Simulating an empirical paper in economics
Why publication bias is so common

2014 MAER-Net Colloquium in Athens

Martin Paldam

Based on paper with URL:
http://www.martin.paldam.dk/Papers/Meta-method/Simulating-pub-bias.pdf

Four highlights, with a preview of answers

- 1. How to simulate economic research and hence simulate 'realistic' funnels?
- Use economic theory on choice problem of researchers: The fit-size diagram ➔ textbook choice

- 2. How different is polishing (fit) and censoring (size)?
- They are amazingly similar

- 3. How many simulations should you make?
- Go on till pattern in results is smooth! I did $70 \ x \ 10^6$ regs.

- 4. Is the PET or PEESE better?
- They are rather similar: The big step is from the mean to either of the two

The format of the research process of an empirical paper estimates parameter β

I. Intuition ➔ theory ➔ Qualitative prediction: **$\beta > 0$**

II. Theory ➔ Estimating model: The β-term + cp controls + other controls: **Concentrate on β-term**
cp controls for β heterogeneity: Deleted ➔ noise.
Other controls taken to be noise. Thus big noise

III. Search among model estimates: **Generate *J* estimates**

IV. Choose the main one to publish: ***SR* choice rules**

Textbook choice:
PPF, production possibility frontier, and
IC, Indifference curves ➔ ***J* & *SR***

- Production function for research results, DGP/EM produces the *J*-set, which is the choice set – its rim is:
- PPF, **production possibility frontier:** PPF = PPF(***J***)
- Researchers + journals have priors for ***fit*** and ***size*** of results
- IC, **indifference curves**, for size and fit
- ***SR***s, **selection rules:** for results published

Where does the ***fit*** and the ***size*** priors come from:

- Many different priors, may be OK, **but problem if:**
- **MP** **main prior**: Joint for too many researchers

- **MP1**: Prior for clear results, ***fit*** (t-ratio), afflicts us all
- **MP 2**: Standard economic theory, ***size*** (normally sign)
- **MP 3**: Political-moral beliefs, ***size***
- **MP 4**: Interest of big sponsors, ***size***
- PS: The paper assumes two main priors:
- (1) **a fit prior + a size prior** - not its origin
  PS: Use a fit-size diagram

- Prior 5: Prior results of researchers: Past MP ➔ future MP

Method: Simulations calibrated by meta-analysis.

It studies the β-literature: i.e., the N estimates that pertain to be of the same β

- **Meta-analysis:** 750 $\pm$ 250 meta-studies in economics
  The average meta study analyzes about 50 papers
- Hence, about 40,000 papers coded.
- So you can calibrate simulations to look reasonable

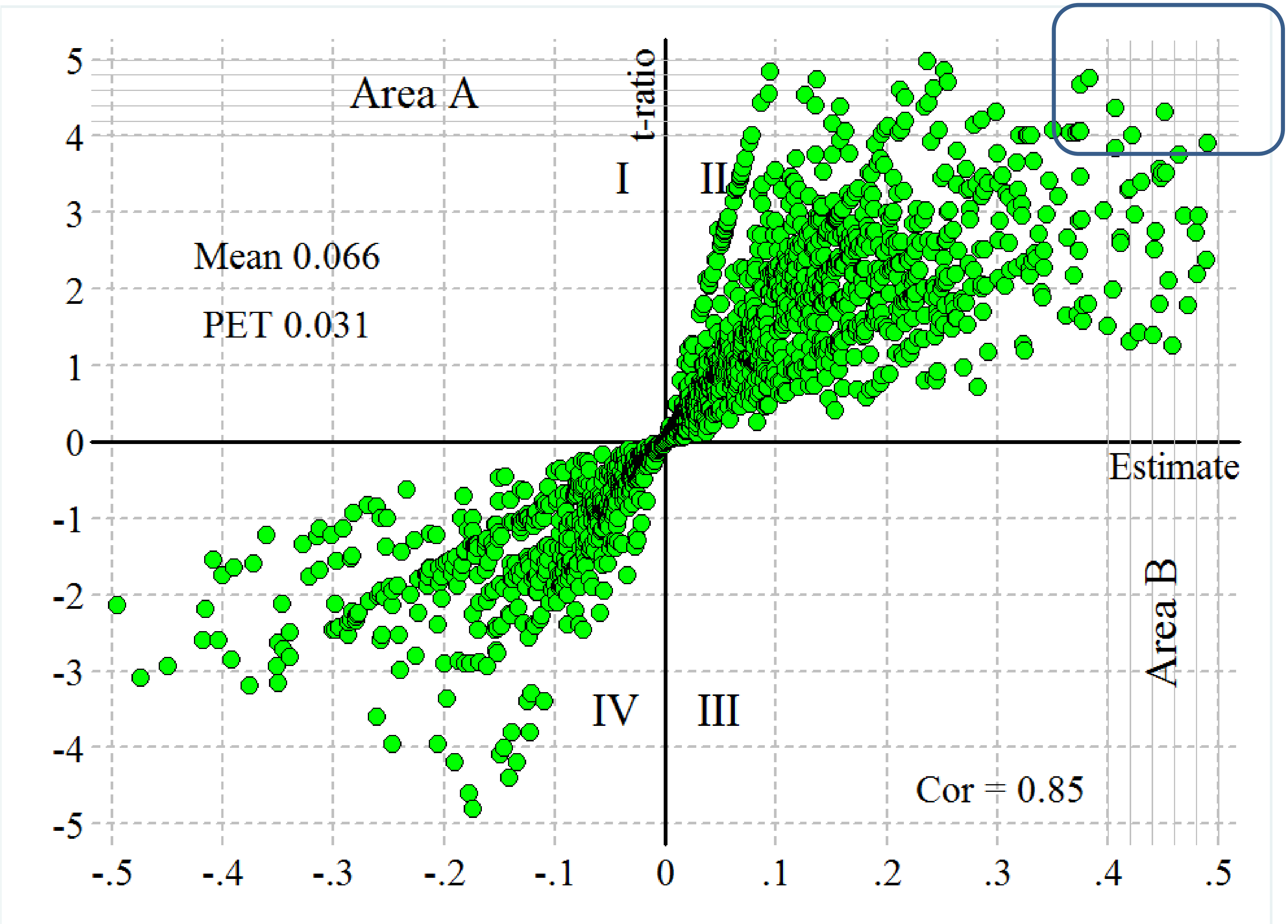
- I take two results to generalize:
  Big variation + frequent bias

From analytical solution to simulations

- **Two years ago**: I presented a theory explaining *J* based on marginal costs and benefits of running regressions.
- Published in *Econ Journal Watch* 10(2), 136-56
- Showing: marginal costs have dropped ➔ *J* must rise
- I looked at SRs (selection rules) I could solve analytically. I could solve a few, but missed important ones

- **Today**: *J* is exogenous – to study effect of different *J*s
- I make 5 *SR*s that I believe are the main ones in practice.
- These SRs are simulated on same *J*-sets so easy to compare
- How does J-sets look: PS look at *N*-set (hmm)

The example: the 1,777 published estimates of aid effectiveness. PS 90 extreme deleted



