The Causal Effect of Facemask-wearing on Symptomatic Infection from COVID-19: Observational Data Evidence from 26 European Countries and Israel

Dimitris Christelis
University of Glasgow

Loretti I. Dobrescu
University of New South Wales

Anna Sanz-de-Galdeano
University of Alicante

COVID-19 and masks

- Important, controversial, and politicized issue

- Need evidence, obtained as transparently as possible

- Why would masks protect from Covid-19?
  - make it more difficult that droplets (eventually transformed into aerosols) from others' mouths entering through the nose/mouth
  - wearing a mask traps the moisture that would otherwise escape when exhaling. This raises the amount of moisture in one's airways. Moisture is essential to promote the natural clearance process of the airways. This helps keep the lungs free of contamination by dust, pollutants, and viruses. Masks may aid this process by raising the humidity.

Causality

- Causal statements are very difficult to make:
  - mask wearing is endogenous – those who wear masks are likely also more cautious on average, which can drive down the probability of infection
  - lack of credible exogenous variation
  - treatment spillovers (violation of the SUTVA) are very likely because mask wearing by one person can affect the probability of infection of another – results from standard methods (including RCTs) likely invalid when not accounting for the problem
  - there could be reverse causality from infection to mask wearing
  - cross-sectional data
  - very large cross-country variation in measures taken, work arrangements, transportation, population density etc.

COVID-19 and masks

- Most of the literature relies on methods that use selection on observables (OLS/LDV panel data, matching, etc.) or DiD, which relies on untestable parallel trends assumption. Evidence is mixed, but mostly points to beneficial effect of masks

- Abaluck et al. (Science, 2022) perform an RCT in Bangladesh (about 340,000 participants)
  - to mitigate the spillover problem randomization is done at the village level – statements about effects can only be made at the village level
  - distances between villages as low as 2km
  - they find a lower infection rate of 8 pp in villages in which masks were distributed

Contribution

- Derive informative identification regions for the average treatment of mask-wearing on symptomatic COVID infection at the individual level using partial identification (PI) methods in observational data
- PI results are robust to both endogeneity and violation of the SUTVA
- We rely on weak, and thus credible assumptions

Data

- Survey of Health, Ageing and Retirement in Europe (SHARE). Special module administered by phone in June 2020 in 26 European countries and Israel to samples representative of the 50+ population

- Countries: Germany, Sweden, Netherlands, Spain, Israel, Italy, France, Denmark, Greece, Switzerland, Belgium, Czech Republic, Poland, Luxembourg, Hungary, Portugal, Slovenia, Estonia, Croatia, Lithuania, Bulgaria, Cyprus, Finland, Latvia, Malta, Romania and Slovakia

- About 42,000 observations

Data (cont.)

- Q1: Since the outbreak of Corona, did you or anyone close to you experience symptoms that you would attribute to the Covid illness, e.g., cough, fever, or difficulty breathing?

- Q2: Have you or anyone close to you been hospitalized due to an infection from the Corona virus?

- Construct a measure of symptomatic infections from Q1 and Q2

- Measurement error: false positives, false negatives

Data (cont.)

- Q: How often did you wear a face mask when you went outside your home to a public space? Was it always, often, sometimes, or never?

  1. Always
  2. Often
  3. Sometimes
  4. Never

- Construct a binary indicator for mask wearing by combining 1,2 (=1) and 3,4 (=0)

