The causal effect of UK firms' investment on their productivity

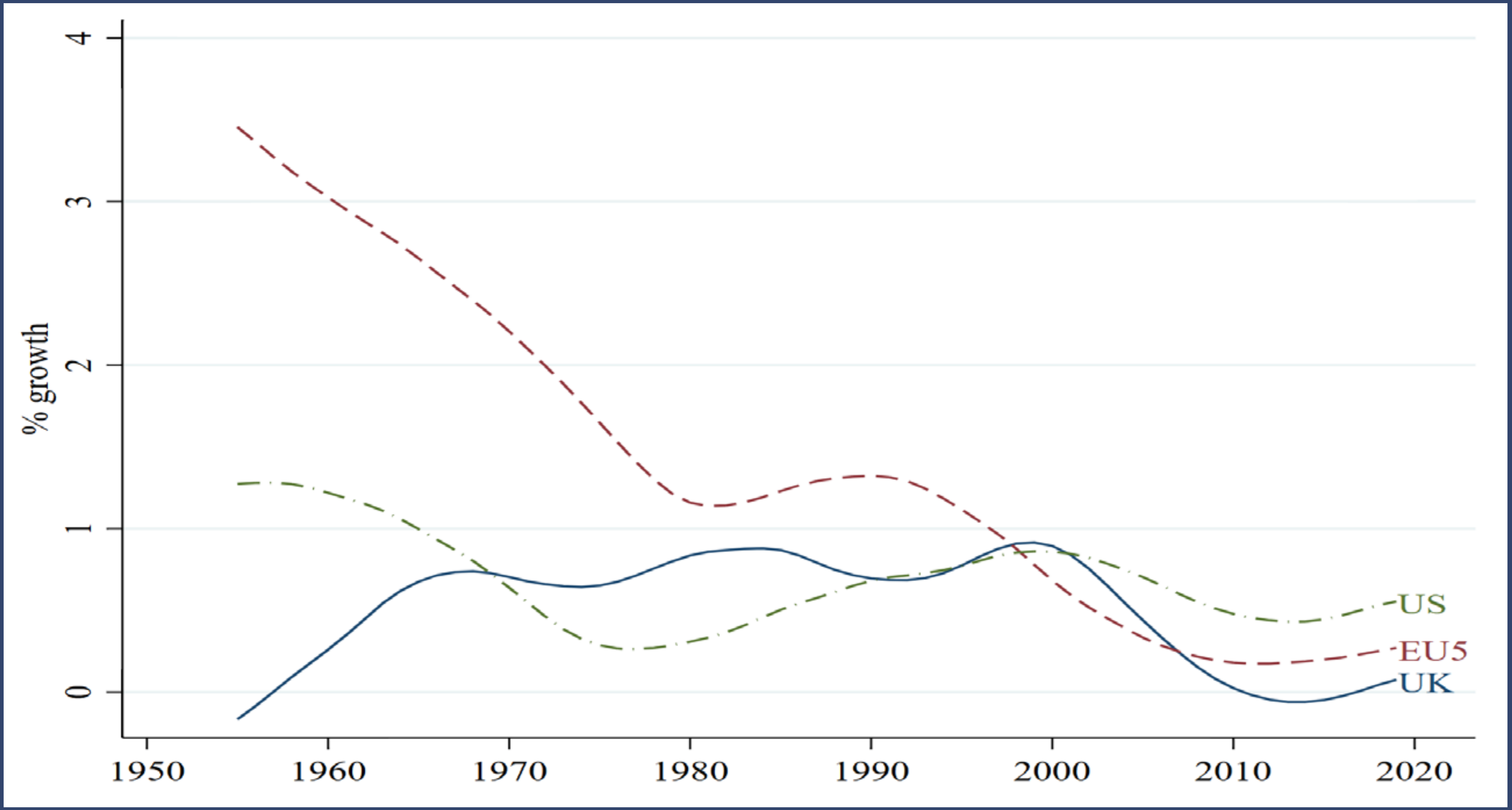
Dimitris Christelis
University of Glasgow