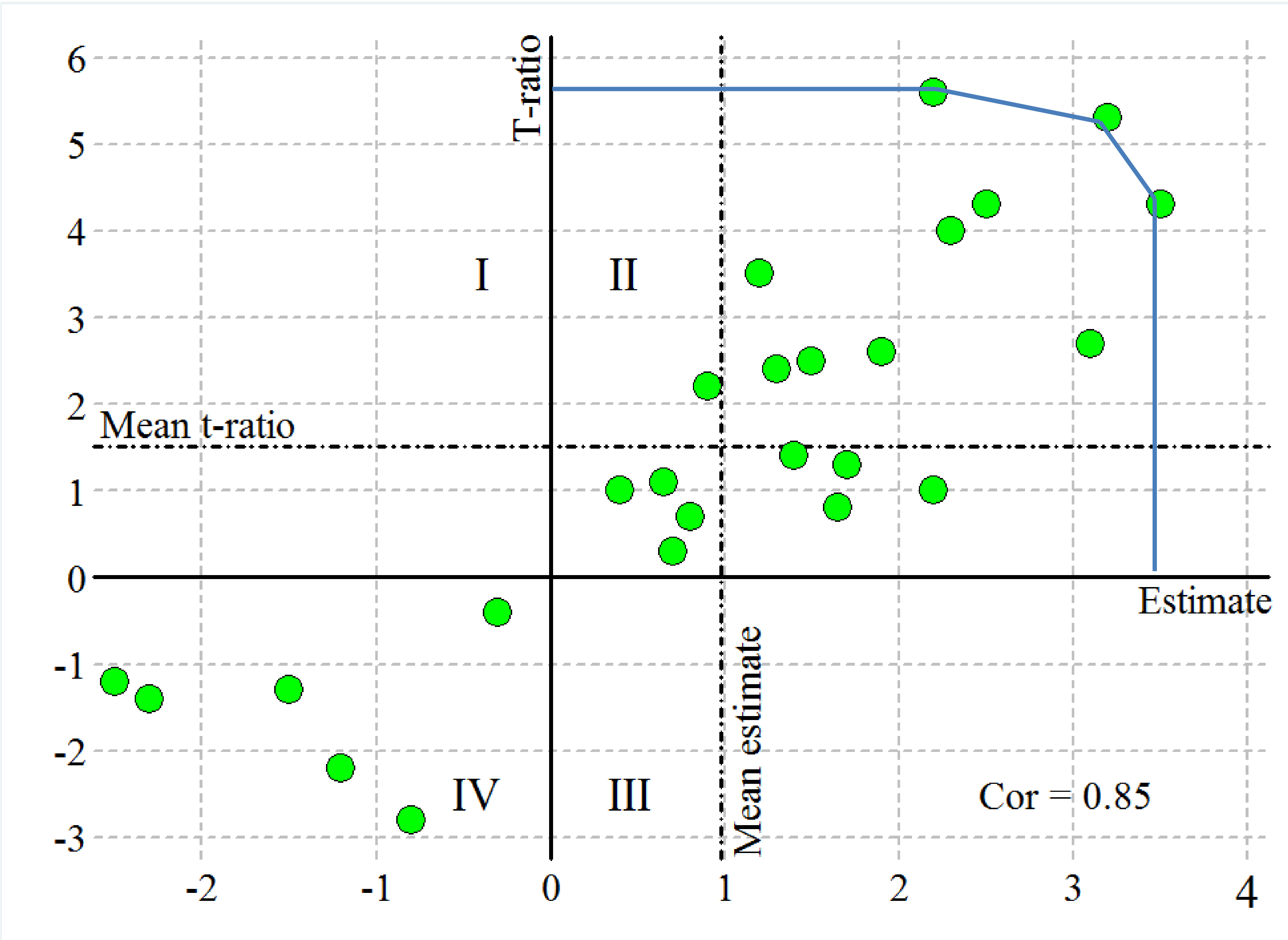
Are fit and size related?

1. Correlation $r(b_i, t_i) = 0.85$ is strong relation

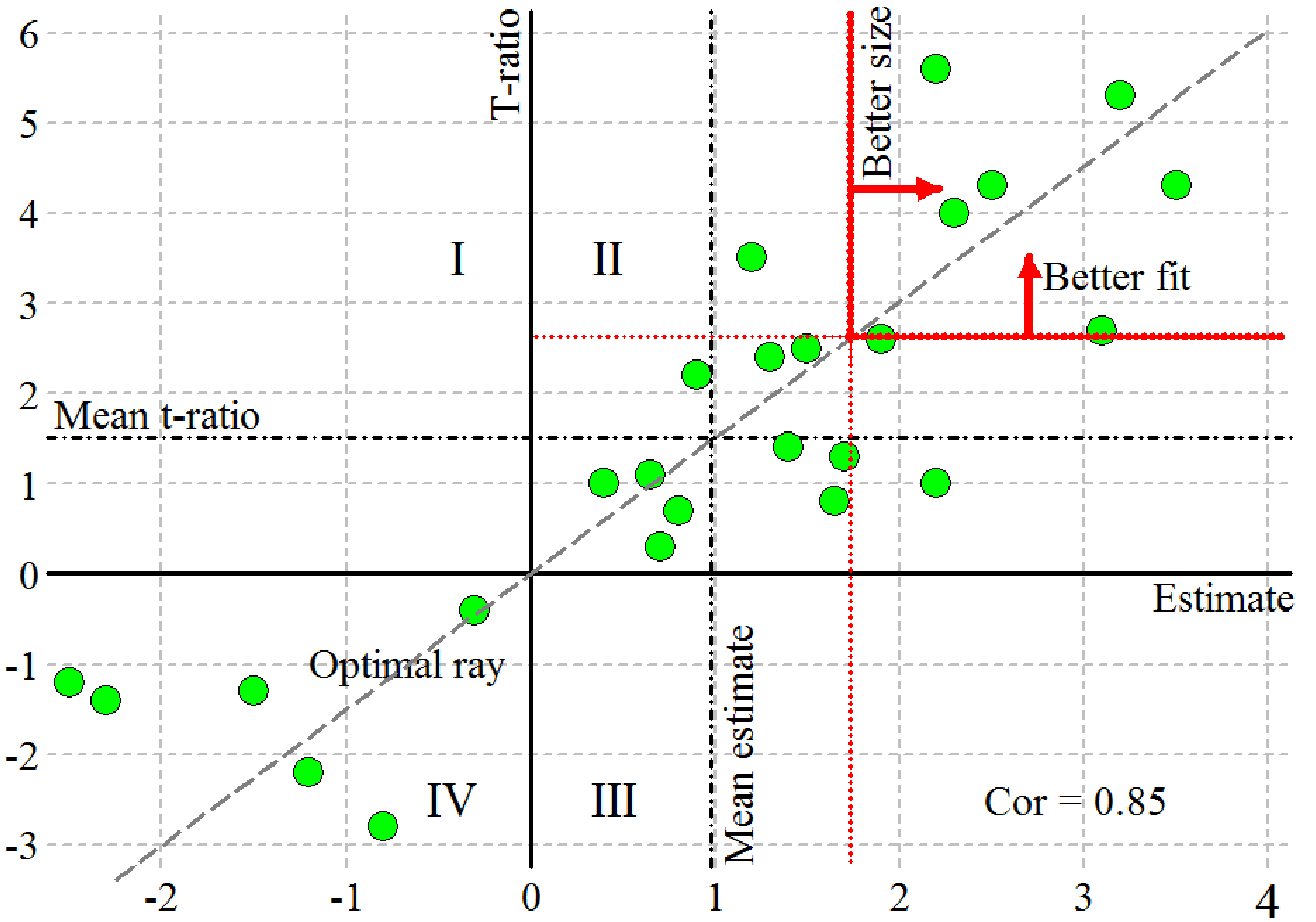
2. Area A: marked with horizontal lines: High fit
   Presumably chosen by fit-prior SR2
3. Area B: marked with vertical lines: Large size
   Presumably chosen by size prior
4. Overlapping area both high fit and size
   very few points. Thus, weak relation

5. Hmmm: Conflicting evidence

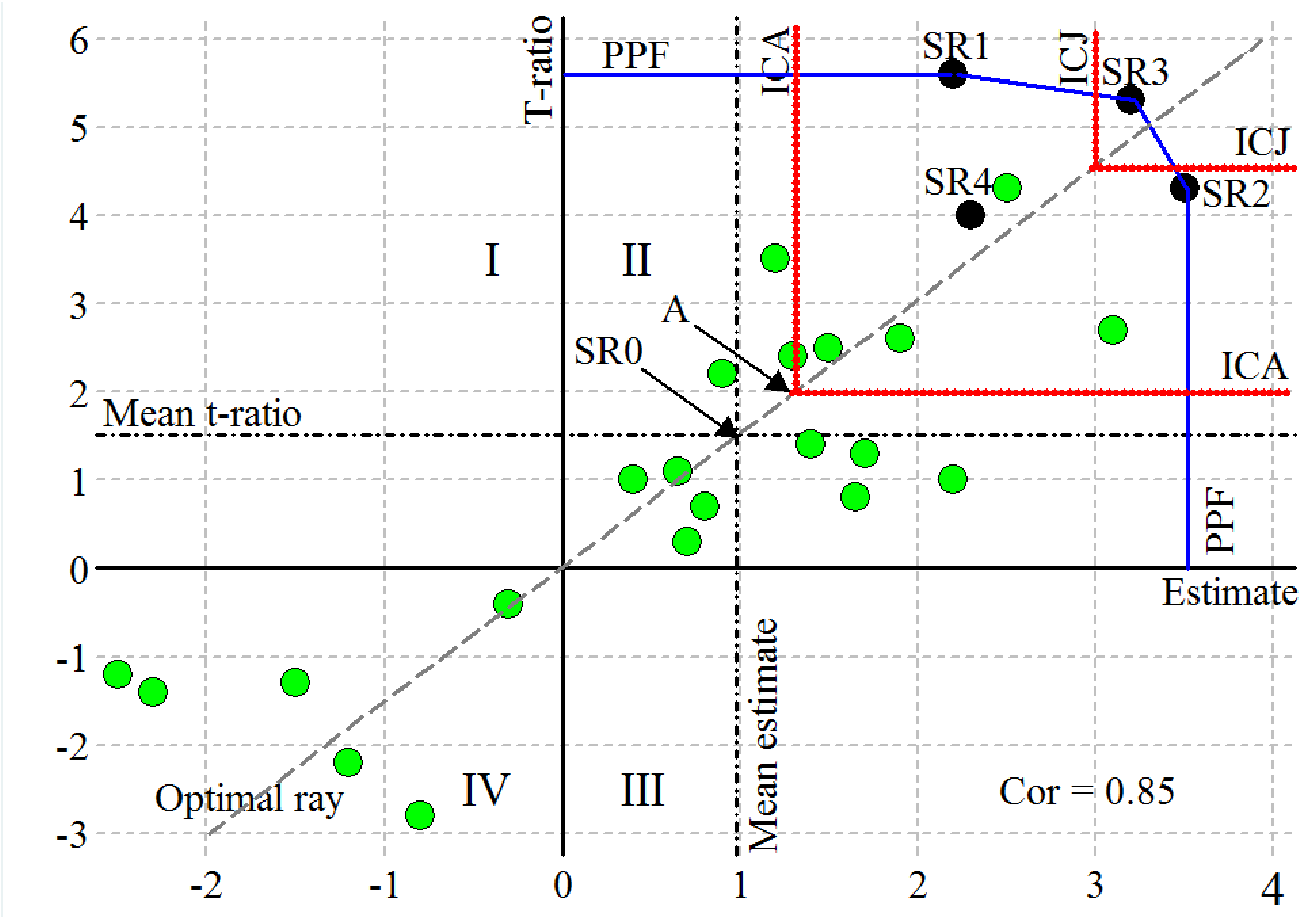
The 25 regressions of the *J*-set. $\beta = 1$. Rim is PPF



