Lecture 9: Heterogeneity

Tomas Havranek

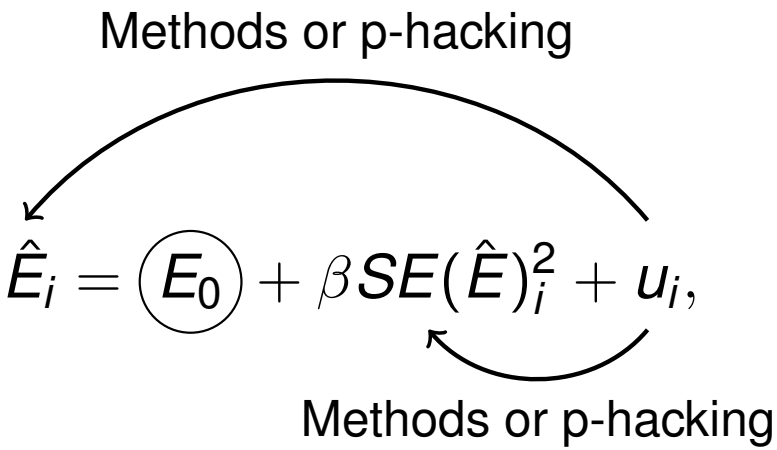
Charles University, CEPR, METRICS

Research Synthesis in Economics and Finance,
University of Canterbury

March 17, 2025

Identification of publication bias

PEESE:



corr$(SE, u) \neq 0 \Rightarrow \hat{\beta}$ and $\hat{E}_0$ biased.

Natural solution: N instrumenting $SE(\hat{E})_i^2 \rightarrow$ MAIVE.

Or add covariates!

Classical heterogeneity measure

Common measure: estimate the relative heterogeneity of underlying effects *(Higgins and Thompson, 2002)*

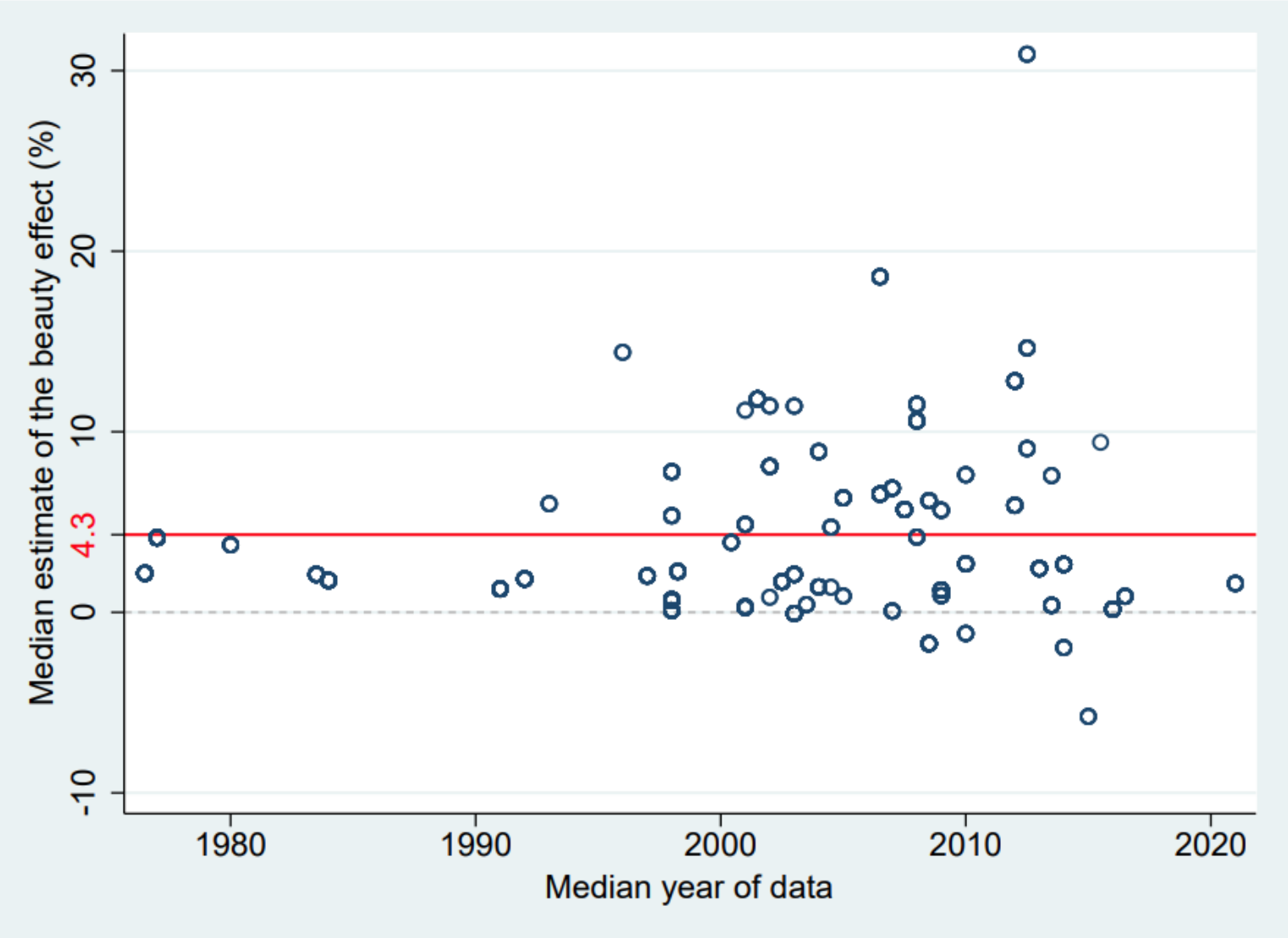
$$I^2 \equiv \frac{\text{underlying variance of effects}}{\text{observed variance of estimates}} = \frac{\tau^2}{\tau^2 + \bar{v}}$$

- $I^2 = 0.75 \sim 1.0$ ... *"considerable heterogeneity" in medicine (Cochrane Review Guideline)*
- $I^2 \simeq 0.9$ ... common in economics meta-analyses (e.g., *Vivalt, 2021*)