Ali Gokhan Karabilgin
University of Glasgow

John Tsoukalas
University of Glasgow

Motivation

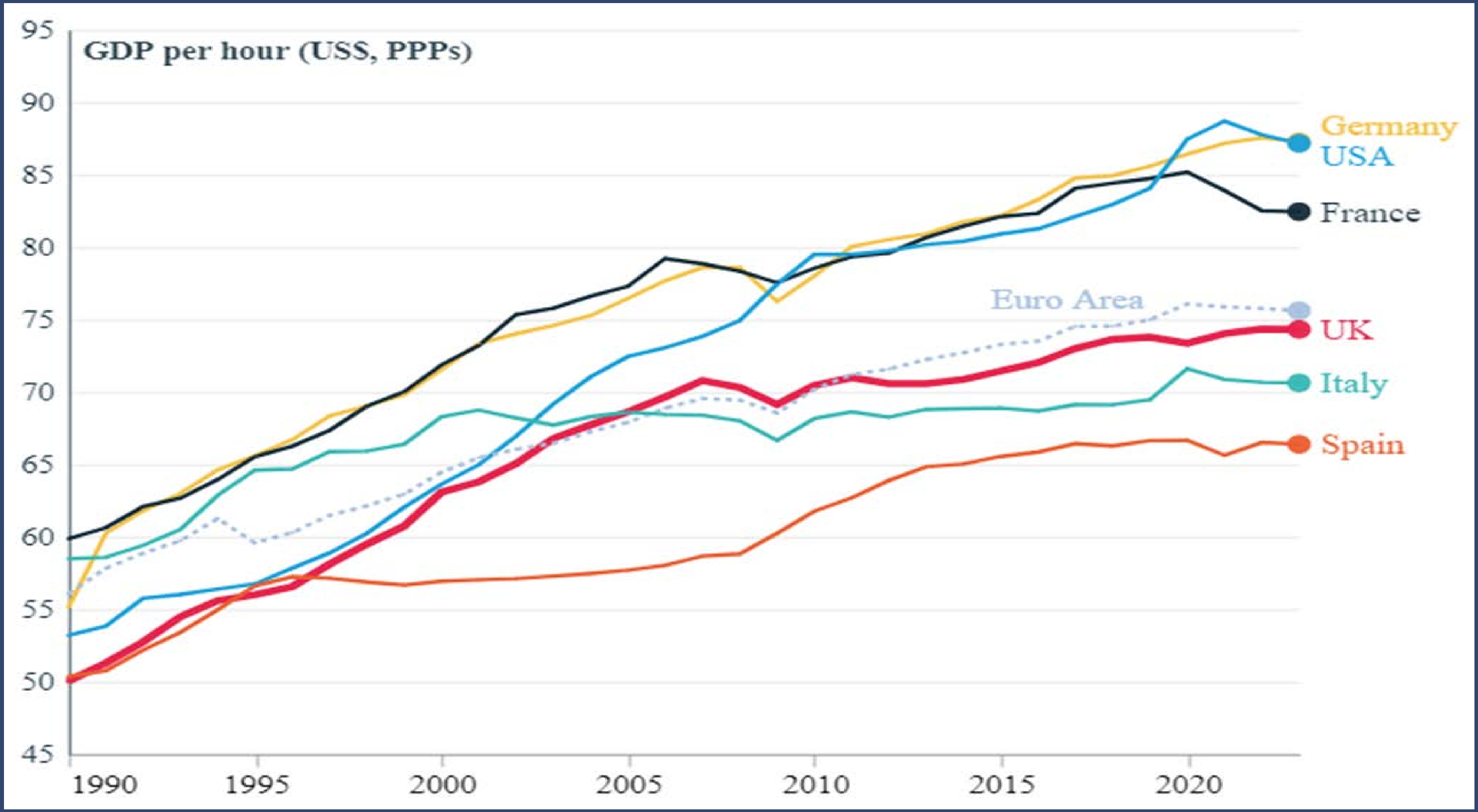
Trend TFP growth in the UK, Northern Europe, and U.S.



Source: J. Fernald, R. Inkaar (2022) The UK Productivity “Puzzle” in an International Comparative Perspective. Working Paper No. 020,The Productivity Institute.
The EU-5 group’s (Germany, France, the Netherlands, Belgium, and Finland, ranked by their 2010 GDP).

Motivation (cont.)

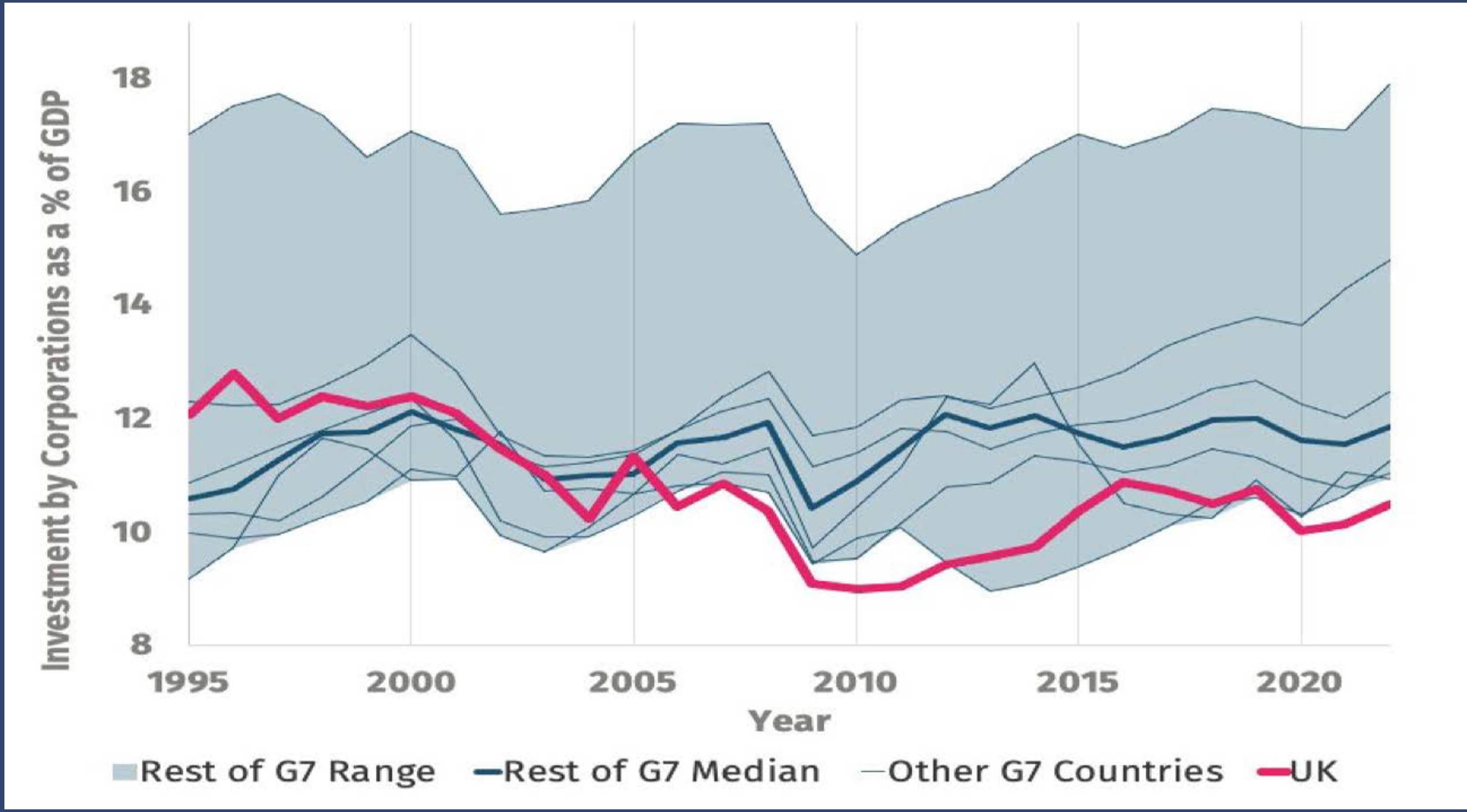
GDP per hour worked (in US$, PPP converted), 1990-2023