Indifference curves: (i) horizontal, (ii) vertical, (iii) kinked
Rays: better the further out. Choice of optimal ray



Combining size and fit: The PPF/IC framework



Note on $r(b_i, t_i)$

- Correlation chosen to be as high as before

- All SRs give different points, but:
- But easy to make examples where they are the same
- Especially for small values of J

- Evidence still conflicting but it suggests
- For $J = 1$ the same, as J grows it will grow

Simulation program in stata by ***Jan Ditzen***.

Uses the Matryoshka set-up: Show 5 levels

Experiments: Uses 6 levels

- **Level 1:** *R* experiments: $R = 1$ illustration, $R = 1'000$ production
- **Level 2:** Seven *J*s: $J = 1, 5, 10, 15, 25, 34, 50$. $\Sigma = 140$
- **Level 3:** One *N*-set of 500 selected: Sample $m = 21$ to 520
- **Level 4:** One selected regression by each of five *SR*s
- **Level 5:** One *J*-set. Using same *m* and *J*
- **Level 6:** One regression on *m* simulated data
- **Output 6:** Regression output: (b_j, s_j, Df_j)
- **Output 5:** *J* output: (b_j, s_j, Df_j), $j = 1, \ldots, J$
- **Output 4:** 5 selected *SR*-regressions: $(b_i, s_i, t_i, p_i, Df_i)$
- **Output 3:** 5 *N*-set: Each gives one *SR*-funnel
- **Output 2:** A funnel-set of $7 \times 5 = 35$ funnels
- **Output 1:** The two cases:
- Case 1: $R = 1$ gives 35 funnels
- Case 2: $R = 1'000$ give 5 tables with 7 rows

PS: The PET is made to adjust for censoring *SR*2

- **Two important questions**

- **Q1:** Does the PET work for *SR*1, *SR*3 and *SR*4?

- **Q2:** How different are the outcomes for the four *SR*s?
  - Notably: How different is *SR*1 and *SR*2 ?
    The two extremes. Dream: They are the same!

Some of the nitty-gritty

- Data generating process: DGP $y_t = \beta\, x_t + \varepsilon_t$ where $\beta = 1$
- Estimation model (OLS): EM $y_t = b\, x_t + u_t$
- Variation: m = 21, …, 520, $\varepsilon_t = N(0, \sigma_\varepsilon^2)$ and $x_t = N(0, \sigma_x^2)$
- To get enough variation $\sigma_\varepsilon^2 = 10$ and $\sigma_x^2 = 2$

The numbers in the simulations:

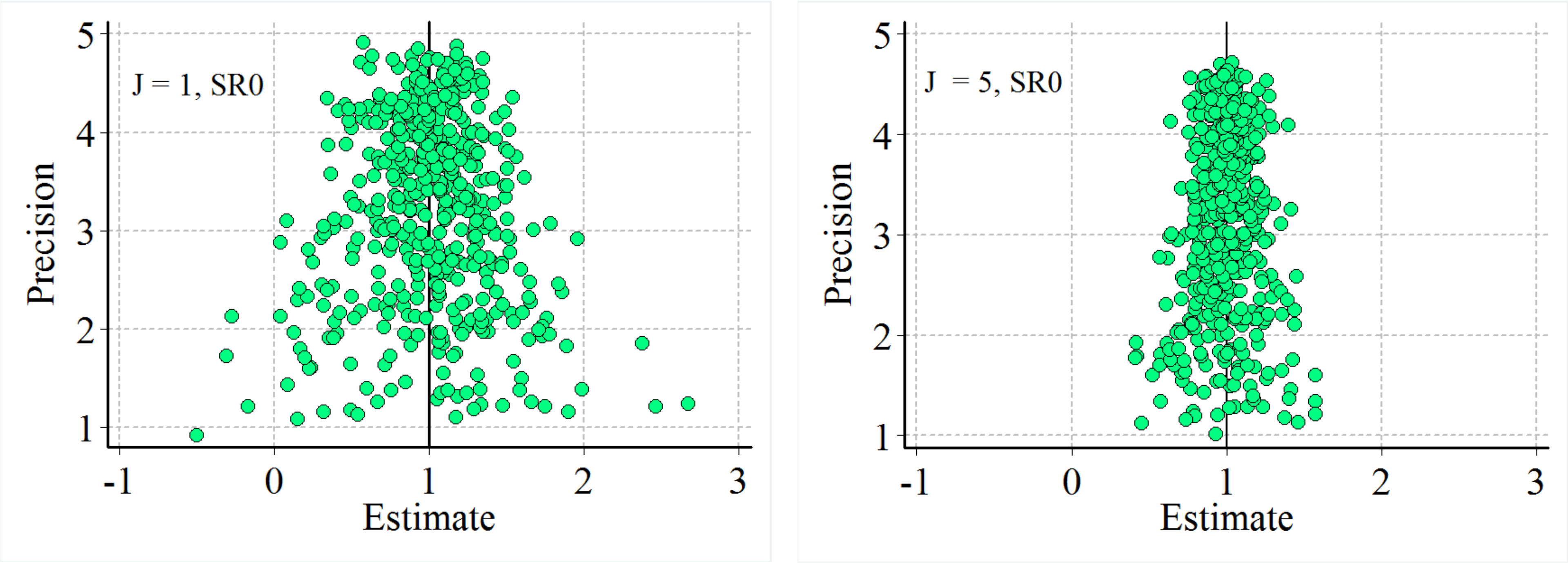
- R = 1 funnel-set (5 *SR*s, 7 *J*s) is 35 funnels
  PS ΣJ = 140 so 140 x 500 = 70,000 regressions
- R = 1,000: gives 1,000 *x* 70,000 = 70 *x* 10^6 regressions.
- One week for a strong pc's working day and night
- PS: Tom Stanley prefers 10,000 funnel sets!
- But: If the points for one *SR* is smooth for different *J*s
  Then you use all the regressions to 'justify' each other

Results for SR0. Select the mean or median

- Why: You plan the best set of regressions, run them and report the average + the std.

- PS: for J = 1, the ideal funnel. Its width corresponds to the t-ratios, and it is nicely symmetrical
- Thus $\underline{b} \approx \beta_M \approx \beta$
- And when J goes up the width falls with $\sqrt{J}$, but same t's
- Thus the funnels become leaner and leaner

SR0 baseline:
For J = 1 (ideal) and 5



Is $SR0$ realistic?

- We know: empirical funnels are wide relative to t-ratios

  Ideal funnels have $J = 1$, they should be wider as J rises
- But $SR0$ gives funnels that become more and more narrow relative to the t-ratios
- Thus, $SR0$ must be rare in practice

- **Now to the two extreme *SR*s**

  Remember dream!