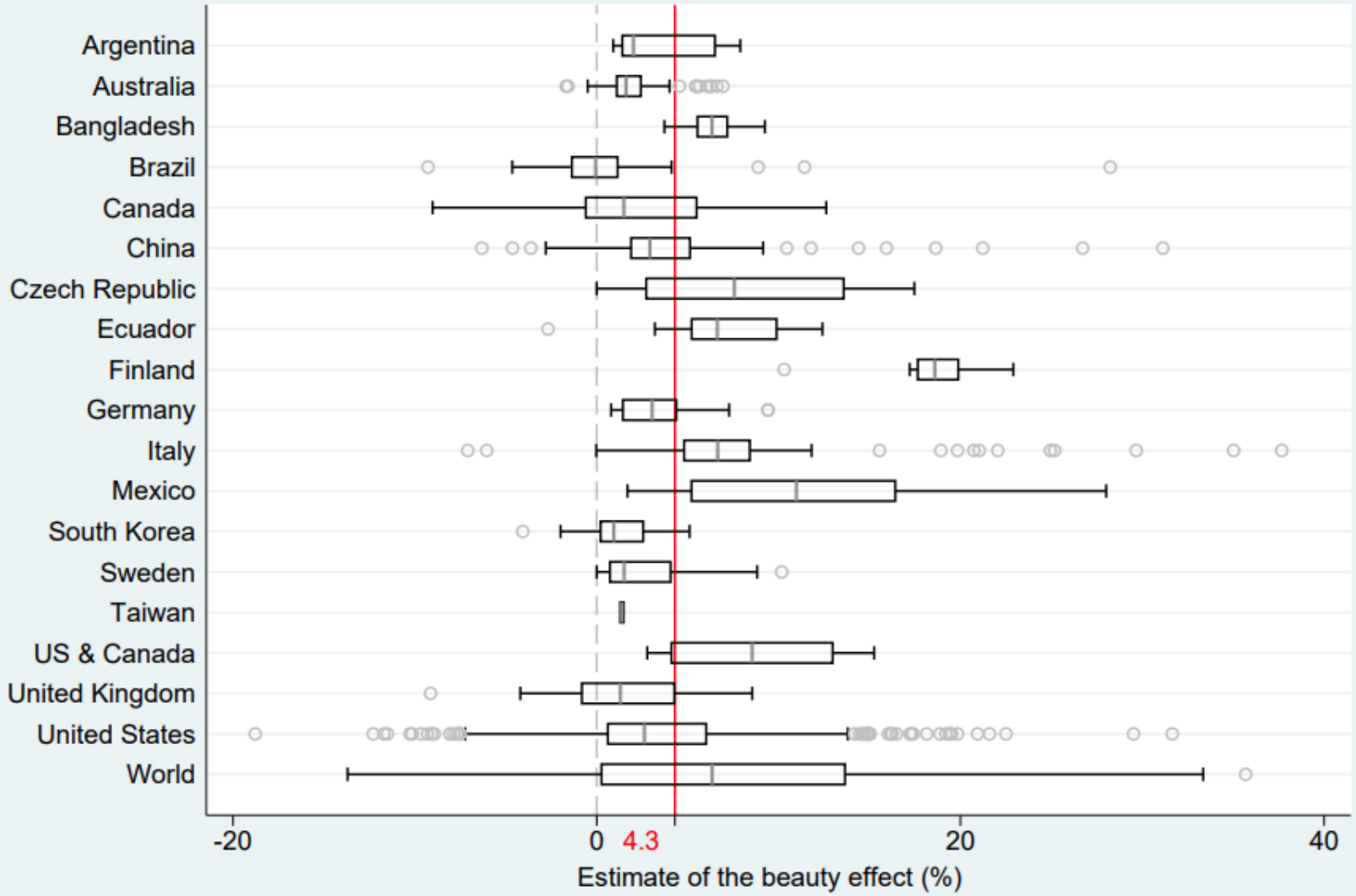
Unresolved concern:

- no universally agreed criteria for "too much" heterogeneity
- undesirable pattern: with more precise studies, higher I^2

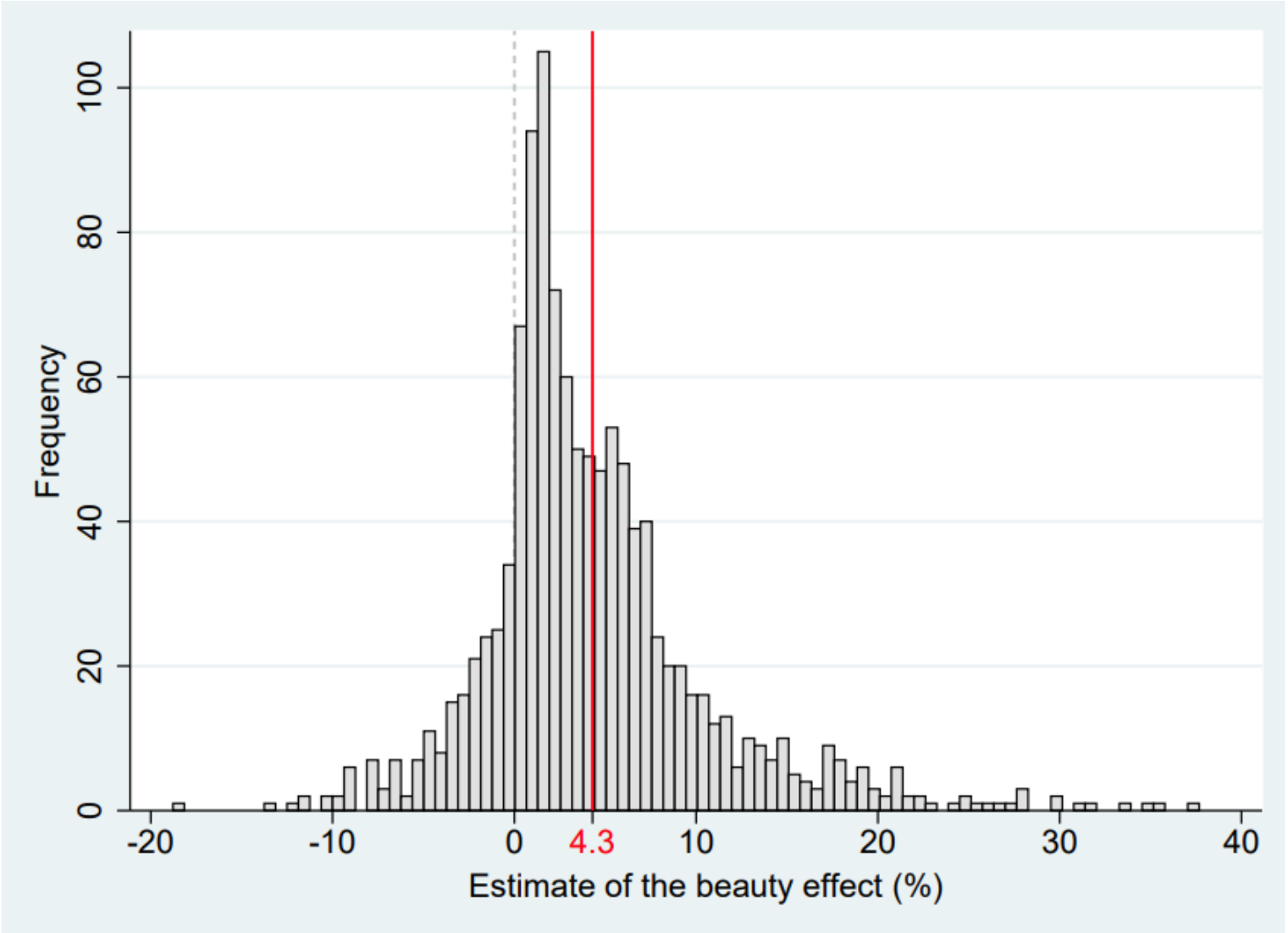
Guiding example: beauty and success



Heterogeneity both within and across



Wide dispersion



Measurement of beauty

	Est.	Stud.	Unweighted Mean	Unweighted 95% conf. int.	
Measurement of beauty					
Interviewer-rated beauty	526	24	4.33	3.80	4.86
Photo-rated beauty	481	37	4.28	3.71	4.85
Software-rated beauty	104	8	5.49	4.15	6.83
Self-rated beauty	56	6	2.04	0.79	3.29
Dummy beauty	460	23	3.63	3.05	4.22
Categorical beauty	699	52	4.70	4.24	5.16
Beauty premium	954	67	4.48	4.07	4.89
Beauty penalty	205	16	3.33	2.56	4.10