Source: Ark, Bart van & O'Mahony, Mary, 2024, What explains the UK's productivity problem? Insights Blog, The Productivity Institute, Retrieved from https://www.productivity.ac.uk/news/what-explains-the-uks-productivity-problem/

Motivation (cont.)

Corporate investment as a % of GDP



Source: IPPR, 2024, Rock bottom: Low investment in the UK economy

The link between investment and productivity

- Labor productivity (LP):
  - more capital generally makes workers more productive
  - feature of all standard production functions

- Total factor productivity(TFP):
  - not as easy to reason about as LP because TFP is unobservable and a residual
  - embodied technical change in more recent capital vintages allows management and workers to be more productive for any given amount of capital – could an issue regarding the measurement of capital services
  - increased output, facilitated by increased investment, could lead to accelerated learning by doing
  - investment in physical capital can lead to larger TFP when combined with investment in intangibles and/or organizational change

Causality

- Causal statements are very difficult to make in our context:

  - investment is an endogenous decision: firms that invest more could also be on average, better run, or have better (expected) profit opportunities – systematically different

  - anticipated productivity gains could affect current investment

Causality (cont.)

- Panel data: cannot deal with time-varying unobservables affecting both investment and productivity (e.g., management quality, profit opportunities, new export markets). Panel data also require strict exogeneity

- IV: exogenous instruments very hard to find, LATE interpretation

- Diff-in-diff: high investment firms likely have different future productivity trajectories from low investment firms due, for example, to differences in management quality, profit opportunities even under low investment – violation of parallel trends assumption. Estimates ATT, not ATE

Contribution

- We derive informative identification regions for the average treatment effect (ATE) of UK firms' investment on their productivity using partial identification (PI) methods in an observational data context

- PI results are robust to endogeneity

- PI relies on weak, and thus credible assumptions

Literature

- Numerous papers document positive associations between investment (often measured in spikes) and productivity using panel data methods (e.g., Nilssen, 2009)

- Gradzewicz (2021), using Polish data and a diff-in-diff methodology, finds that LP increases by ≈3% after a spike, driven by capital deepening

- Karmakar et al. (2024), using UK firms and panel data methods, find that intangible investment raises TFP (≈ 9.7 % for firms in the top quartile of intangibles); tangible investment has no effect on TFP

Data

- FAME database – standard source of firm data for the UK
- Data from 2001 to 2023 – we stop at 2019 to exclude the labor hoarding effects of the pandemic period
- Balance sheet and profit loss statements available
- Output: gross value added
- Labor input: number of employees (not hours)
- TFP is calculated using the methodology in Ackeberg et al (2015, a modification of Olley and Pakes, 1996)

Data (cont.)

- We examine tangible investment (lagged, scaled by lagged capital)

- Since firms can dispose of capital, investment is net of disposals and thus can be negative (about 10% of our sample has negative investmnet)

- Crucially, disinvestment can be rational – downsizing of firms can lead to productivity gains and better performance