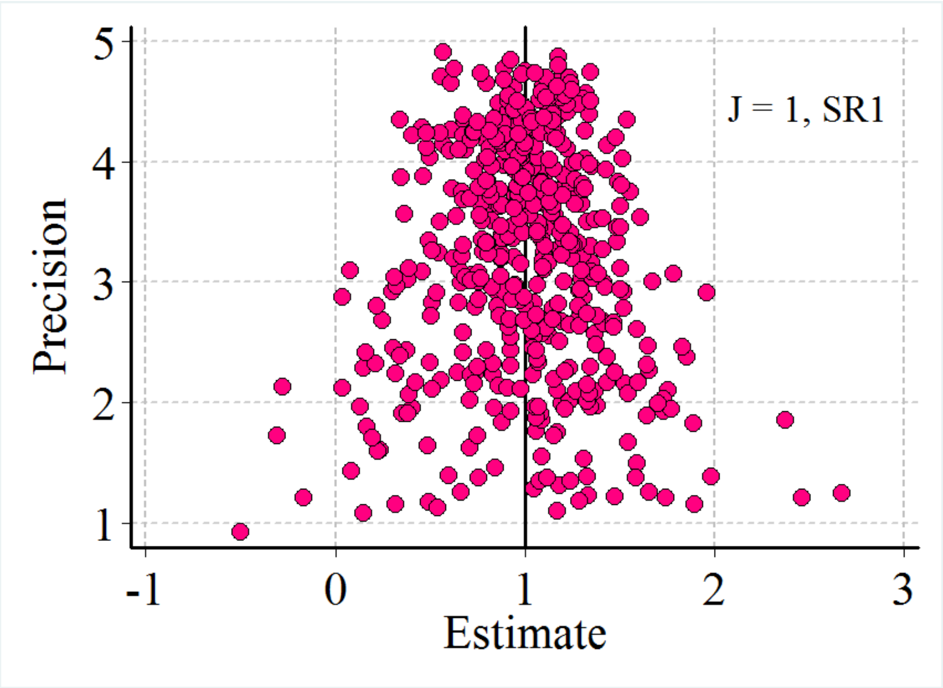
SR*1**: Best ***fit, highest t-ratio
SR*2**: Best ***size, highest b-estimate

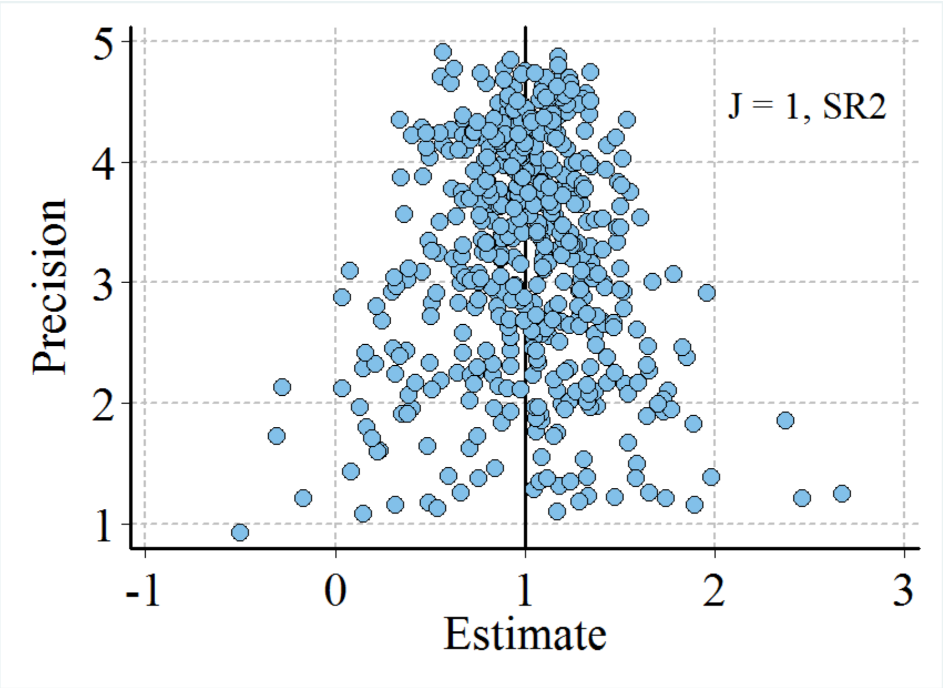
- Shown as a cartoon: Illustrated by 1 funnel for each J = 1, 10, 25 and 50

- *SR*1 is a little tricky: Drawn with p over b. When J goes up so do p but b goes up as well: $t = b/s = bp$ ➔ $p = t/b$, so t and b rise almost the same.

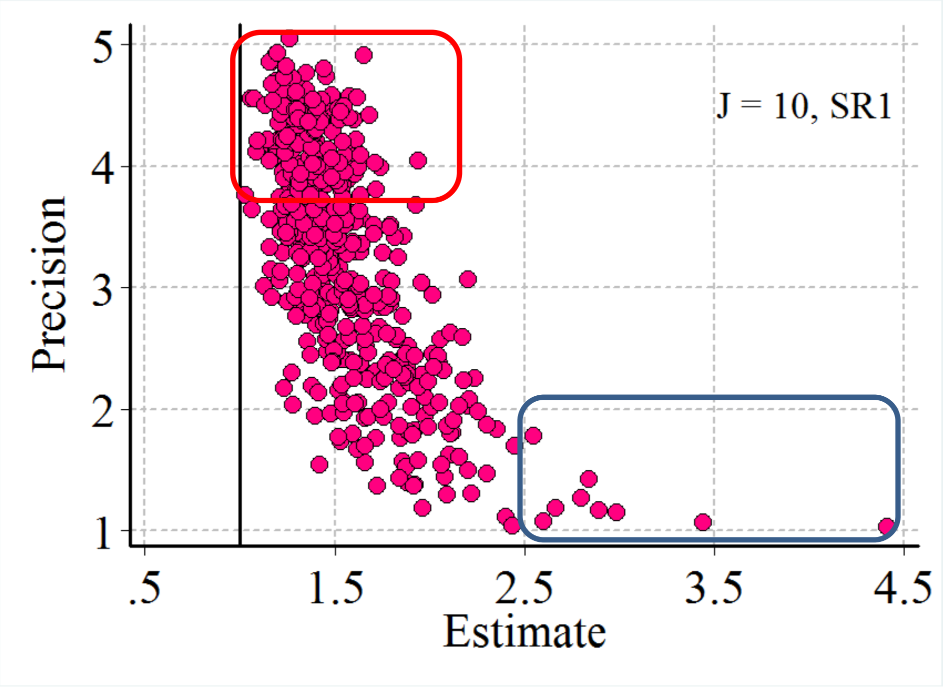
- For J up both funnels more sausage-like.
- Most different:
- For small b's where you still get some high t's and
- For small t's, where you still get some high b's

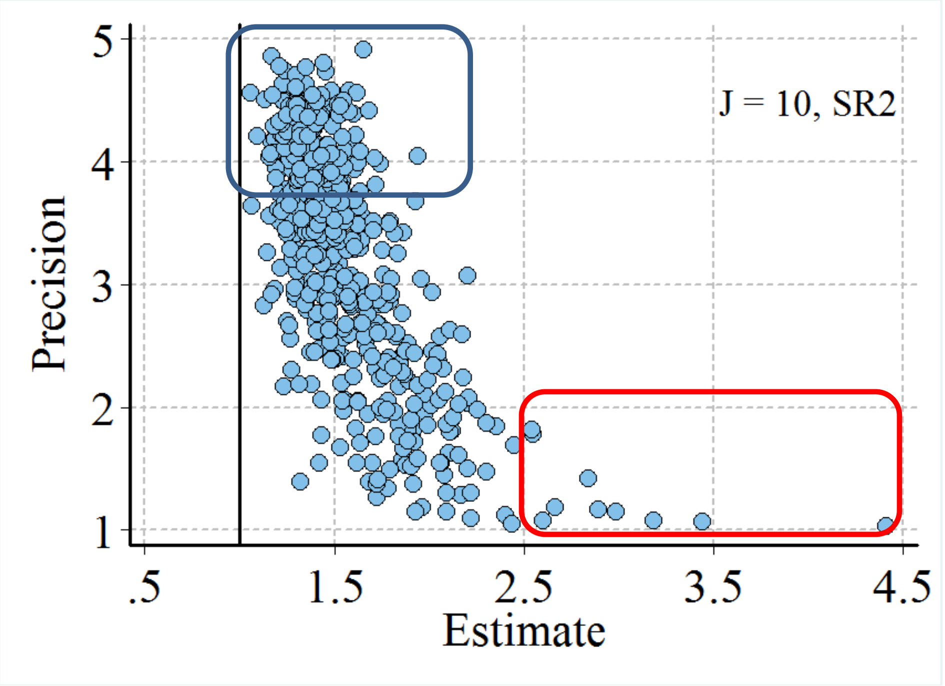
Comparing *SR*1 (polishing) and *SR*2 (censoring):
For J = 1. Here the two funnels are the same
This is the ideal funnel (same as before)