Country	Had symptoms consistent with COVID-19	Was hospitalized due to Covid-19	Tested positive for COVID-19	Wears masks always or often	Number of observations
Belgium	0.0371	0.0031	0.0075	0.6392	3,290
Bulgaria	0.0159	0.0031	0.0031	0.9110	634
Croatia	0.0002	0.0000	0.0009	0.6284	1,283
Cyprus	0.0141	0.0000	0.0190	0.7473	434
Czech Republic	0.0151	0.0011	0.0008	0.9886	2,247
Denmark	0.0280	0.0021	0.0107	0.0290	1,896
Estonia	0.0157	0.0015	0.0030	0.3190	3,428
Finland	0.0272	0.0000	0.0000	0.0511	1,354
France	0.0679	0.0027	0.0132	0.8524	1,864
Germany	0.0163	0.0004	0.0073	0.9252	2,491
Greece	0.0068	0.0000	0.0000	0.8140	2,703
Hungary	0.0013	0.0000	0.0052	0.9172	698
Israel	0.0383	0.0003	0.0062	0.9287	1,044
Italy	0.0057	0.0006	0.0033	0.9775	2,325
Latvia	0.0194	0.0000	0.0060	0.2252	818
Lithuania	0.0109	0.0007	0.0013	0.9644	994
Luxembourg	0.0306	0.0006	0.0069	0.9671	776
Malta	0.0100	0.0047	0.0000	0.8836	416
Netherlands	0.0341	0.0000	0.0117	0.0436	736
Poland	0.0058	0.0018	0.0013	0.9529	2,567
Portugal	0.0065	0.0002	0.0124	0.9760	902
Romania	0.0054	0.0064	0.0085	0.9261	1,089
Slovakia	0.0000	0.0008	0.0031	0.9364	800
Slovenia	0.0183	0.0000	0.0063	0.9348	2,523
Spain	0.0610	0.0045	0.0095	0.9619	1,533
Sweden	0.1237	0.0023	0.0287	0.0189	1,283
Switzerland	0.0338	0.0014	0.0070	0.2550	1,630
All countries	0.0301	0.0017	0.0077	0.8162	41,758

Partial identification

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

- Starting premise:
  - the treatment is endogenous and there is no exogenous variation that can help identify the ATE
  - SUTVA is violated

- 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:

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

Simulated, fully specified models

- A simulated model allows one to evaluate the counterfactual $E[y(d) \mid x, w \neq d]$, typically in non-closed form as a policy function

- Treatment effect is calculated as the difference in the, typically, average outcomes when policy functions are evaluated at two different (sets of) values of the choice/parameter of interest

- Models make many assumptions (about optimizing behaviour, parameter values, functional forms, extent of heterogeneity, work, transportation and leisure networks, exogeneity of errors, global maxima); not clear how each one affects the results

Regression models

- One can also evaluate the counterfactual $E[y(d) | x, w \neq d]$ using a regression model, by evaluating the estimating equation giving everybody two different values of the treatment, and keeping everything else constant

- In a linear model, the treatment effect is given by the relevant regression coefficient

- Regression models can, in principle, evaluate structural relationships, even as best linear approximations

- Regression requires exogenous variation, choices on which to include, assumptions about unobservables, functional forms, etc.

Partial identification

- PI, instead of evaluating $E[y(d)|x, w \neq d]$ using various strong assumptions, put bounds around it

- 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, rests on few and credible assumptions, and it is absolutely transparent how each assumption affects results.

Partial Identification as a non-RCT (cont.)

- Bounding the counterfactual term implies bounding also $E[y(d)|x]$,

$$LB^M(d|x) \leq E[y(d)|x] \leq UB^M(d|x)$$

- Bounding $E[Y(d)]$ implies bounding also the ATE:

$$LB^M(d_2|x) - UB^M(d_1|x) \\ \leq E[y(d_2)|x] - E[Y(d_1|x)] \leq \\ UB^M(d_2|x) - LB^M(d_1|x)$$

Partial Identification (cont.)

- For example,

$$0.2 \leq E[Y(0)|x] \leq 0.3$$
$$0.1 \leq E[Y(1)|x] \leq 0.15$$

$$0.1 - 0.3 = -0.2$$
$$\leq E[Y(1)|x] - E[Y(0)|x)] \leq$$
$$0.15 - 0.2 = -0.05$$

- 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 (i.e., 0 and 1), which typically produce uninformative identification regions

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

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

Assumption 1: Monotone Treatment Response (Manski, Econometrica, 1997)

- More frequent mask wearing does not increase the probability of infection, ***on average***, ***keeping everything else constant***
  - weak monotonicity, not sufficient to exclude 0 from the ATE identification region

$$E[y(1)|x, w=d] \leq E[y(0)|x, w=d], \text{ for every } d$$

- Plausible assumption – difficult to think why mask wearing would increase infection prevalence. Has strong effect on results

Assumption 1: Monotone Treatment Response (Manski, Econometrica, 1997)

- Threat to the validity of MTR risk compensation
  - Wearing masks makes it more likely to take risks
  - No empirical evidence for this behaviour on average

- Example of MTR operation: the counterfactual mean infection prevalence of those who wear masks seldom, had they worn them often. Under MTR, an upper bound for this counterfactual term is the actual infection rate of those who wear masks seldom

- This MTR upper bound (=0.042) is smaller than the overall maximum (=1), and thus identification regions become narrower and thus more informative

Monotone treatment response (cont.)