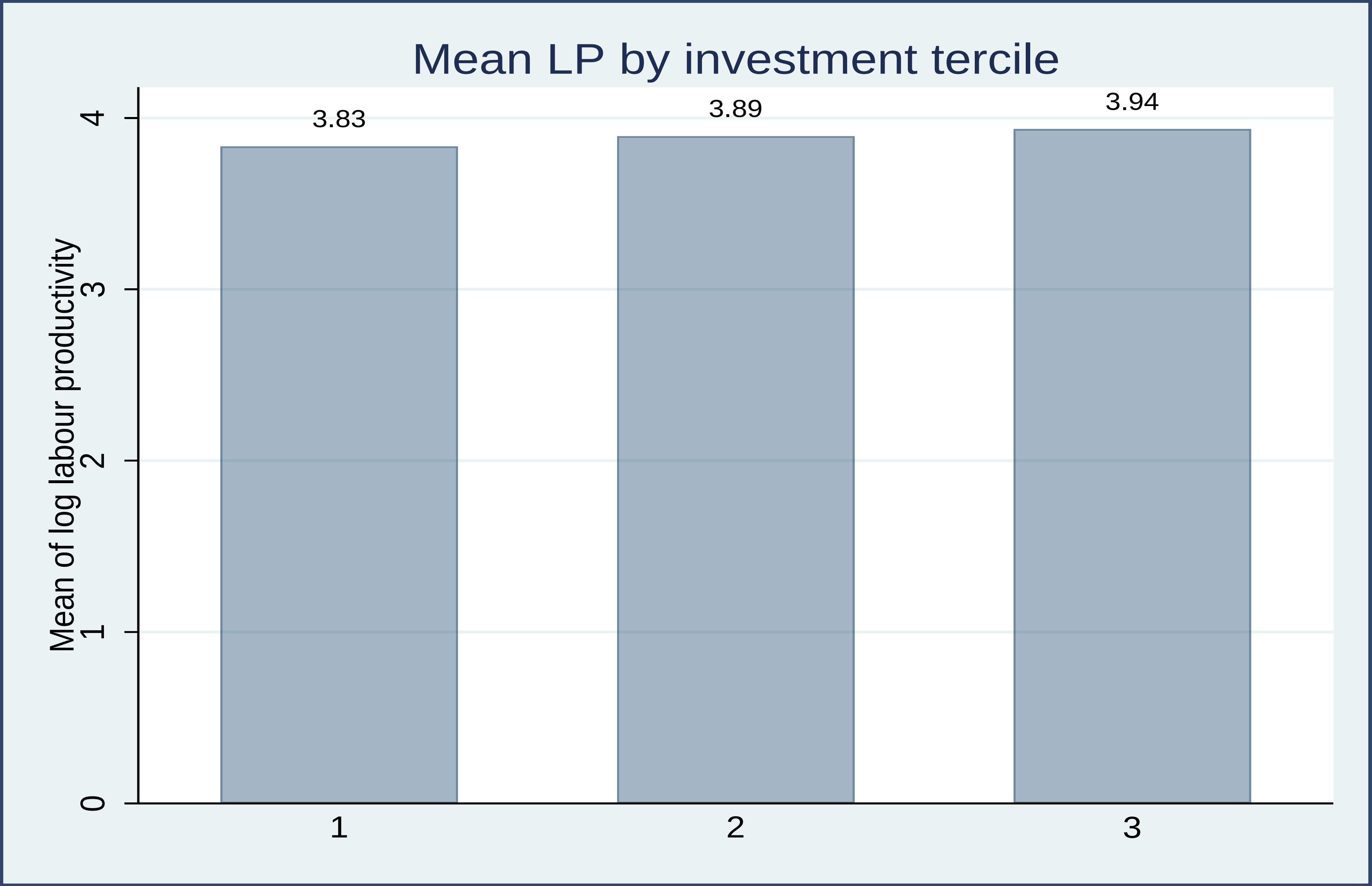
- We only look at firms with positive gross investment – can lead to further endogeneity, but PI accommodates endogeneity of any form

Data (cont.)

- We exclude agriculture, finance, public sector firms, and the household sector

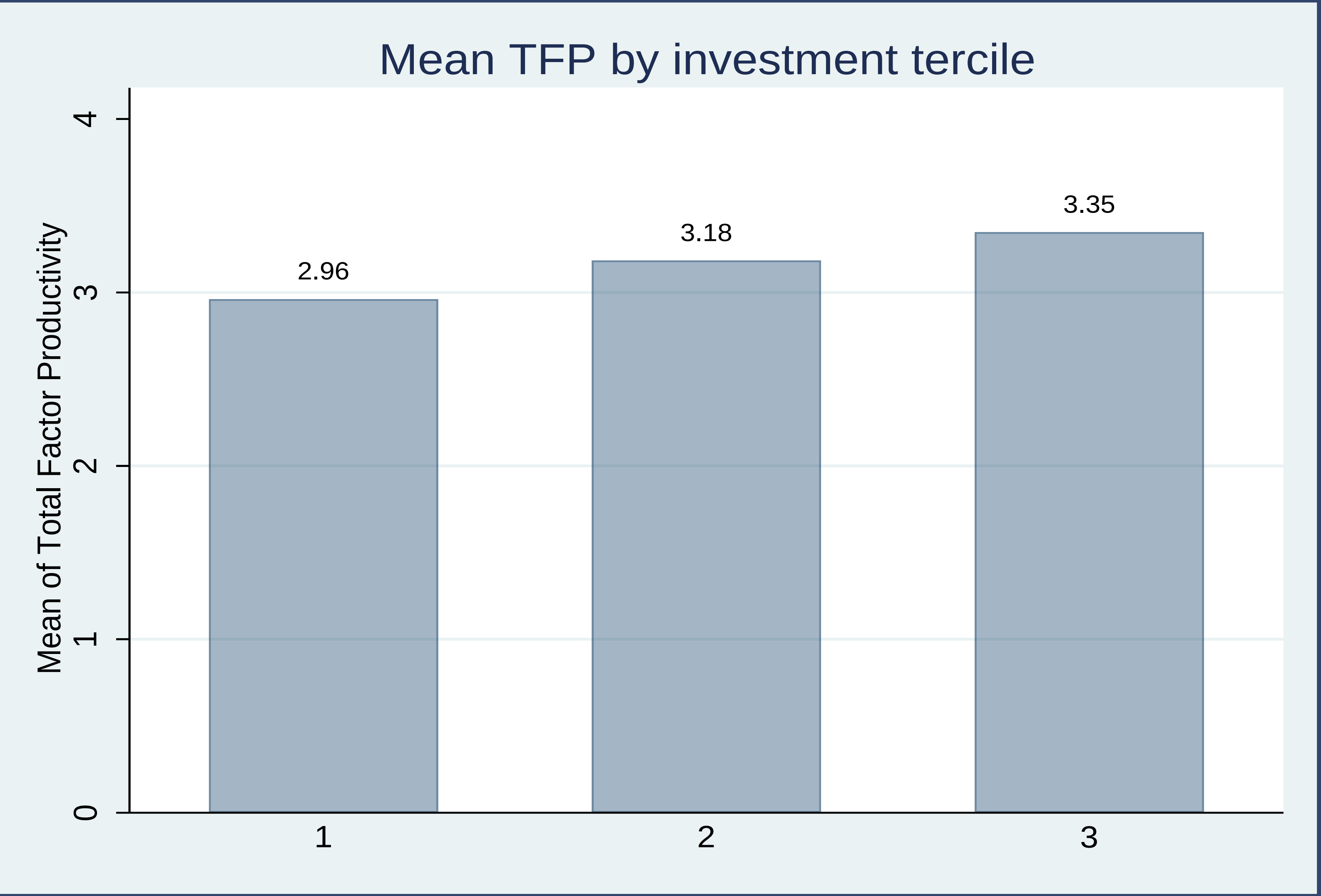
- Number of observations: about 358,400 for LP, about 205,450 for TFP

Data (cont.)