Measurement of success

			Unweighted		
	Est.	Stud.	Mean	95% conf. int.	
Measurement of success					
Earnings	688	43	4.33	3.92	4.75
Study outcomes	189	9	3.19	1.98	4.41
Teaching & research outcomes	139	8	4.95	4.05	5.85
Athletic success	33	2	6.28	3.03	9.54
Electoral success	37	4	7.00	4.26	9.75
Other outcomes	73	6	2.98	2.24	3.73

Data characteristics

	Est.	Stud.	Unweighted Mean	Unweighted 95% conf. int.	
Data characteristics					
Male subjects	375	44	3.66	3.10	4.22
Female subjects	447	48	4.10	3.45	4.75
Mixed gender	337	36	5.21	4.57	5.85
High-skilled workers	335	27	4.30	3.78	4.81
Prostitutes	55	4	8.55	7.27	9.82
Other dressy occupations	138	15	4.92	3.89	5.96
Non-dressy occupations	966	51	3.94	3.55	4.34
Western culture	874	47	3.85	3.46	4.23
Other cultures	285	21	5.60	4.72	6.48
Panel data	998	55	3.86	3.47	4.25
Cross-section	161	13	6.87	5.90	7.83

Estimation technique

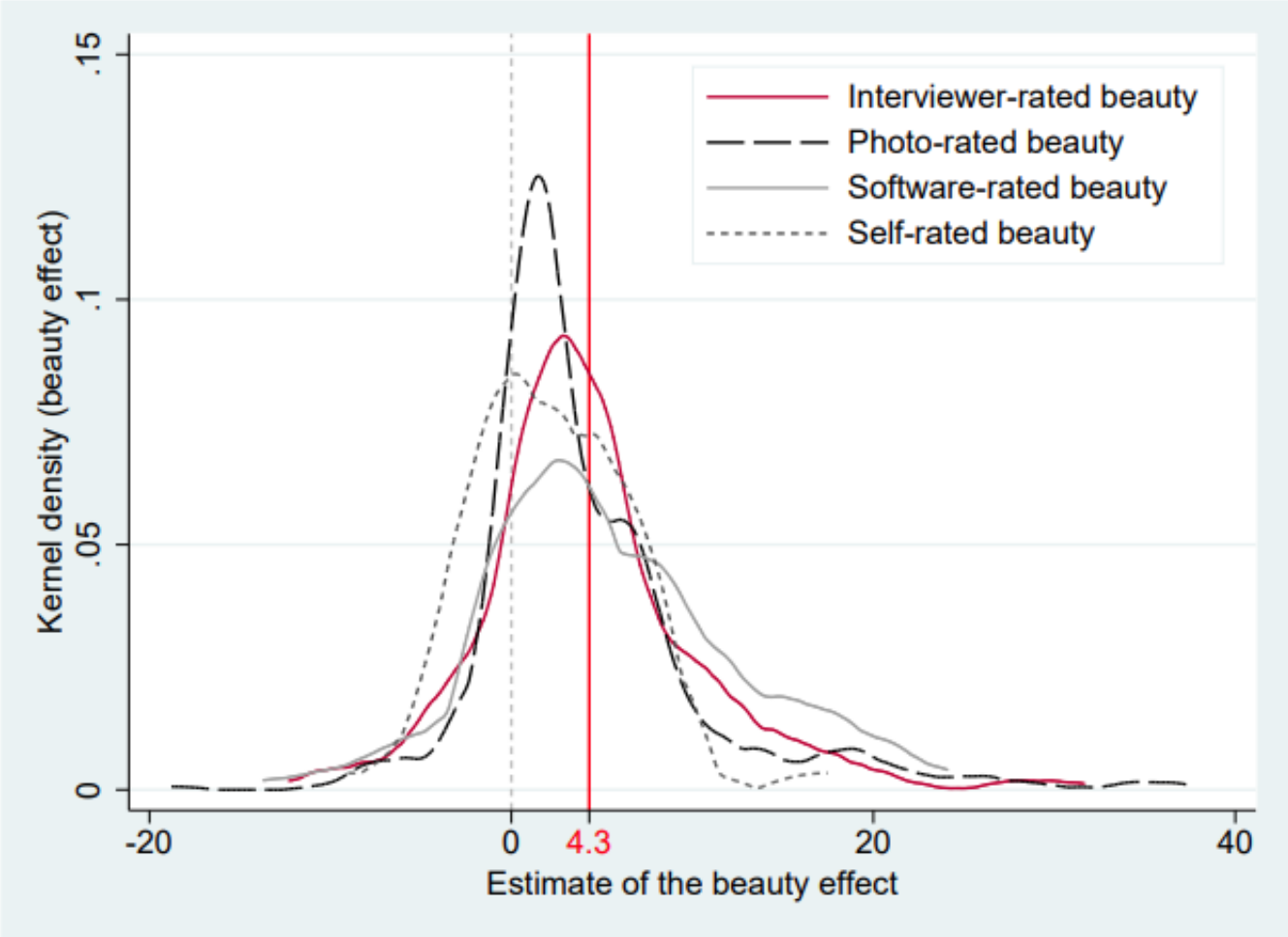
	Est.	Stud.	Unweighted Mean	Unweighted 95% conf. int.	
Estimation technique					
Ordinary least squares	903	60	4.17	3.77	4.56
Instrumental variables	83	10	4.86	3.07	6.64
Difference-in-differences	12	2	1.24	0.63	1.84
Other method	161	13	4.84	3.78	5.90
Cognitive skill control	445	26	2.77	2.17	3.36
No cognitive skill control	714	53	5.22	4.78	5.66

Publication characteristics

	Est.	Stud.	Unweighted Mean	Unweighted 95% conf. int.	
Publication characteristics					
Published study	1,027	58	4.19	3.80	4.57
Unpublished study	132	9	4.99	3.96	6.03
High-quality peer review	151	7	5.40	4.56	6.25