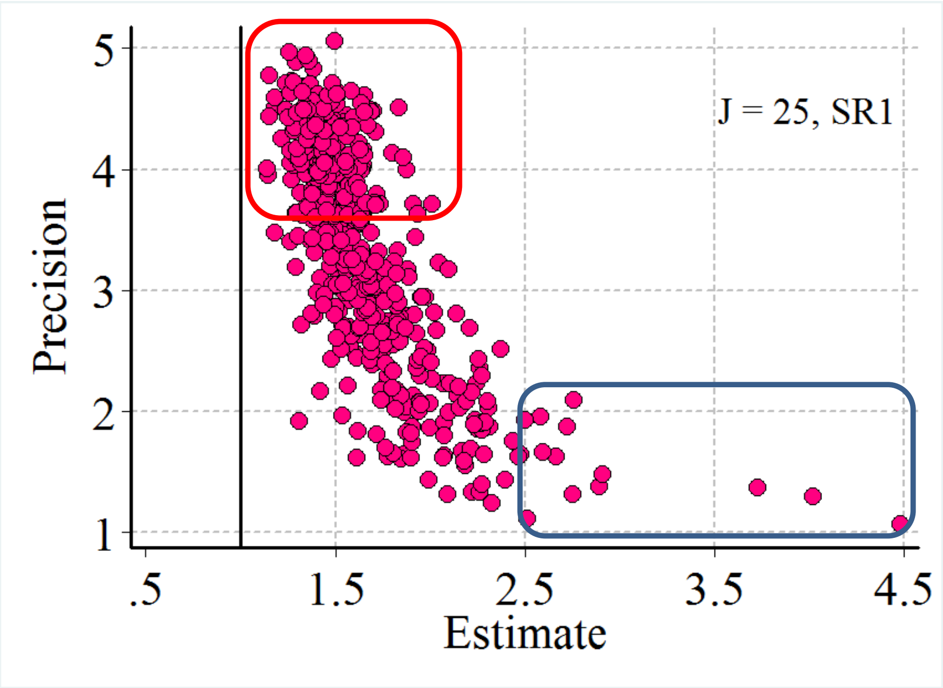


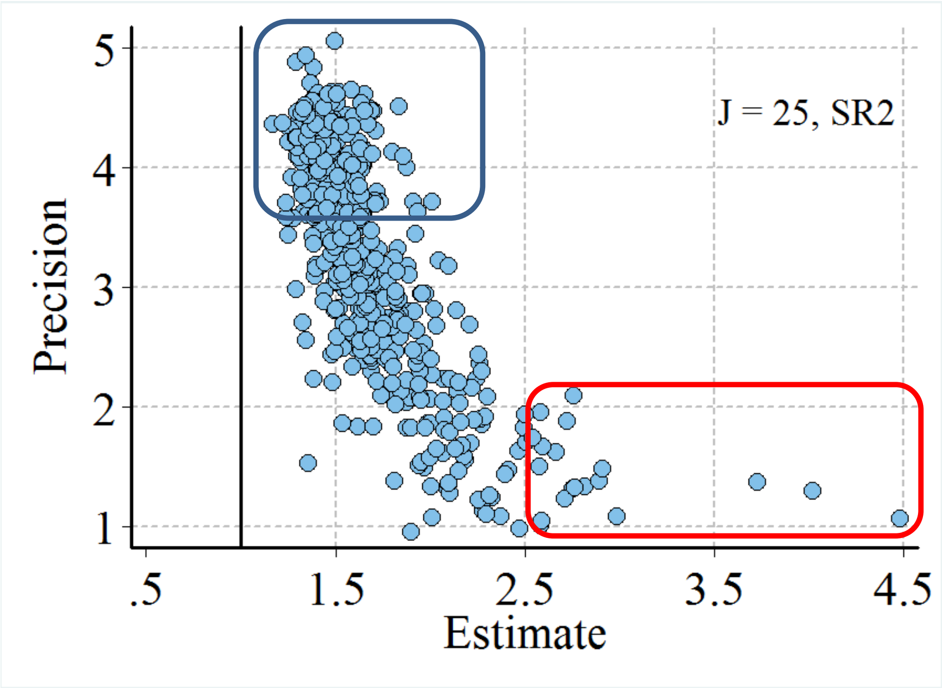
Comparing SR1 (polishing) and SR2 (censoring): For J = 10. The two funnels still similar



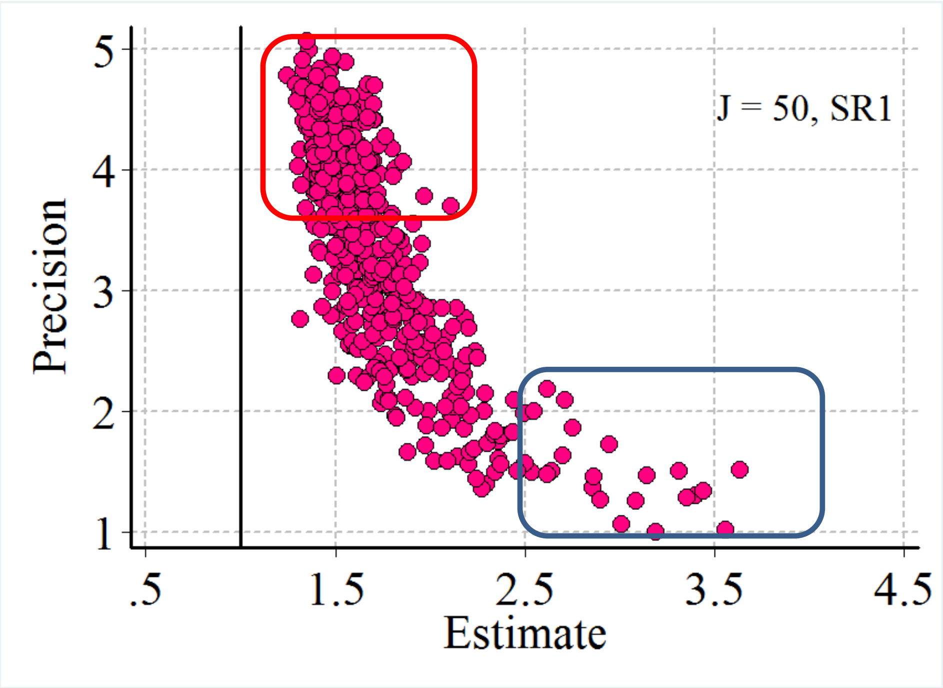


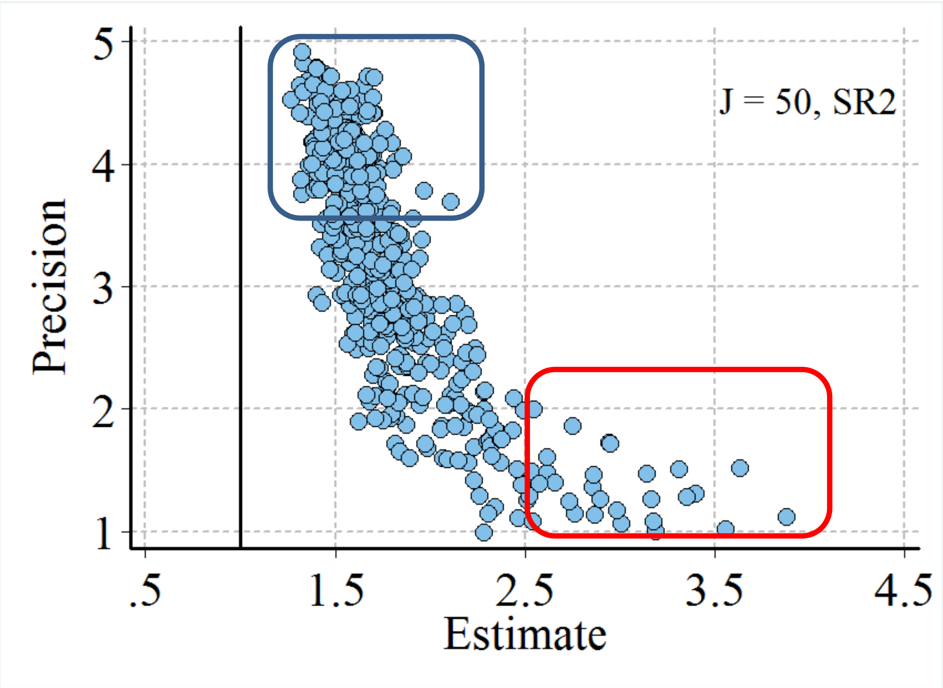
Comparing SR1 (polishing) and SR2 (censoring): For J = 25. Differences starts to grow but they are not big!





Comparing SR1 (polishing) and SR2 (censoring): For J = 50. The two funnels still similar





(Table 4). SR1, the polished funnel Undershoots a bit

(1)	(2)	(3)	(4)
J	$\underline{b}$	β_M	β_F
1	**1.00**	**1.00**	0.00
5	**1.42**	**0.98**	1.22
10	**1.54**	**0.97**	1.62
15	**1.61**	**0.97**	1.82
25	**1.69**	**0.96**	2.08
34	**1.73**	**0.96**	2.22
50	**1.78**	**0.95**	2.39

(Table 5). SR2, the censored funnel. Overshoots a bit

(1)	(2)	(3)	(4)
J	$\underline{b}$	β_M	β_F
1	**1.00**	**1.00**	0.00
5	**1.43**	**1.00**	1.16
10	**1.57**	**1.01**	1.51
15	**1.64**	**1.01**	1.69
25	**1.73**	**1.02**	1.91
34	**1.78**	**1.02**	2.02
50	**1.84**	**1.03**	2.16

Results from *SR*2 look remarkably like *SR*1
Bias in mean that grows with *J* to 80 %

- PET bias from negative (for *SR*1) to positive (for *SR*2)

- PS: *SR*2 is what the PET is made for, and it works well
  about 2-3 % wrong only!