Mean LP by investment tercile
Mean of log labour productivity
0
1
2
3
4
3.83
3.89
3.94
1
2
3

Data (cont.)



Mean TFP by investment tercile
Mean of Total Factor Productivity
0
1
2
3
4
2.96
3.18
3.35
1
2
3

Partial identification

- We use partial identification methods (Manski, 1990; Manski, 1997; Manski and Pepper, 2000)

- Starting premise:
  - the treatment is endogenous (any form of endogeneity can be accommodated) and there is no exogenous variation that can help point identify the ATE

- PI puts bounds around counterfactual outcomes, and thus around the average treatment effect (ATE), using credible assumptions

Partial identification

- Average treatment effect (outcome Y, observed treatment w, generic treatment value d, controls x) :

$$E[Y(d_2)|x] - E[Y(d_1)|x]$$

- Average potential outcome:

$$E[Y(d)|x] = E[Y(d)|x, w=d]P(w=d|x)$$
$$+ E[Y(d)|x, w\neq d]P(w\neq d|x)$$

Partial identification (cont.)

- Observed sample analogues of $E[Y(d)|x,w=d]$, $P(w=d|x)$, $P(w\neq d|x)$

- Unobserved counterfactual $E[Y(d)\,|\,x,w\neq d]$

- Exogeneity implies $E[Y(d)\,|\,x,w\neq d]=E[Y(d)|x,w=d]$. Thus

$$E[Y(d_2)|x]-E[Y(d_1)\,|x]=$$
$$E[Y(d_2)|x,w=d_2]-E[Y(d_1)|x,w=d_1]$$

Problem of the counterfactual

- Two approaches:
  - statistical models
  - models modelling explicitly decision-making

- Both approaches rely on exogeneity
  - error distributions are assumed to have no connection to observables or other parameters in the model
  - the same estimating equation/decision rule can be applied irrespective of actual treatment received
  - all that is left out of the model is unsystematic

- PI postulates that one cannot included all systematic factors in the model; hence exogeneity cannot be assumed, which leads to $E[Y(d) \mid x, w \neq d] = E[Y(d) \mid x, w = d]$

Partial identification

- PI, instead of assuming $E[Y(d)|x,w\neq d]=E[Y(d)|x,w=d]$, put bounds around $E[Y(d)|x,w\neq d]$. Endogeneity for any reason can be accommodated as all reasons lead to the same implication: $E[Y(d)|x,w\neq d]\neq E[Y(d)|x,w=d]$

- Bounds on $E[Y(d)|x,w\neq d]$ imply bounds on average potential outcomes, and bounds on the differences in average potential outcomes, that is, average treatment effects

- PI is non-parametric, relies on few and credible assumptions, and is absolutely transparent how each assumption affects results.

Partial Identification (cont.)

- For example,

$$2 \leq E[Y(d_1)|x] \leq 3$$
$$5 \leq E[Y(d_2)|x] \leq 7$$

Then,

$$5 - 3 = 2$$
$$\leq E[Y(d_2)|x] - E[Y(d_1)|x)] \leq$$
$$7 - 2 = 5$$

- To narrow identification regions, we need lower bounds that are as large as possible, and upper bounds that are as small as possible

Partial Identification (cont.)

- Without assumptions, the only credible bounds on the counterfactual terms are the minimum and maximum possible values, which typically produce uninformative identification regions

$$E[Y(d) \mid x, w=d] P(w=d|x) + y_{min} P(w \neq d|x)$$
$$\leq E[Y(d)|x] \leq$$
$$E[Y(d) \mid x, w=d] P(w=d|x) + y_{max} P(w \neq d|x)$$

- Using assumptions, we will replace y_{min} with something larger, and y_{max} with something smaller

Assumption 1: Monotone Treatment Response (Manski, 1997)

- Higher investment does not decrease productivity, ***on average***, ***keeping everything else constant.*** For $d_2 > d_1$ and $\forall d$

$$E[Y(d_2)|w = d] \geq E[Y(d_1)|w = d)]$$

- Weak monotonicity, thus equality allowed - not sufficient to exclude zero from the ATE identification region

- Plausible assumption – difficult to think why higher investment would reduce productivity, on average and conditional on investment being positive. Has strongest effect on ATE lower bounds

- How it works:

$$E[Y(d_1)|w = d_2)] \leq E[Y(d_2)|w = d_2]$$

Assumption 2: Monotone Treatment Selection (Manski and Pepper, 2000)

- Higher investment firms have weakly higher productivity than lower investment firms for any common level of investment, ***on average.*** For $d_2 > d_1$ and $\forall d$

$$E[Y(d)|w = d_2] \geq E[Y(d)|w = d_1)]$$

- Weak monotonicity, thus equality allowed. MTS characterises the direction of treatment endogeneity

- Plausible assumption – higher investment firms likely to have better investment opportunities and more far-sighted management – on average and conditional on positive investment. Has strongest effect on ATE upper bounds

- How it works:

$$E[Y(d_1)|w = d_2)] \geq E[Y(d_1)|w = d_1]$$

Joint test of MTR+MTS (Manski and Pepper, 2000)

- Very important to test as many assumptions as possible for the credibility of results

- Manski and Pepper (2000) show that MTR+MTS implies that

$$E[Y(d_2)|w=d_2] \geq E[Y(d_2)|w=d_1] \geq E[Y(d_1)|w=d_1)]$$

The first inequality holds due to MTS, and the second due to MTR.