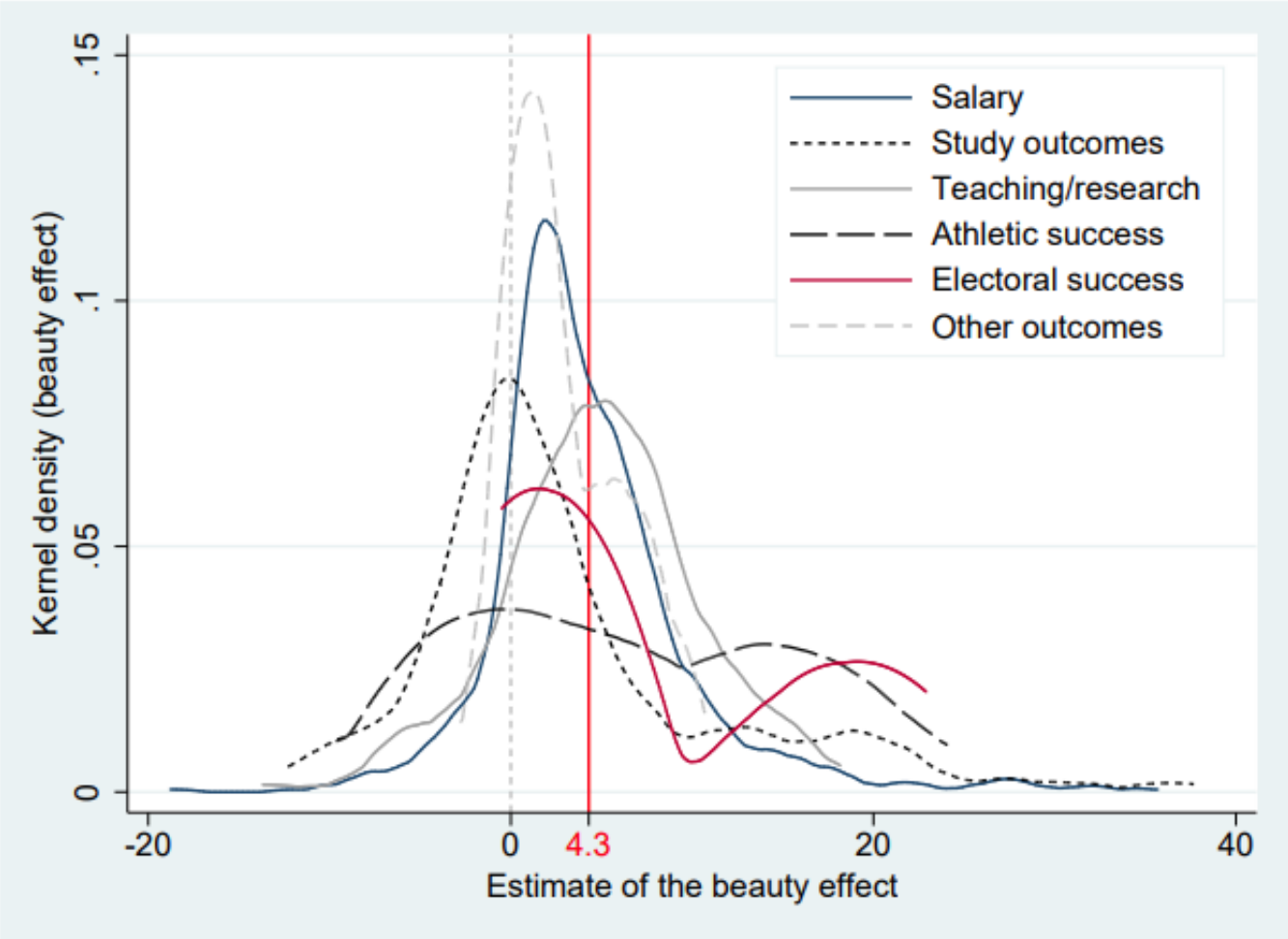
Beauty measurement doesn't matter

(a) Beauty measurement



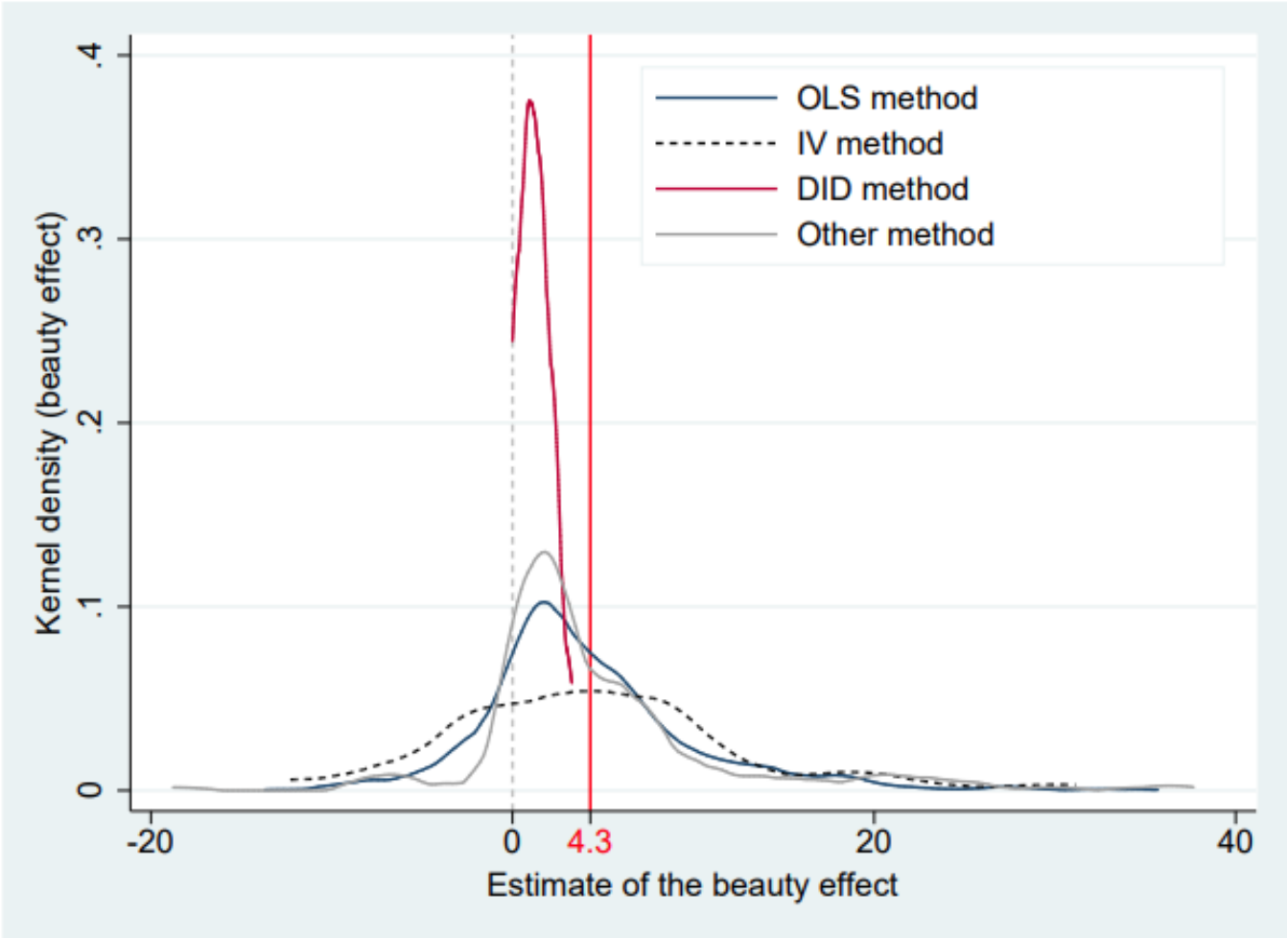
Success measurement doesn't matter

(b) Success measurement



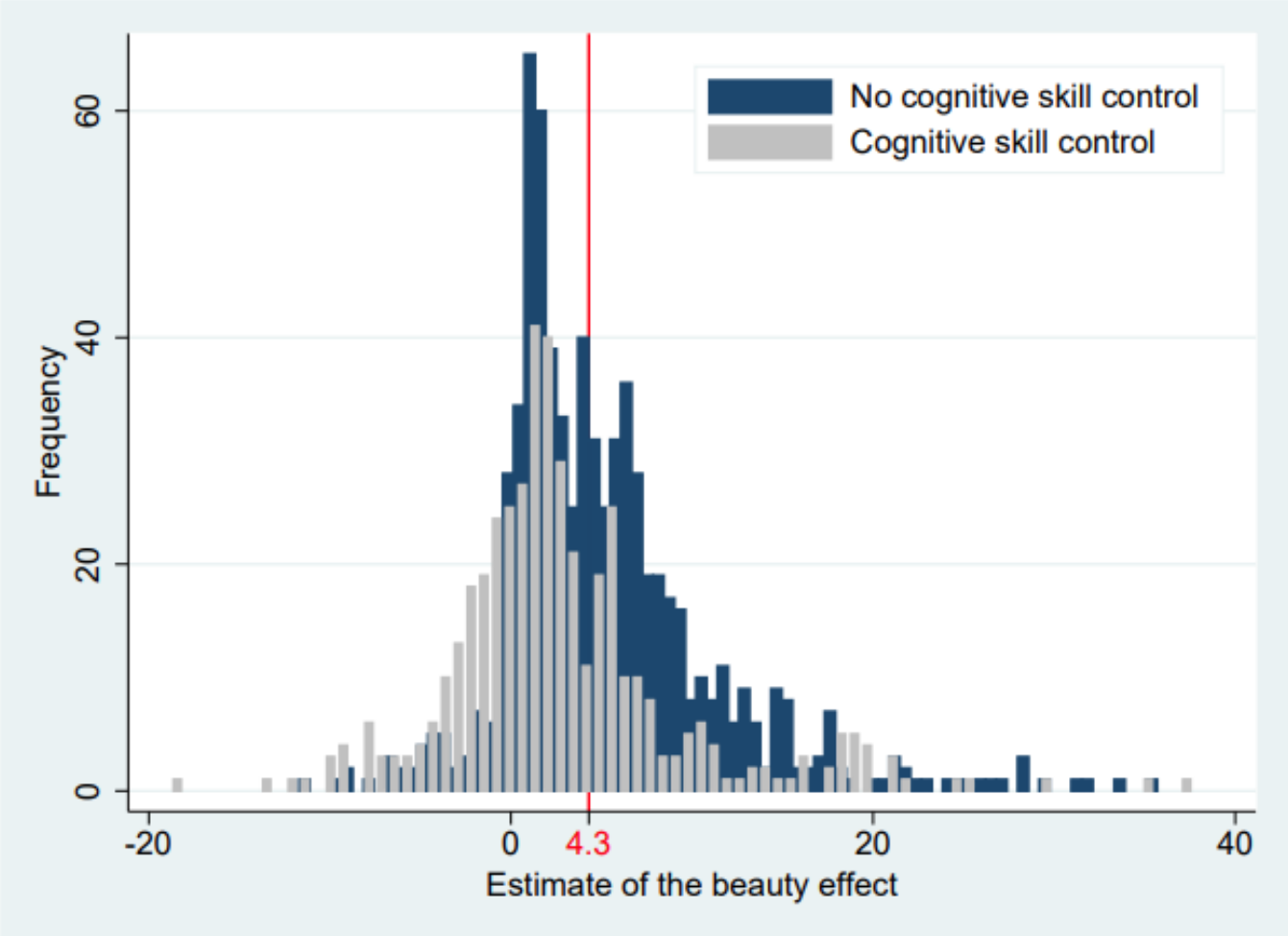
Method doesn't matter

(c) Method choice



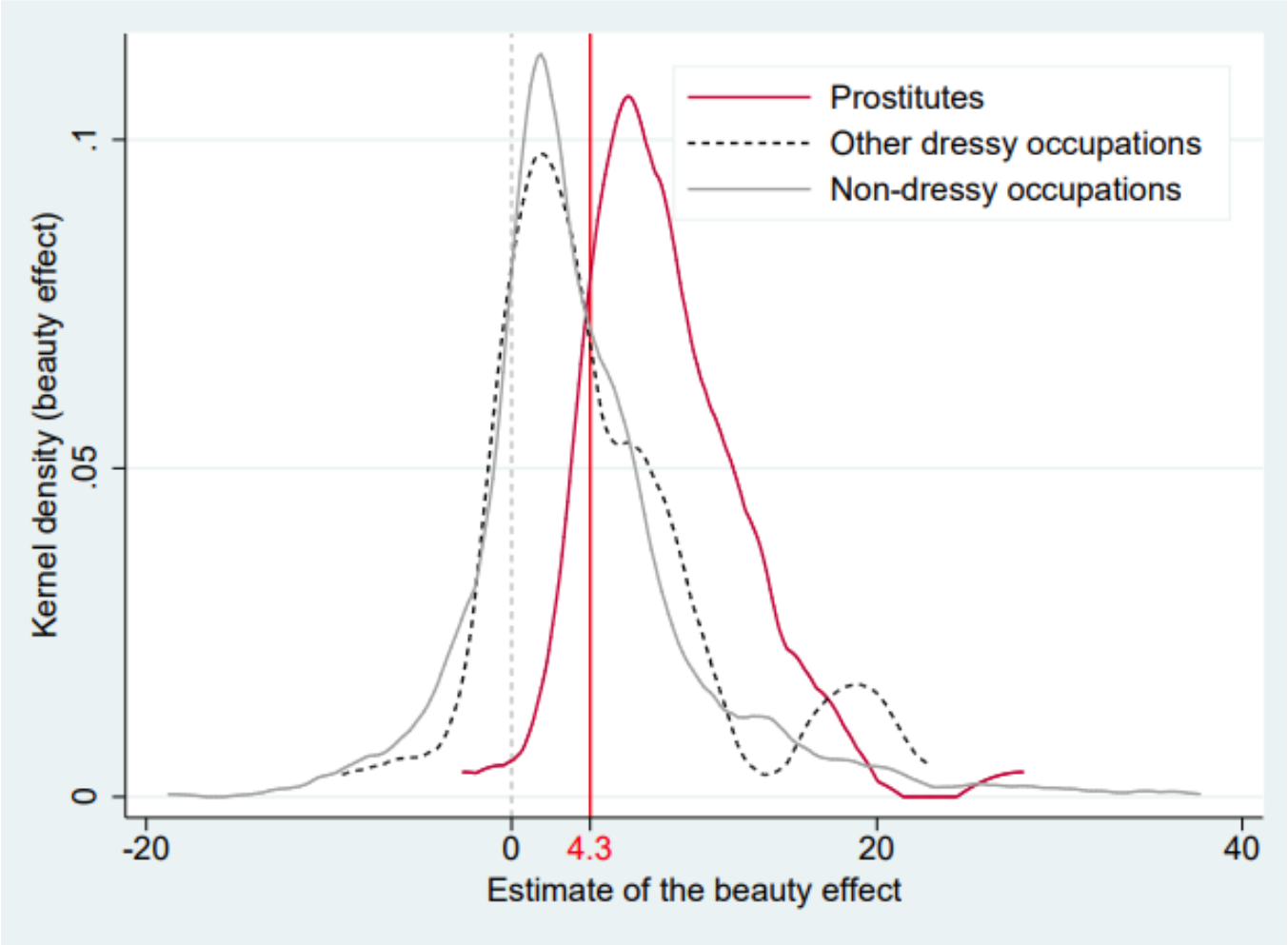
Ability control matters

(g) Ability control

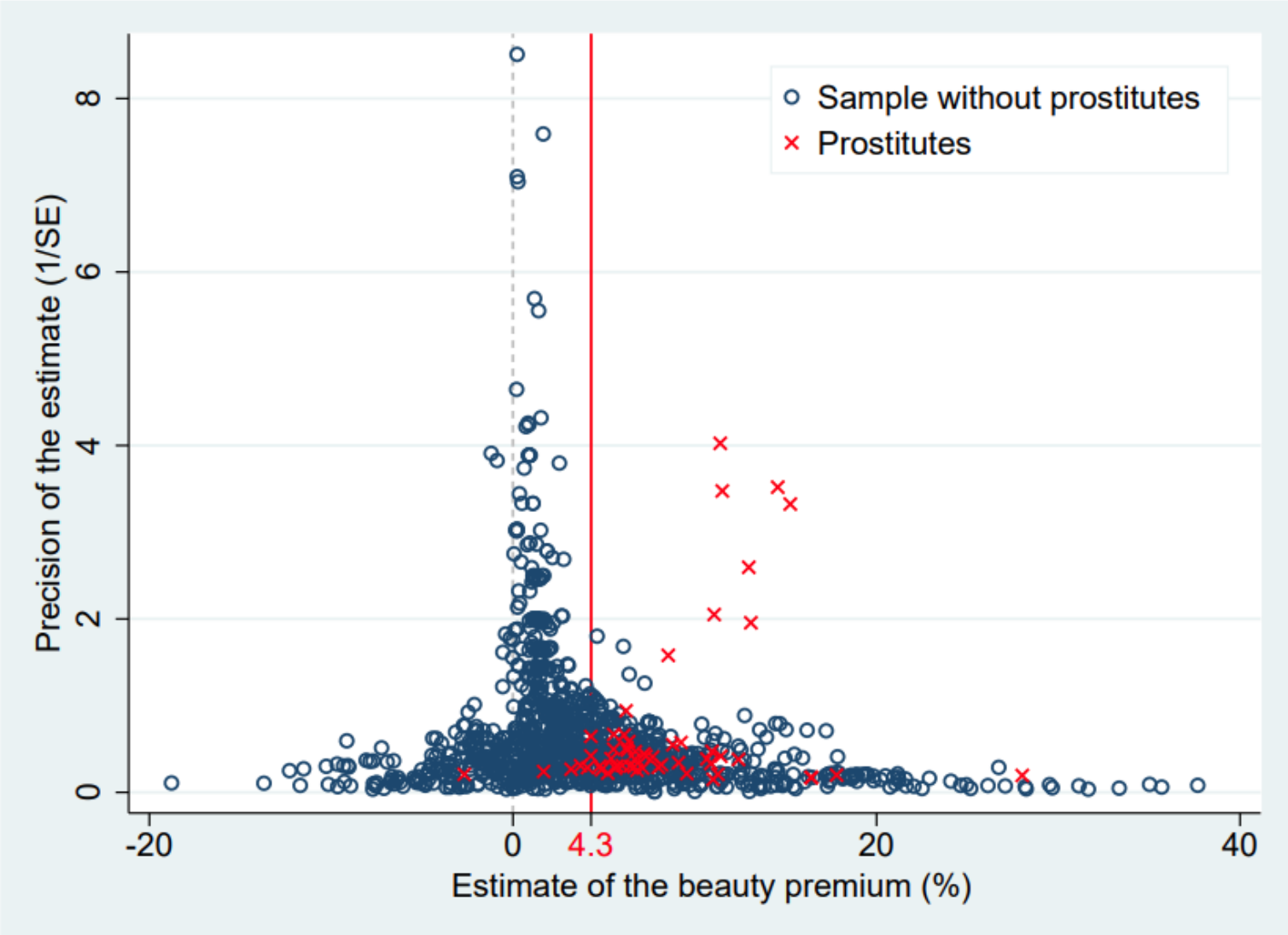


Occupation matters

(d) Occupation



Sex workers stand apart



Option 1: subsamples

Part 1. Subsample of prostitutes

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias (*standard error*)	0.148 (0.776) [-2.956, 2.671]	1.467 (0.782)	-0.756 (1.032)	-1.567 (0.967) {-7.405, -0.514}	-0.468** (0.210) [-3.545, 0.383]
