{1(t)}, \boldsymbol{X}_{1(t)}, \boldsymbol{X}_{2(t)})$ are exogeneous covariates and $\boldsymbol{GX}_{2(t-1)}$ is an IV in all IV analyses. The elements of $\boldsymbol{X}_{k(t)}$, $k = 1, 2$, are: gender, age, gender-age interaction, birth era, birth year, smoking status, number of siblings, and (for $k = 1$ only) the geographic distance between residential locations of ego and alter at tie-formation. All models include dyad fixed effects. $\boldsymbol{GX}_{2(t-2)}$ and $Y_{1(t-1)}$ are added to $\boldsymbol{Z}_{(t)}$ as indicated in the two left-most columns.

[b] The F-statistic is for the overall effect of the IV, $\boldsymbol{GX}_{2(t-1)}$, in the first-stage equation. The critical value of the Cragg-Donald F-statistic, which quantifies the power of an IV, at the 20% level ranges from 6.71 to 6.77 across the models.

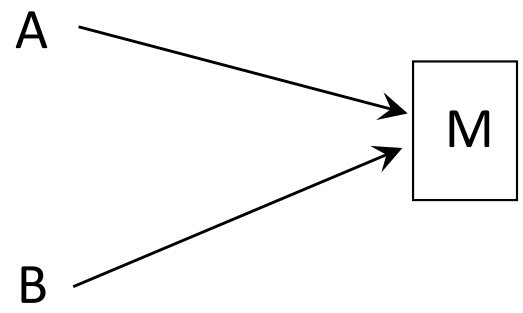
From Causation to Association

Two variables can be associated for three structural reasons in the population:

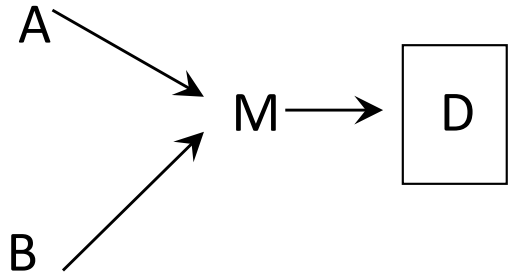


1. Causation: A causes B

2. Confounding: A and B share a common cause; common cause <u>is not</u> conditioned on.



3. Selection bias: A and B share outcome; common outcome <u>is</u> conditioned on. (→M← is a "collider")



Conditioning on descendant of collider also leads to selection bias

□ = conditioning

Reading Associations Off a Graph

Essence: Use Wright's path rules for linear models.

More generally: Path Principles (d-separation [Verma & Pearl 1988])

- All associations are transmitted along paths
- But not all paths transmit association

A path does not transmit association if it

1. Contains a causal chain A→M→B or a confounding fork A←M→B and the middle node M is conditioned on.
2. Contains a collider A→M←B and neither the middle node M, nor any descendants of M are conditioned on.

- Otherwise: path does transmit association.

Key points: Conditioning on a collider A→[M]←B opens a path.

Conditioning on a non-collider A←[M]→B blocks the path.