- $UB^{MTR}(1) = LB^{MTR}(0) = E(y)$

- Hence, $UB_{ATE}^{MTR}(1-0) = 0$

- After combining with MIV assumption,

$$UB_{ATE}^{MTR+MIV}(1-0) < 0$$

- MTR can be tested in conjunction with Monotone Treatment Selection – the test does not reject the joint validity

Assumption 2: Monotone Instrumental Variable ((Manski and Pepper, Econometrica, 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

- They serve to partition the sample space conditional on their values and search for maximum lower bounds and minimum upper bounds
- Need to narrow the identification region
  - UB 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

Monotone instrumental variable (MIV)

- 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

- Our MIVs are: i) not keeping distance from others when going out; ii) meeting often with groups of 5 or more people outside the household. Both should be weakly positively associated with the prevalence of infection

How do MIVs work? (cont.)

- 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

Assumption 3: Bounded variation (Manski and Pepper, 2018)

- The BV assumption puts a bound on how much the counterfactual term can differ from the observed term
- Example of BV operation: the counterfactual mean infection prevalence of those who wear masks seldom, had they worn them often. Under BV, the lower bound for this counterfactual term can be smaller than the actual infection prevalence of those who wear masks seldom by at most a multiple k of the standard deviation of the outcome (=0.17).
- This lower bound is smaller than the overall minimum and thus identification regions become narrower and thus more informative
- BV does not affect the ATE upper bound, i.e., by at least how much the probability of infection drops due to mask wearing) but only the ATE lower bound (i.e., by at most how much the probability of infection drops)
- We show results for several values of k (60%, 50%, 40% of a SD) and let readers decide which they think is most plausible

Non-individualistic response (Manski, 2013)

- $E[y(d^J)]$, i.e., average potential outcomes depend on the treatments received by others in a set J (here assumed to be the population)
- Problem: $E[y(d^J)|w = d] \neq E(y|w = d) = E[y(d)|w = d]$
- Reinforcing interactions (RI) assumption: for any two treatment vectors d_1^J and d_2^J such that $d_2^k \geq d_1^k \ \ \forall\, k \in J$,

$$E[y(d_2^J)|w = d] \leq E[y(d_1^J)|w = d]$$

- MTR and RI operate in the same direction: credible in this context

Non-individualistic responses (cont.)

- Just as with MTR, $UB^{RI}(1) = LB^{RI}(0) = E(y)$

- Hence, $UB^{RI}_{ATE}(1-0) = 0$

- After combining with MIV, $UB^{RI+MIV}_{ATE}(1-0) < 0$

- MIV operates with RI exactly as with MTR

PI results for consumption – individualistic response

Assumptions	Estimates: Lower bound	Estimates: Upper bound	Lower Bound Low 95% CI	Lower Bound Upper 95% CI	Upper Bound Low 95% CI	Upper Bound Upper 95% CI
Panel A: Individualistic response						
Exogenous treatment selection	-0.0136		-0.0277		-0.0020	
No assumptions	-0.8000	0.2000	-0.8095	-0.7902	0.1905	0.2098
MTR	-0.8000	0.0000	-0.8095	-0.7902	0.0000	0.0000
MIV2	-0.6135	0.2000	-0.6574	-0.5642	0.1905	0.2098
MTR + MIV2	-0.6161	-0.0076	-0.6607	-0.5682	-0.0173	-0.0008
MTR + MIV2 + BV60	-0.0825	-0.0076	-0.0940	-0.0704	-0.0173	-0.0008
MTR + MIV2 + BV50	-0.0685	-0.0076	-0.0786	-0.0574	-0.0173	-0.0008
MTR + MIV2 + BV40	-0.0548	-0.0076	-0.0639	-0.0445	-0.0173	-0.0008
Number of observations			40,954			

PI results for consumption – non-individualistic response

Assumptions	Estimates		Lower Bound Low 95% CI	Lower Bound Upper 95% CI	Upper Bound Low 95% CI	Upper Bound Upper 95% CI
	Lower bound	Upper bound				
No assumptions	-1	1	-1	-1	1	1
RI	-1	0	-1	-1	0	0
MIV2	-1	1	-1	-1	1	1
RI + MIV2	-1	-0.0076	-1	-1	-0.0173	-0.0008
RI + MIV2 + BV60	-0.0825	-0.0076	-0.0940	-0.0704	-0.0173	-0.0008
RI + MIV2 + BV50	-0.0685	-0.0076	-0.0786	-0.0574	-0.0173	-0.0008
RI + MIV2 + BV40	-0.0548	-0.0076	-0.0639	-0.0445	-0.0173	-0.0008