- **Now *SR*3**: The best combination of fit and size
- It is almost the average of the results for *SR*1 and *SR*2

(Table 6). Selection rule SR3. Super fit of β_M

(1)	(2)	(3)	(4)
J	$\underline{b}$	β_M	β_F
1	**1.00**	**1.00**	0.00
5	**1.43**	**0.99**	1.16
10	**1.56**	**0.99**	1.51
15	**1.63**	**0.99**	1.69
25	**1.72**	**0.99**	1.91
34	**1.76**	**0.99**	2.02
50	**1.82**	**0.99**	2.16

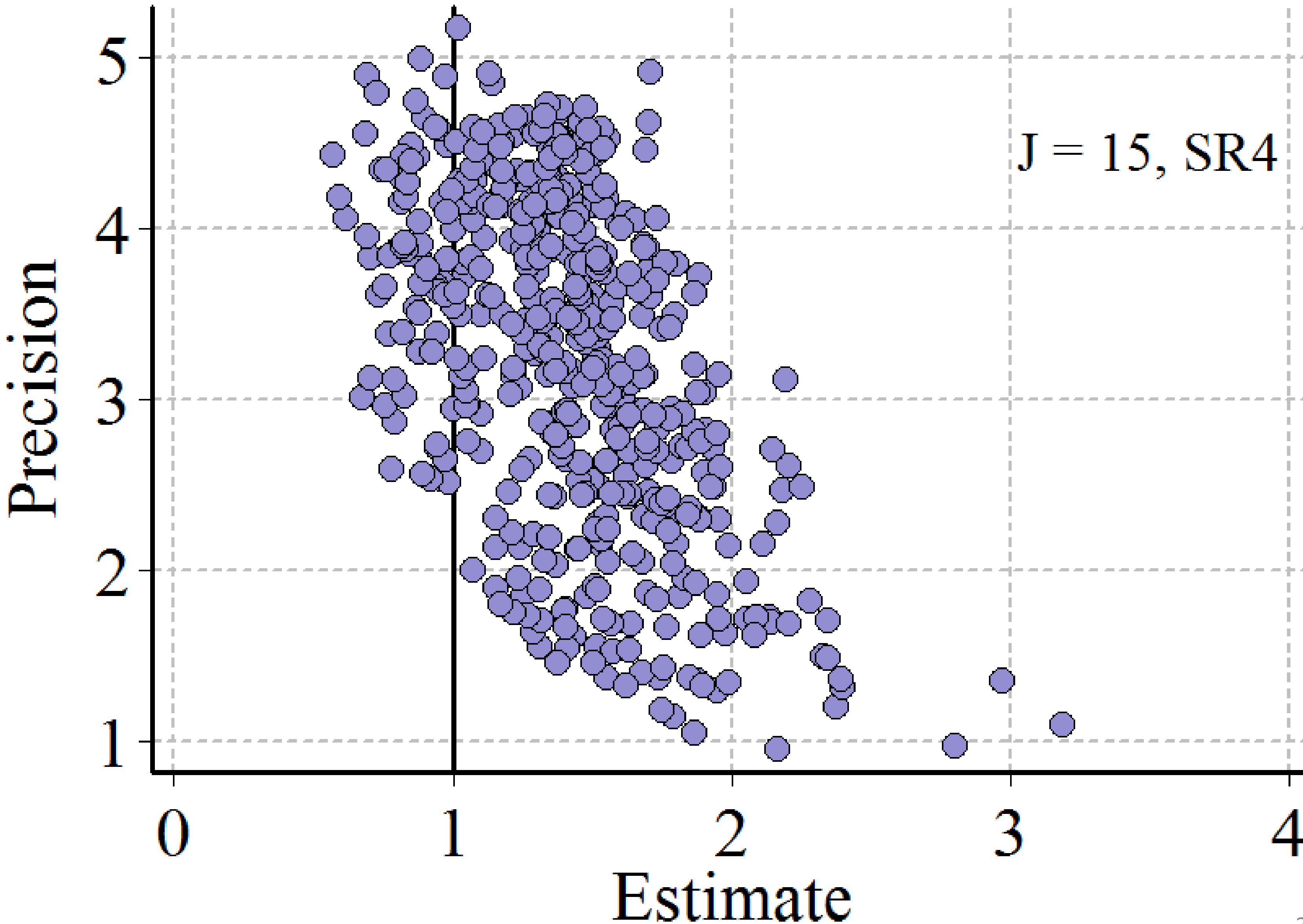
Results for *SR*3 and *SR*4 funnels and tables in paper

- *SR*3 is a compromise between *SR*1 and *SR*2. As they look the same, so does *SR*3. As the PET bias is negative for *SR*1 and positive for *SR*2 it is really small for *SR*3: typically -1 % (small overshooting)

- *SR*4 is different as J is endogenous. For $J = 5$ it looks like the previous. As J rises it becomes a mixture, and there are some values below 1 all the way up.