- Very easy to test this condition regarding weak monotonicity of observed conditional mean outcomes, which clearly holds in our data

Assumption 3: Monotone Instrumental Variable (Manski and Pepper, 2000)

- MIVs can be weakly monotonically associated with the outcome, and are thus much easier to find
- They do not identify the LATE but the ATE – nothing to do with compliers
- An instrument z is monotone if

$$z_1 < z_2 \Rightarrow E[Y(d)|Z = z_1] \leq E[Y(d)|Z = z_2]$$

- Assumption is about potential outcome $E[Y(d)]$, hence it is untestable. It also reflects a simple correlation, not a causal relationship

Monotone instrumental variable (MIV)

- MIV condition allows a directed search in the sample space, through conditioning on the MIV, for larger lower bounds and smaller upper bounds

- Our MIV: twice lagged GI/K
  - Assumption: given last year's GI/K, the higher GI/K was two years ago, the weakly higher current LP and TFP are
    - Weak assumption, denotes only a weakly positive association, on average, between GI/K and productivity two years apart

PI results – 3d vs 1st tercile of GI / K

Assumptions	Labor Productivity		Total Factor Productivity	
	Lower bound	Upper bound	Lower bound	Upper bound
Exogenous treatment selection	0.101		0.386	
No assumptions	-12.188	12.255	-11.907	12.165
MTR	0.000	12.255	0.000	12.165
MTS	-12.188	0.101	-11.907	0.386
MIV	-11.789	6.637	-10.953	6.638
MTR+MTS	0.000	0.101	0.000	0.386
MTR+MIV	0.014	6.637	0.074	6.638
MTS+MIV	-11.789	0.044	-10.953	0.239
MTR+MTS+MIV	0.014	0.044	0.074	0.239
Number of observations	358,393		205,449	

PI results– remaining ATEs under MTR+MTS+MIV

Assumptions	Labor Productivity		Total Factor Productivity	
	Lower bound	Upper bound	Lower bound	Upper bound
Panel A: second vs first tercile of I/K				
MTR+MTS+MIV	0.001	0.038	0.043	0.166
Panel B: third vs second tercile of I/K				
MTR+MTS+MIV	0.000	0.023	0.000	0.140
Number of observations	358,393		205,449	

Conclusions

- We find a positive and economically substantial causal effect of firm investment on both firm LP and firm TFP using micro data from FAME
- Results are obtained via PI methods using weak, and thus credible, assumptions
- Observational data (even cross-sectional ones) can be used to make causal statements, without the use of exogenous variation
- Future work: quantile treatment effects, the effect of intangible investment, incorporation of negative investment firms

THANK YOU!