Effect beyond bias (*constant*)	8.136*** (2.767) [1.525, 15.17]	4.488 (2.164)	11.260* (2.686)	12.880*** (1.392) {1.110, 17.173 }	11.760*** (0.414) [5.245, 13.97]
First-stage robust F-stat				15.8	
Observations	55	55	55	55	55
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	-2.154*** (0.745) [-4.057, -0.568]			-2.126 (1.878) [-4.326, -0.755]	P = 1.215 (0.091)
Effect beyond bias	13.290*** (0.298) [-22.03, 16.87]	12.168*** (0.372)	11.205*** (0.905)	12.214*** (0.349) [-22.49, 17.96]	8.490*** (1.691)
Observations	55	55	55	55	55

Option 1: subsamples

Part 2. Sample without prostitutes

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias	0.391***	0.204*	0.800***	0.875*	0.718***
(*standard error*)	(0.117)	(0.119)	(0.161)	(0.450)	(0.271)
	[0.120, 0.642]			{0.118, 2.078}	[0.061, 1.331]
Effect beyond bias	2.581***	3.289***	1.603*	0.741	2.617**
(*constant*)	(0.447)	(0.453)	(0.875)	(1.728)	(1.061)
	[1.660, 3.501]			{0.048, 3.696}	[0.333, 5.113]
First-stage robust F-stat				16.2	
Observations	1,104	1,104	1,104	1,104	1,104
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	1.610***			1.610***	P = 0.307
	(0.202)			(0.202)	(0.039)
	[1.184, 2.051]			[1.184, 2.051]	
Effect beyond bias	0.152	0.229***	0.008	0.152	0.669***
	(0.118)	(0.07)	(0.798)	(0.118)	(0.231)
	[-0.050, 0.963]			[-0.050, 0.963]	
Observations	1,104	1,104	1,104	1,104	1,104