- I show the case for *SR*4 and $J = 15$

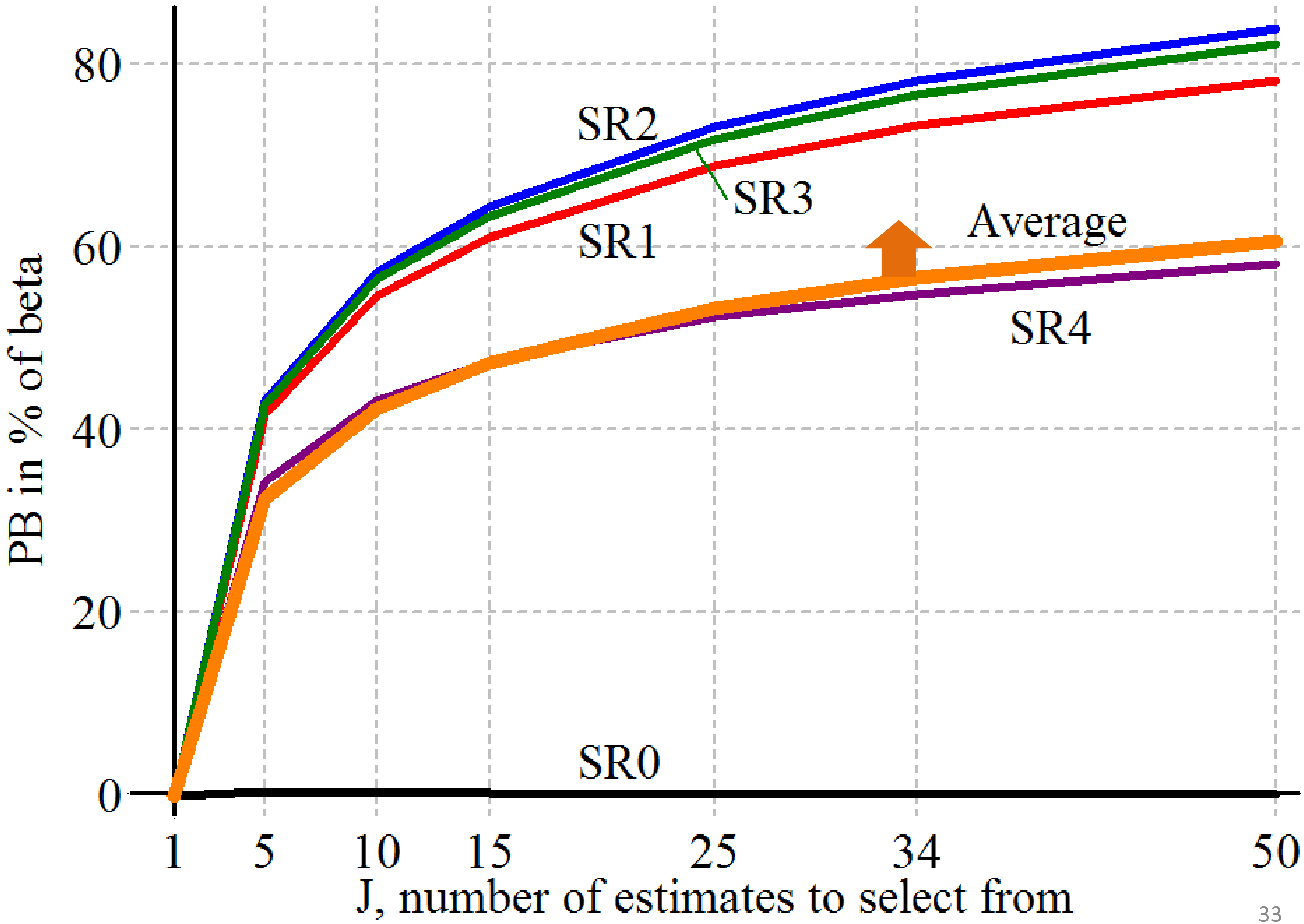
SR4: J = 15



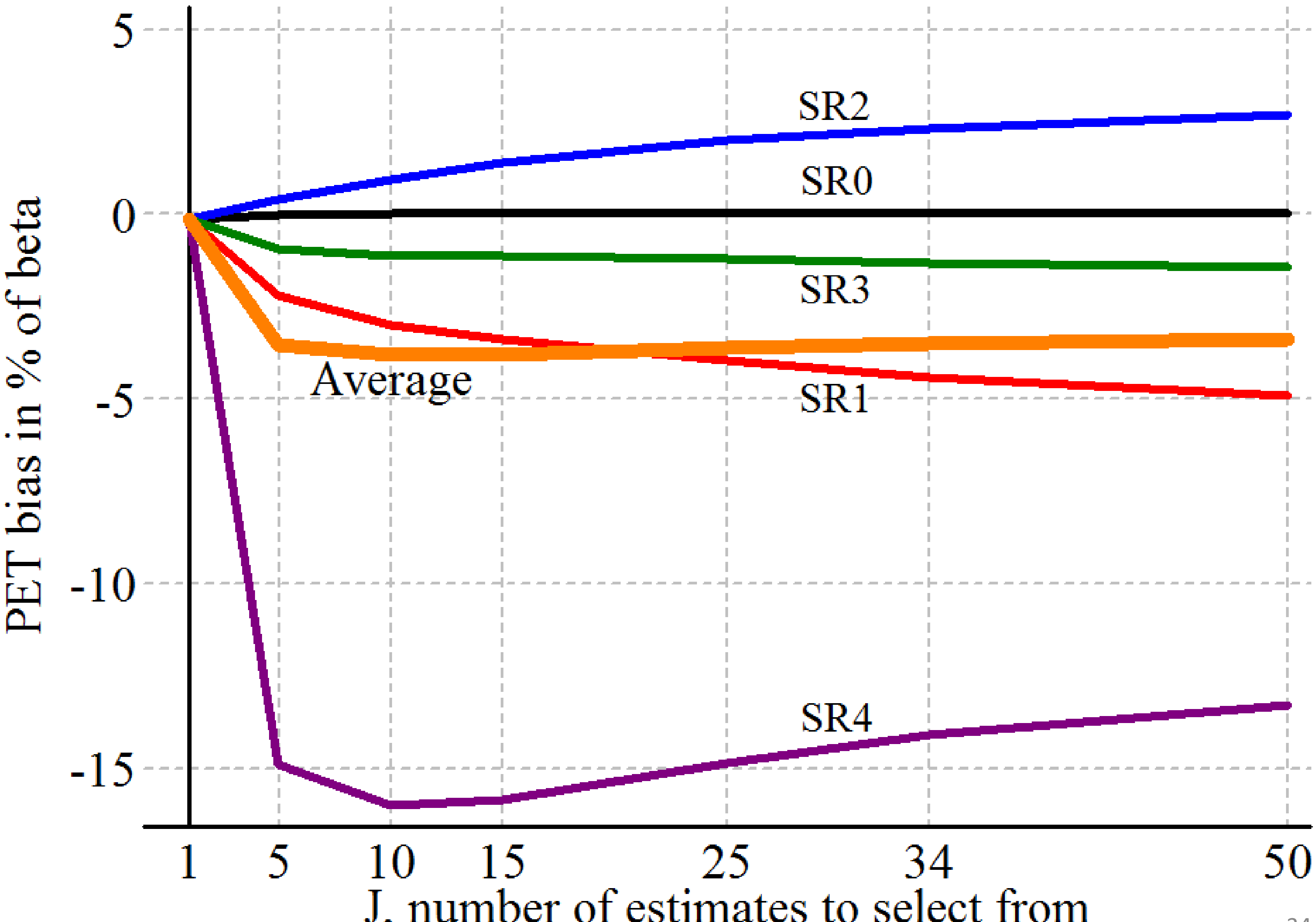
Now I compare graphically:

- One graph covers one statistic PB^T, PB_{PET}, FAT, μ
- The 7 Js are 7 points on the horizontal axis
- Each graph has six curves:
- One curve for each SR + the average
- PS: researchers use different Js and SRs

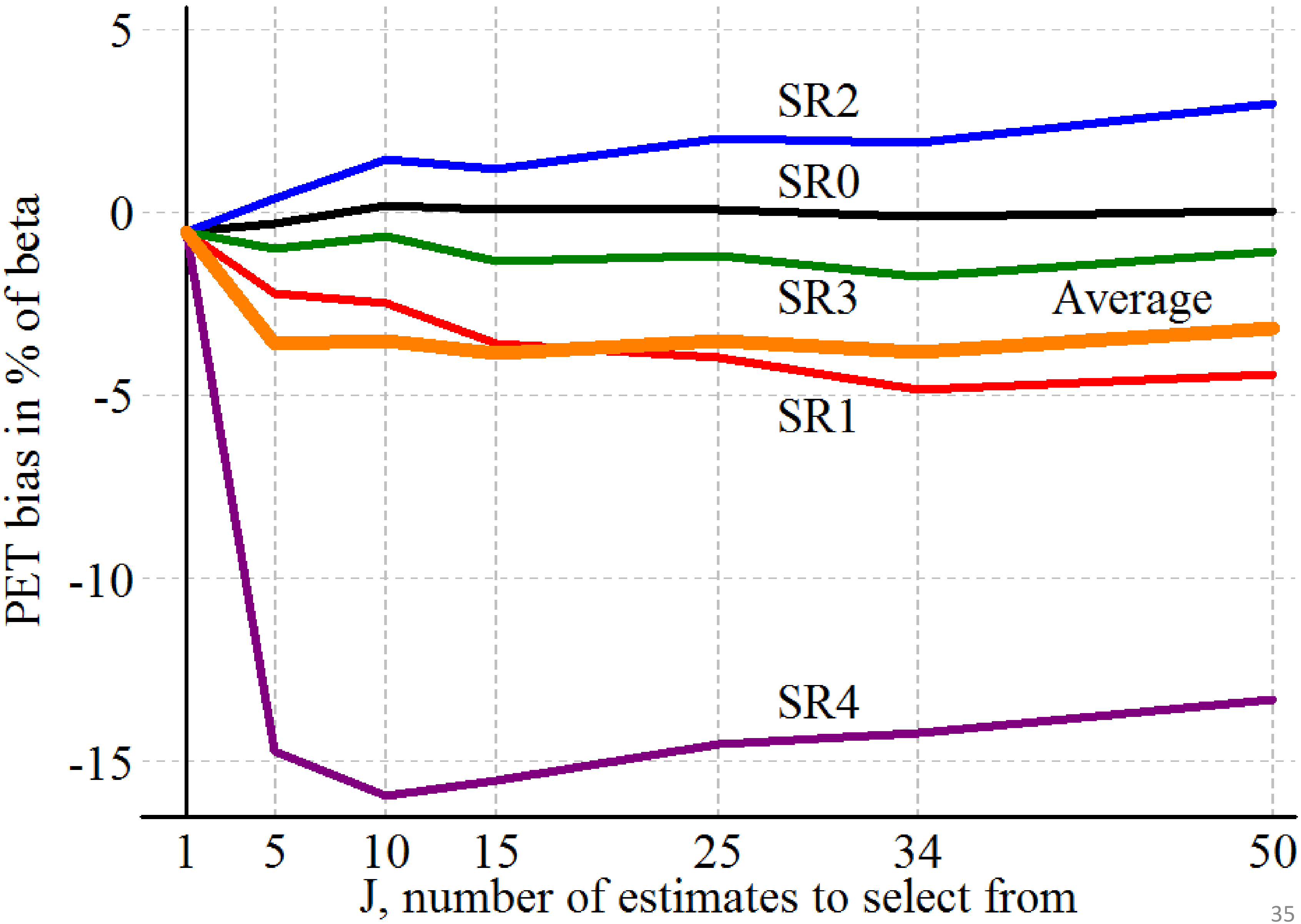
True publication bias of the mean. PS average at 50 to 70%