Conclusions

- We find a negative causal effect of masks on probability of symptomatic infection of at least 0.76% and as much as 5.5% - non-trivial effect when infection prevalence is 3%, and also when symptomatic infection can lead to serious health problems/death

- Results obtained using 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 while taking into account treatment spillovers

- Limitations: increased uncertainty, temporal specificity (virus strains have changed since June 2020)

THANK YOU!

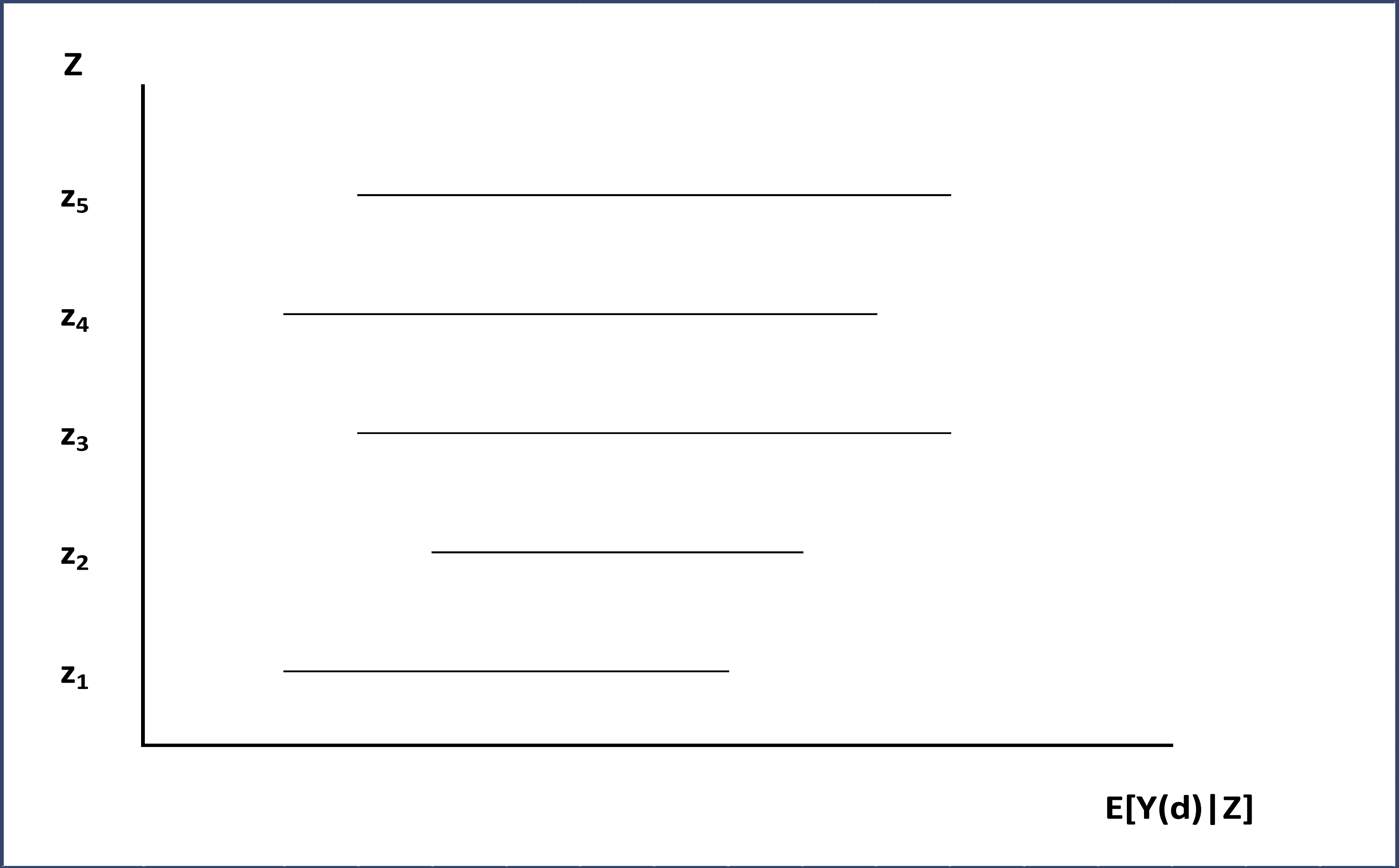
How do MIVs work? - XIVs

- 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.)