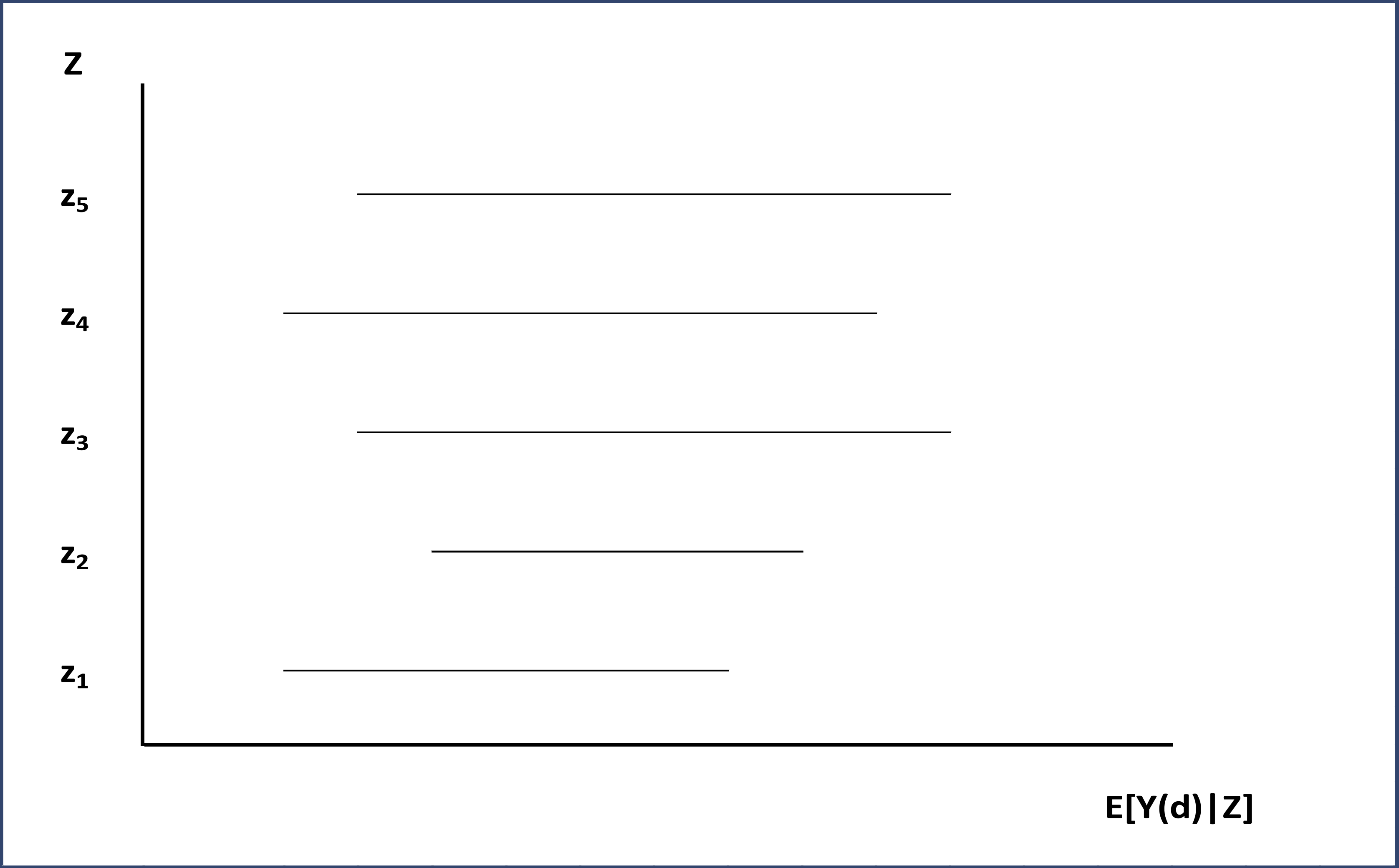
How do MIVs work? - XIVs

- Let us assume for the moment that the instrument is exogenous (XIV)

$$E[Y(d)|Z = z] = E[Y(d)], \forall z$$

- Need to narrow the identification region
  - LB example: $LB(d) = \sum_z LB(d|Z = z)P(Z = z)$
  - need to increase $LB(d|Z = z)$, as $P(Z = z)$ is given by the data
  - Correspondingly, need to decrease $UB(d|Z = z)$

How do MIVs work? - XIVs (cont.)



How do MIVs work? - XIVs (cont.)

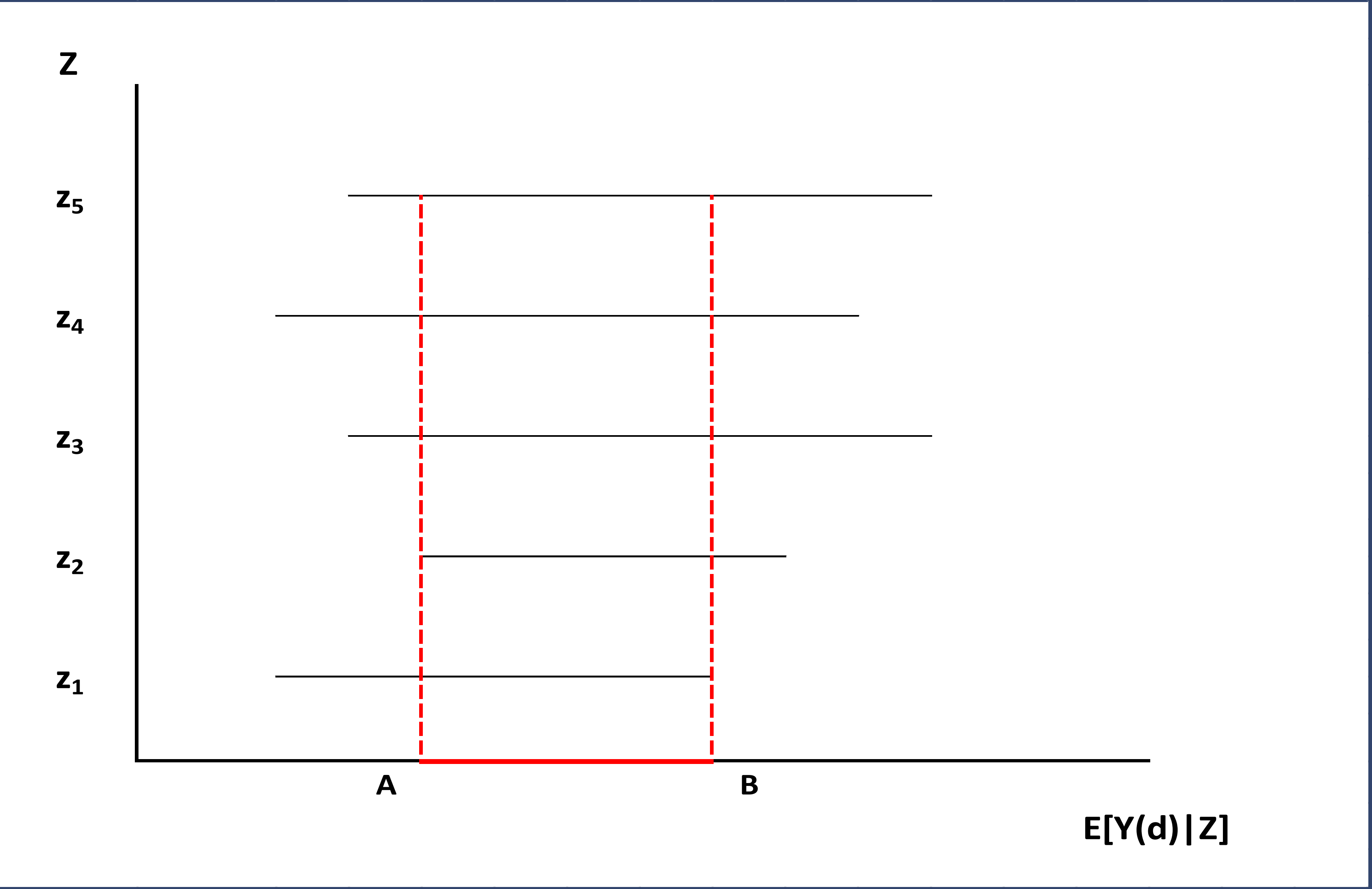
- Since the instrument is exogenous, results from all conditional identification regions are valid for the population

- All regions give valid answers → the intersection of all identification regions

- The intersection is given by the maximum lower bound and the minimum upper bound

$$\max_z LB[d|Z=z] \leq E[Y(d)] \leq \min_z UB[d|Z=m]$$

How do MIVs work? - XIVs (cont.)