PET bias. **Scale up 4 times.** Most + average within ±5%



Same for 100 experiments. Not very smooth



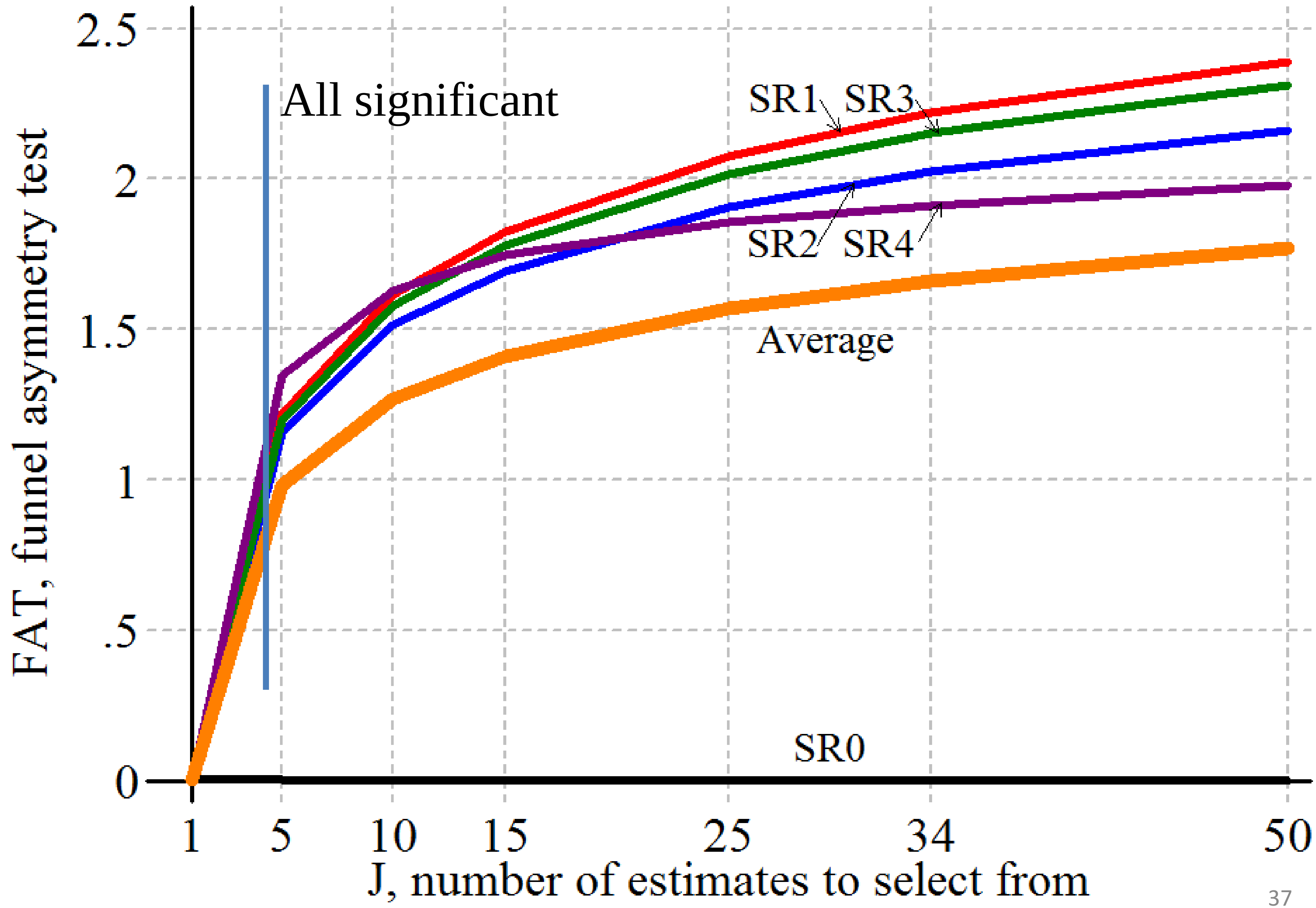
Conclusions on publication bias:

- Selection of ‘best’ results ➔ always gives positive bias
- The bias is substantial: About 50-60 % (incl 20 % SR0)

- The main prior $\beta > 0$ causes $\underline{b} > \beta$
- This confirms: **The exaggeration result**
- Other term: **The theory confirmation bias**

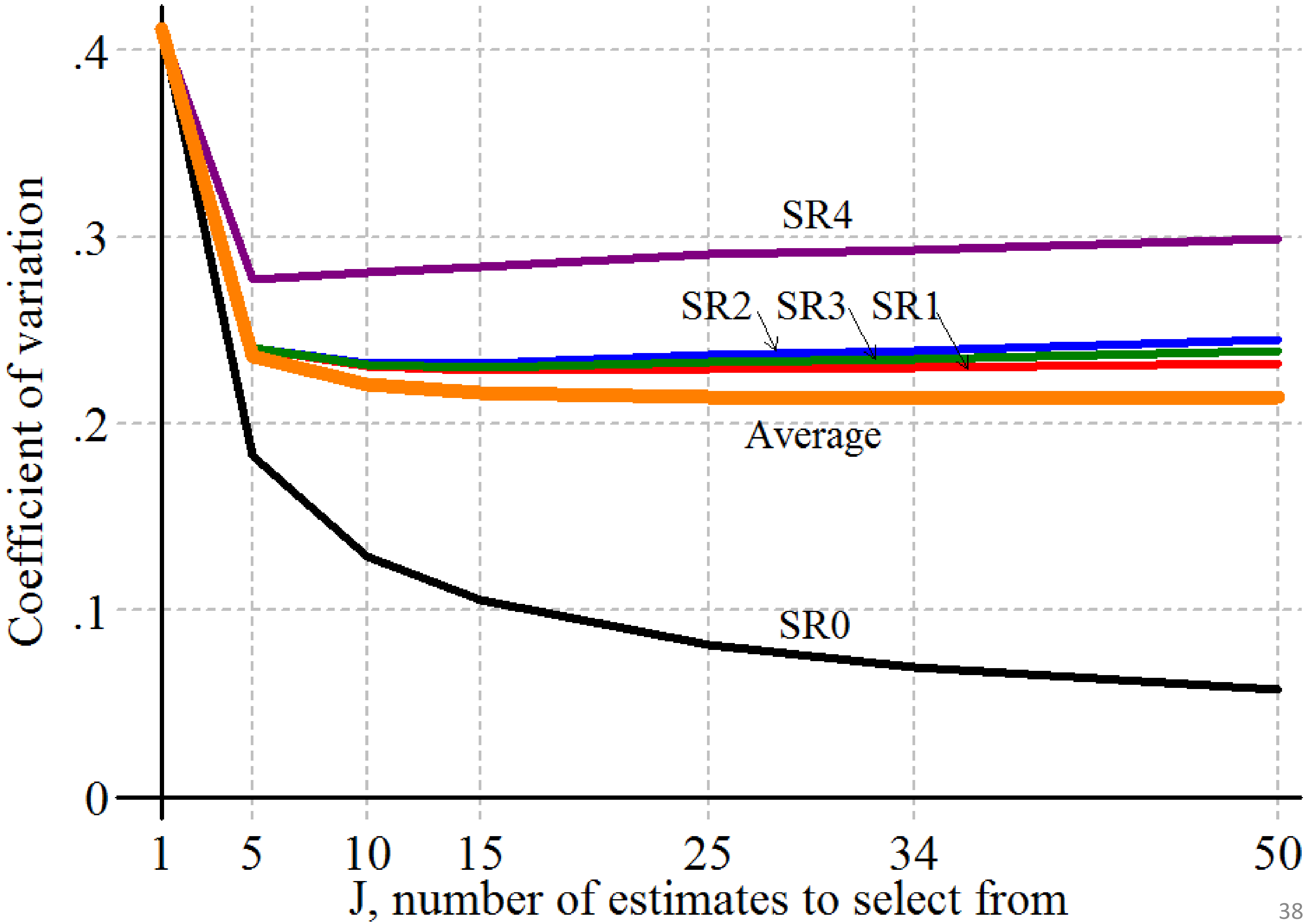
- The PET is much closer to the true value: In average it is less than 10 % of the *PB*, i.e., within 4 % from β

The FAT



All significant
SR1
SR3
SR2
SR4
Average
SR0
FAT, funnel asymmetry test
J, number of estimates to select from
2.5
2
1.5
1
.5
0
1
5
10
15
25
34
50

Width of funnel. A problem. Empirical funnels are wide



My interpretation

- Simulations catch 2/3 of the typical publication bias
- The funnels observed are a mixture of the funnels simulated. So it looks realistic!

- The PET catches the true value of β amazingly well
- It does not matter if the SR is SR1, SR2 or SR3
- The PB found is 1.5 – 1.7
- There is more due to model variation: It is rather PB = 2

PET or PEESE – does it matter?

- Exchange equation in simulations – one more time 70 mill simulated regressions.
- Easy to do, and then you just run your computer for a week. With no stops
- Results are mostly marginally different

- But the PET is normally a little closer to β in average and the PEESE has fewer rejections of true values

Comparing all 35 cases

	Best	***Best***	***Same***	***Don't Reject β = 1***	
	PB_{PET}	PB_{PEESE}	J =1	***PET***	***PEESE***
*SR*0	6	0	1	4	3
*SR*1	2	4	1	0	7 B
*SR*2	3	3	1	0	7 B
*SR*3	5	1	1	2	5
*SR*4	5 B	1	1	1	6
Sum	21	9	5	7	28

Missing/problems:

- Model variation. Difficult to simulate and less transparent.
  I think: It increases biases and μ

- SRs based on models with more coefficients:
  I think: It decreases biases but increase μ

- **The End**