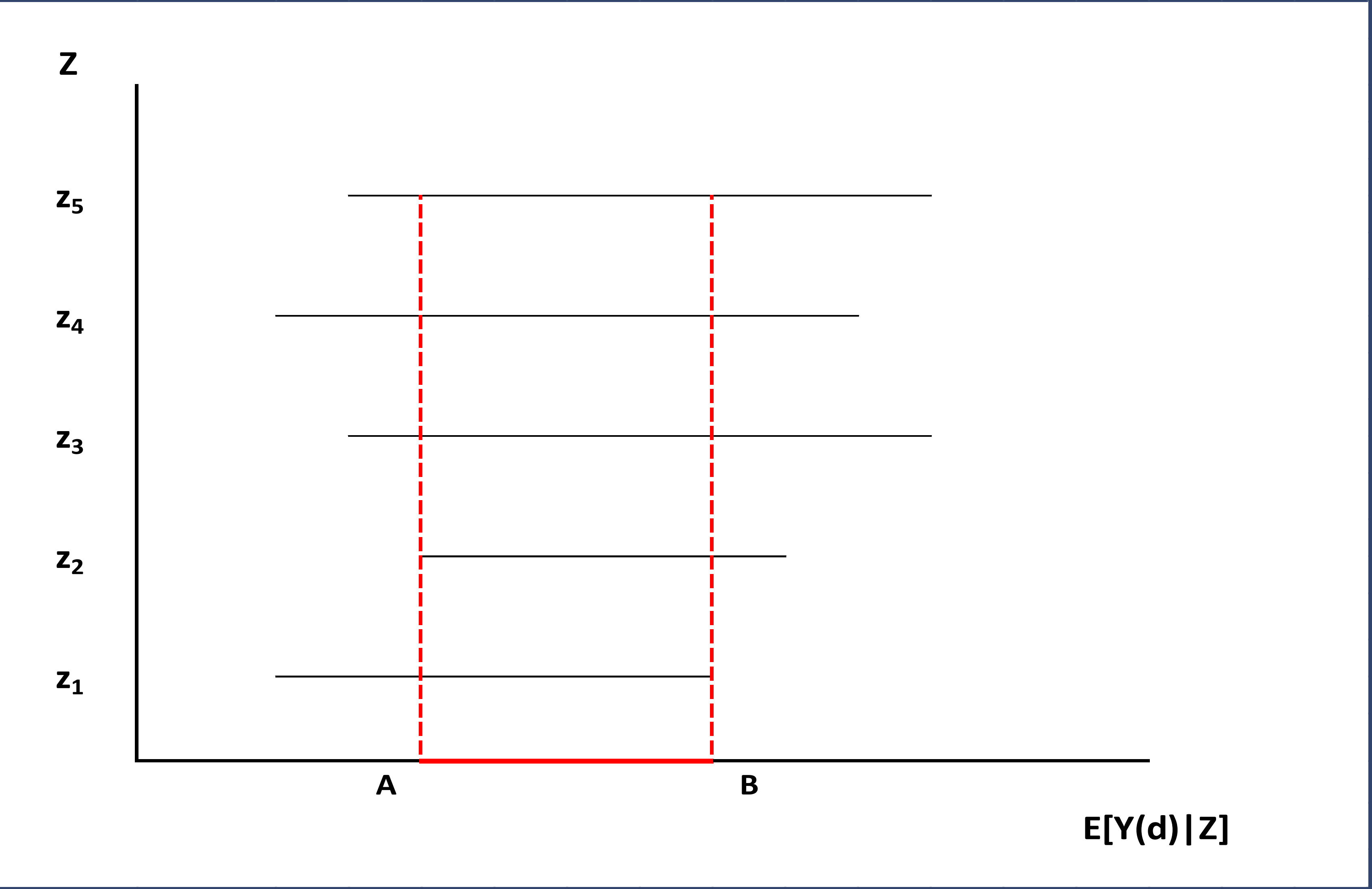
- Since the instrument is exogenous, all conditional identification regions should give the same answer

- Same answer → the intersection of all 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? (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