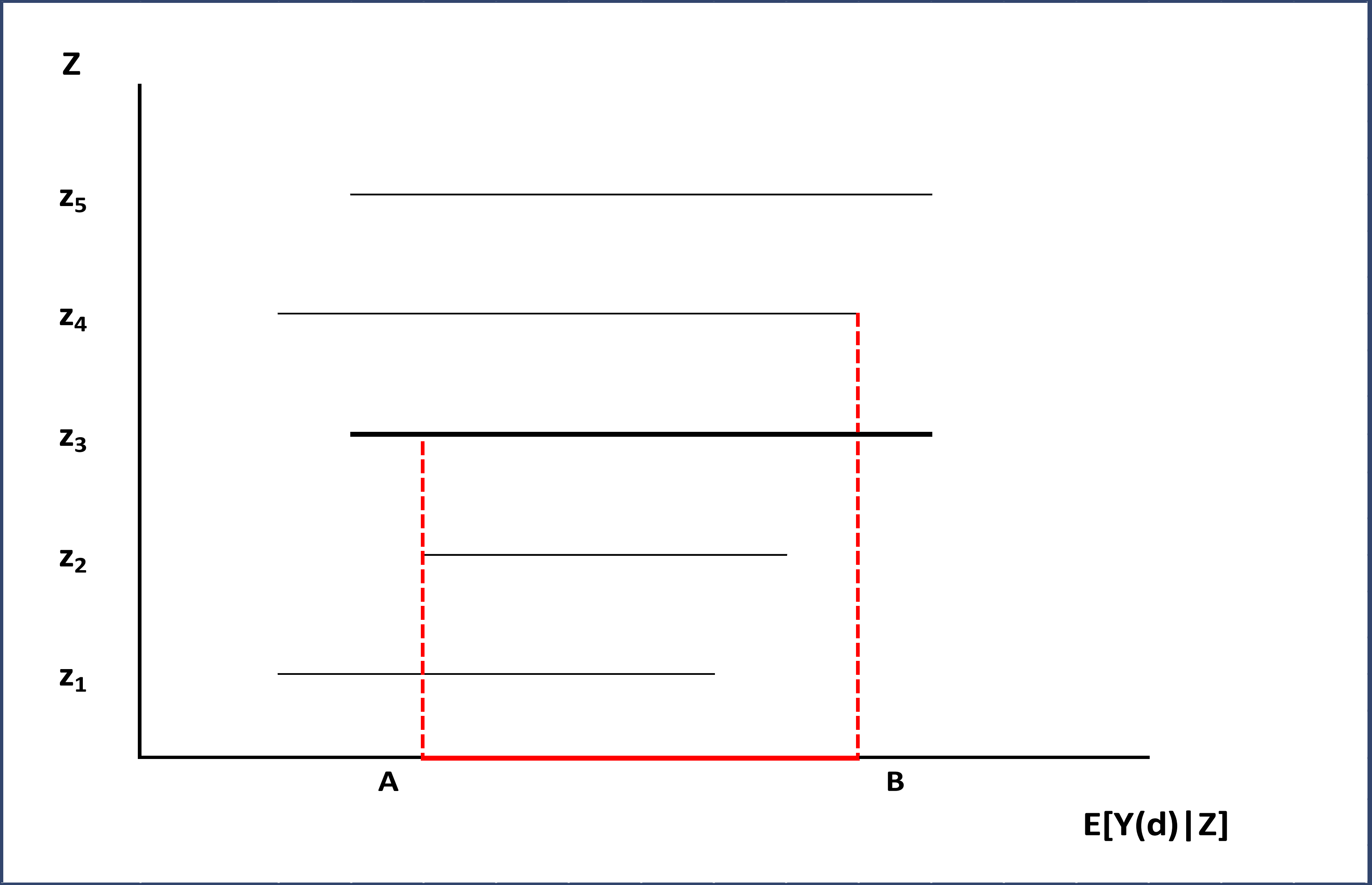


How do MIVs work?

$$z_1 \leq z_2 \rightarrow E[Y(d)|Z = z_1] \leq E[Y(d)|Z = z_2]$$

- For a given MIV value can examine only what happens to
  - lower bounds at lower instrument values
  - upper bounds of higher instrument values

How do MIVs work?



How do MIVs work? (cont.)

- Hence, the final bounds for the identification region conditional on MIV value z_3 are:
  - the maximum lower bound over z_1, z_2, z_3, which is the one corresponding to MIV value z_2
  - the minimum upper bound over z_3, z_4, z_5, which is the one corresponding to MIV value z_4

- Analogous calculations performed for identification regions conditional on MIV values other than z_3

- Finally, overall bounds are computed by integrating out the MIV

How do MIVs work? (cont.)

- Thus, optimization takes place only for a restricted range of values – identification regions less informative

- To be expected, as the MIV assumption is weaker than the XIV one

MIVs (cont.)

- MIVs partition the sample space into cells defined by instrument values and look for minima (maxima) of upper (lower) bounds
- In the case of MIVs, the search is directed
- Typically, the finer the partition (achieved, e.g., by using more than one MIVs) the narrower the identification regions
- Local minima and maxima are perfectly OK – they just lead to wider identification regions
- Different instruments possibly lead to different identification regions – this is fine, as the unobserved average potential outcomes do not change
- Constraint: number of observations