Option 1: subsamples

Part 1. Subsample of penalties

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias	0.165**	-0.354	0.702***	0.578	0.354***
(*standard error*)	(0.0678)	(0.235)	(0.168)	(1.285)	(0.131)
	[0.021, 0.406]			{-1.941, 3.096}	[0.028, 1.395]
Effect beyond bias	2.647***	4.801***	0.447	0.933	2.655**
(*constant*)	(0.733)	(0.977)	(0.877)	(5.565)	(1.153)
	[1.225, 4.378]			{-3.133, 4.998}	[0.038, 5.782]
First-stage F-stat				0.8	
Observations	205	205	205	205	205
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	1.097***			1.097***	P = 0.369
	(0.164)			(0.12)	(0.019)
	[0.709, 1.489]			[0.709, 1.489]	
Effect beyond bias	0.139***	0.244***	0.245	0.139	0.236***
	(0.0531)	(0.021)	(0.323)	(0.097)	(0.067)
	[-0.405, 1.425]			[-0.405, 1.425]	
Observations	205	205	205	205	205

Option 1: subsamples

Part 2. Sample without penalties

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias	0.437***	0.282*	0.745***	0.772*	0.666**
(*standard error*)	(0.139)	(0.160)	(0.172)	(0.440)	(0.281)
	[0.125, 0.745]			{0.032, 1.949}	[-0.044, 1.295]
Effect beyond bias	2.881***	3.448***	2.228**	1.652	3.470***
(*constant*)	(0.555)	(0.585)	(0.914)	(1.580)	(1.202)
	[1.745, 4.058]			{0.068, 4.171}	[0.845, 6.305]
First-stage robust F-stat				17.3	
Observations	954	954	954	954	954
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	1.881***			1.881***	P = 0.300
	(0.246)			(0.246)	(0.037)
	[1.351, 2.378]			[1.351, 2.378]	
Effect beyond bias	0.346	0.38*	0.013	0.346	0.200
	(0.294)	(0.229)	(1.323)	(0.294)	(0.828)
	[-0.051, 2.285]			[-0.051, 2.285]	
Observations	954	954	954	954	954

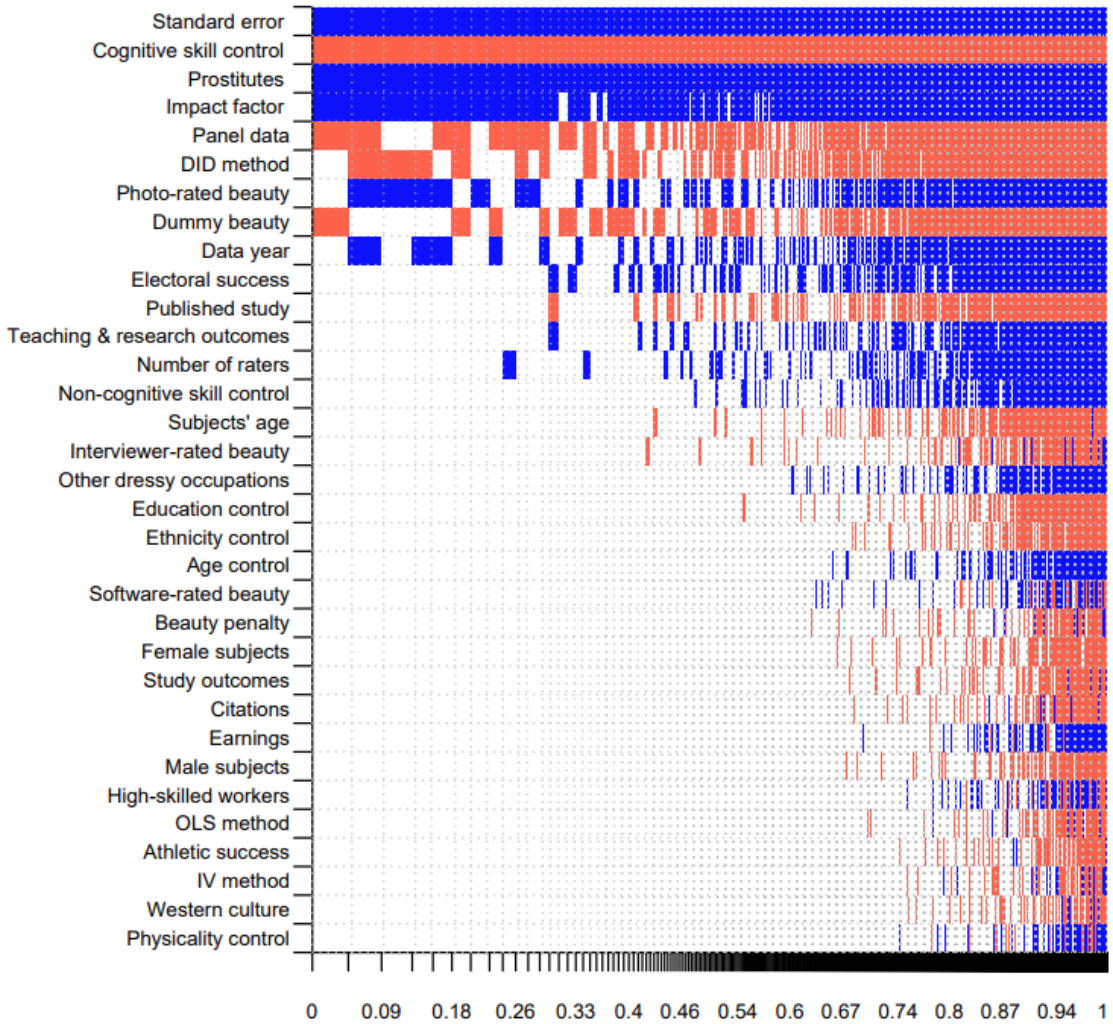
Option 2: Bayesian model averaging

- We want to regress estimates of the beauty effect on variables reflecting heterogeneity
- But too many variables; model uncertainty
- Solution: run regressions with different combinations of the variables
- Give more weight to those that fit the data well and are parsimonious

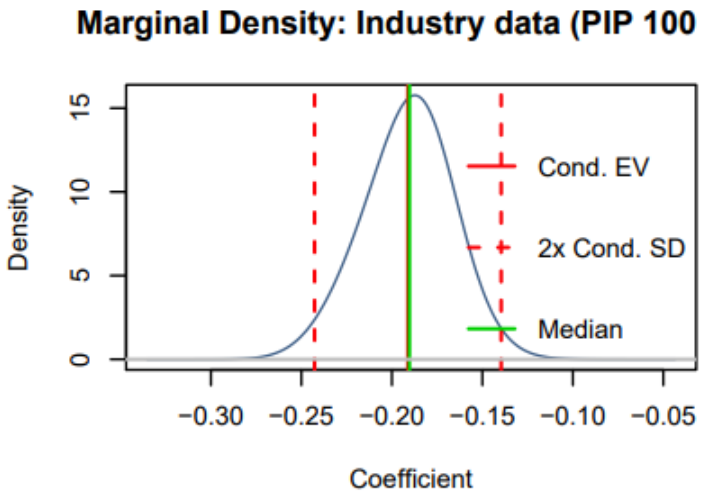
Option 2: Bayesian model averaging

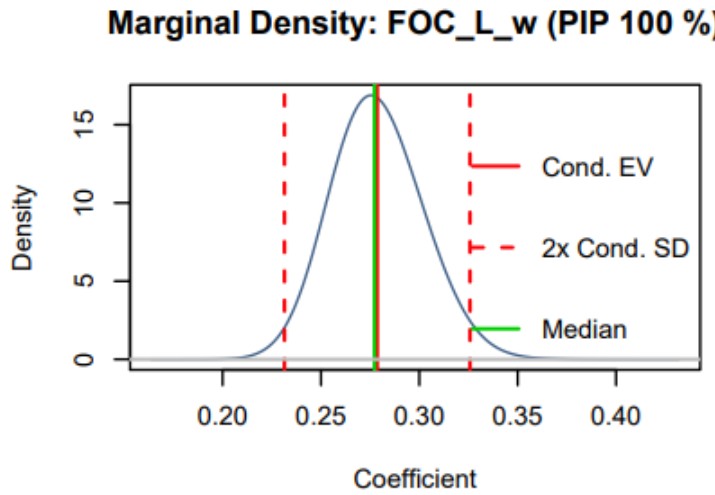
- Model averaging can be frequentist but computationally difficult
- Intuition: run different combinations and weight by adjusted R^2 or information criteria
- In the Bayesian setting we can use the MCMC algorithm
- G-prior: all coefficients are zero. Weight? UIP = the prior has the same weight as one observation
- Model prior: uniform = all models have the same weight
- Dilution prior: models with collinearity get less weight

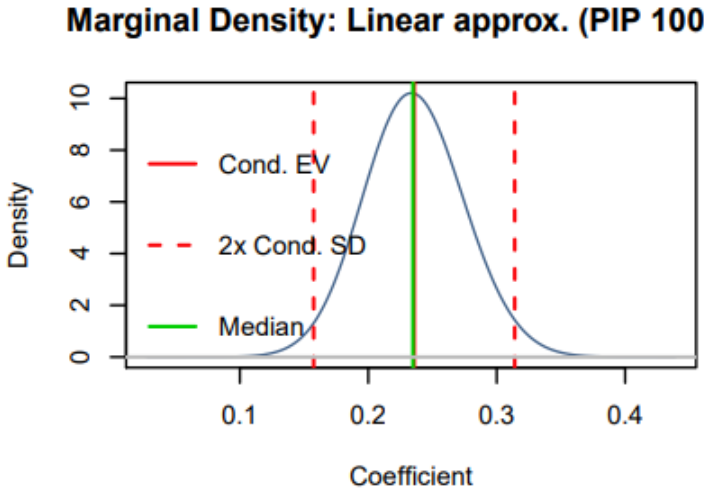
Model inclusion in BMA

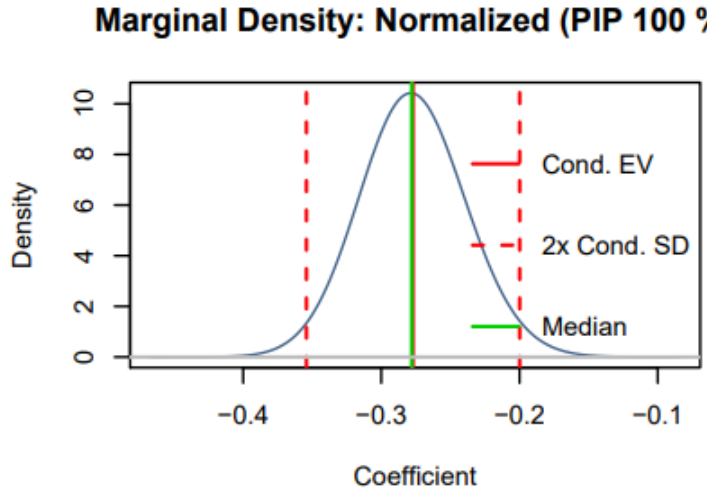


Posterior densities









Regression results

Table 5: Why reported beauty premiums vary

Response variable: Beauty premium	Bayesian model averaging (baseline model)			Stepwise regression (frequentist check)		
	P. mean	P. SD	PIP	Mean	SE	p-value
Constant	2.759	NA	1.000	3.582	0.562	0.000
Standard error	0.435	0.042	1.000	0.426	0.112	0.000
Measurement of beauty						
Interviewer-rated beauty	-0.033	0.204	0.036			
Photo-rated beauty	0.591	0.745	0.424			
Software-rated beauty	0.009	0.149	0.014			
Dummy beauty	-0.485	0.664	0.387			
Beauty penalty	-0.007	0.089	0.012			
Number of raters	0.051	0.140	0.134			
Measurement of success						
Earnings	0.006	0.098	0.010			
Study outcomes	-0.006	0.114	0.011			
Teaching & research	0.238	0.645	0.141			
Athletic success	-0.004	0.104	0.007			
Electoral success	0.670	1.361	0.224			
Data characteristics						
Male subjects	-0.005	0.064	0.010			
Female subjects	-0.006	0.069	0.012			
Subjects' age	-0.075	0.332	0.061			
High-skilled workers	0.002	0.064	0.009			
Prostitutes	4.793	1.058	0.999	4.386	1.199	0.001
Other dressy occupations	0.035	0.227	0.032			
Western culture	-0.001	0.039	0.006			
Panel data	-1.131	1.016	0.606			
Data year	0.337	0.504	0.346			

Best practice (implied estimate)

- From BMA results we know how estimates depend on study design
- Some study designs are better: identification, data, generally quality
- Select best practice based on the literature (quasi-experimental studies, etc.)
- Compute fitted values and confidence intervals (lincom command in Stata)

Best practice (class size effect)

	Effect	95% cred. int.	
STAR experiment	-1.50	-3.04	0.04
Regression discontinuity	-0.18	-1.36	1.00
Instrumental variable	-0.42	-1.59	0.75
Fixed effects	0.01	-1.17	1.19
OLS	0.01	-1.19	1.21
Kindergarten	0.18	-1.10	1.46
Primary school	-0.46	-1.65	0.72
Secondary school	0.18	-0.99	1.35
Disadvantaged students	-0.19	-1.42	1.03
USA	-0.21	-1.30	0.88
Scandinavian countries	-0.21	-1.47	1.05
Other countries	-0.21	-1.43	1.02
Math	-0.21	-1.38	0.97
Reading	-0.21	-1.39	0.96
Writing	-0.20	-1.50	1.11
Languages	-0.21	-1.40	0.99
Other subject	-0.21	-1.43	1.01
Class size = 15	-0.99	-2.18	0.21
Class size = 20	-0.45	-1.62	0.72
Class size = 25	-0.18	-1.35	1.00
Class size = 30	0.01	-1.18	1.19
Class size = 35	0.15	-1.05	1.34
Overall corrected mean	-0.21	-1.38	0.97