Search Distaste at the Highway Stop Level

Jonah B. Gelbach

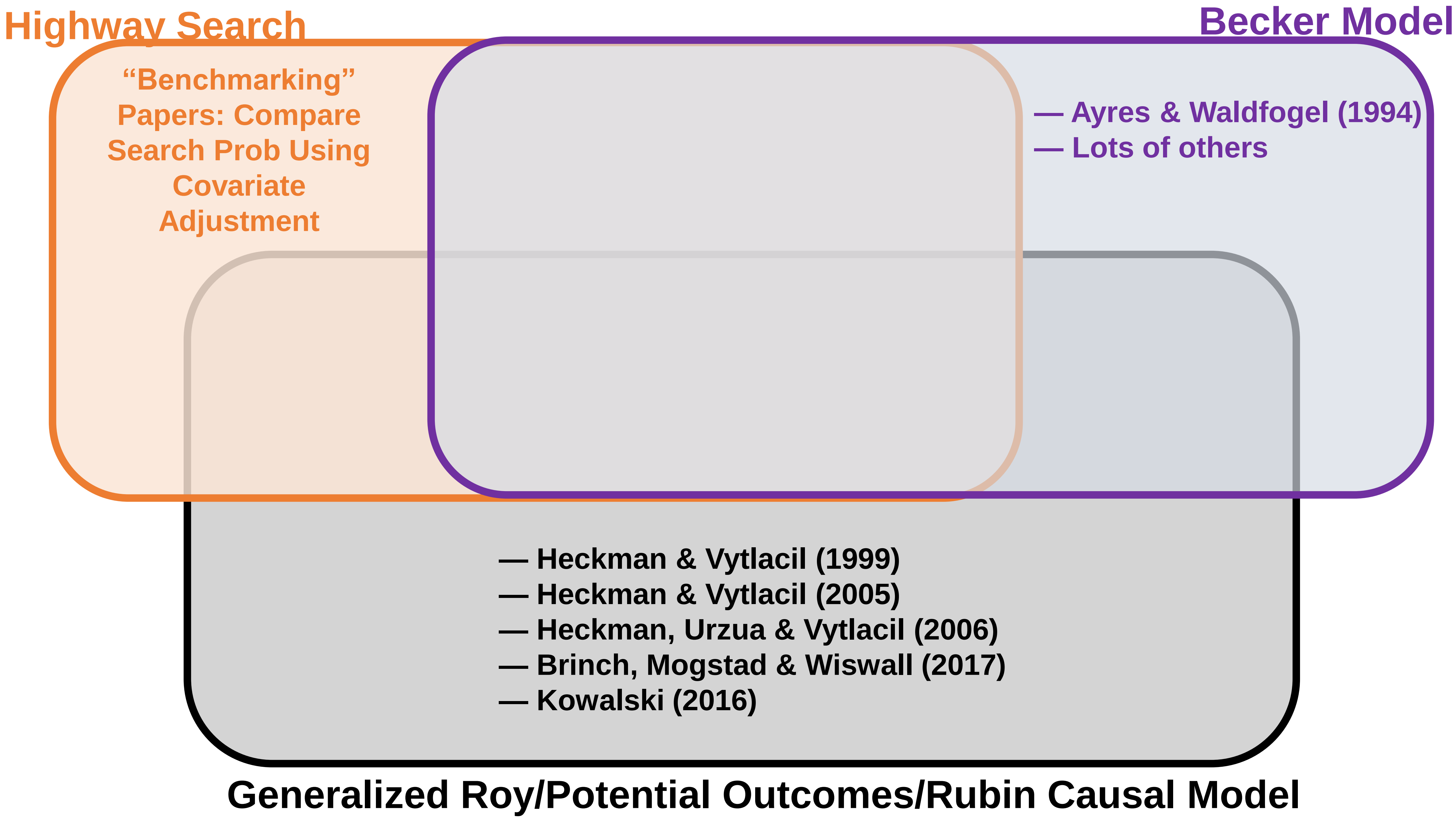
Professor of Law

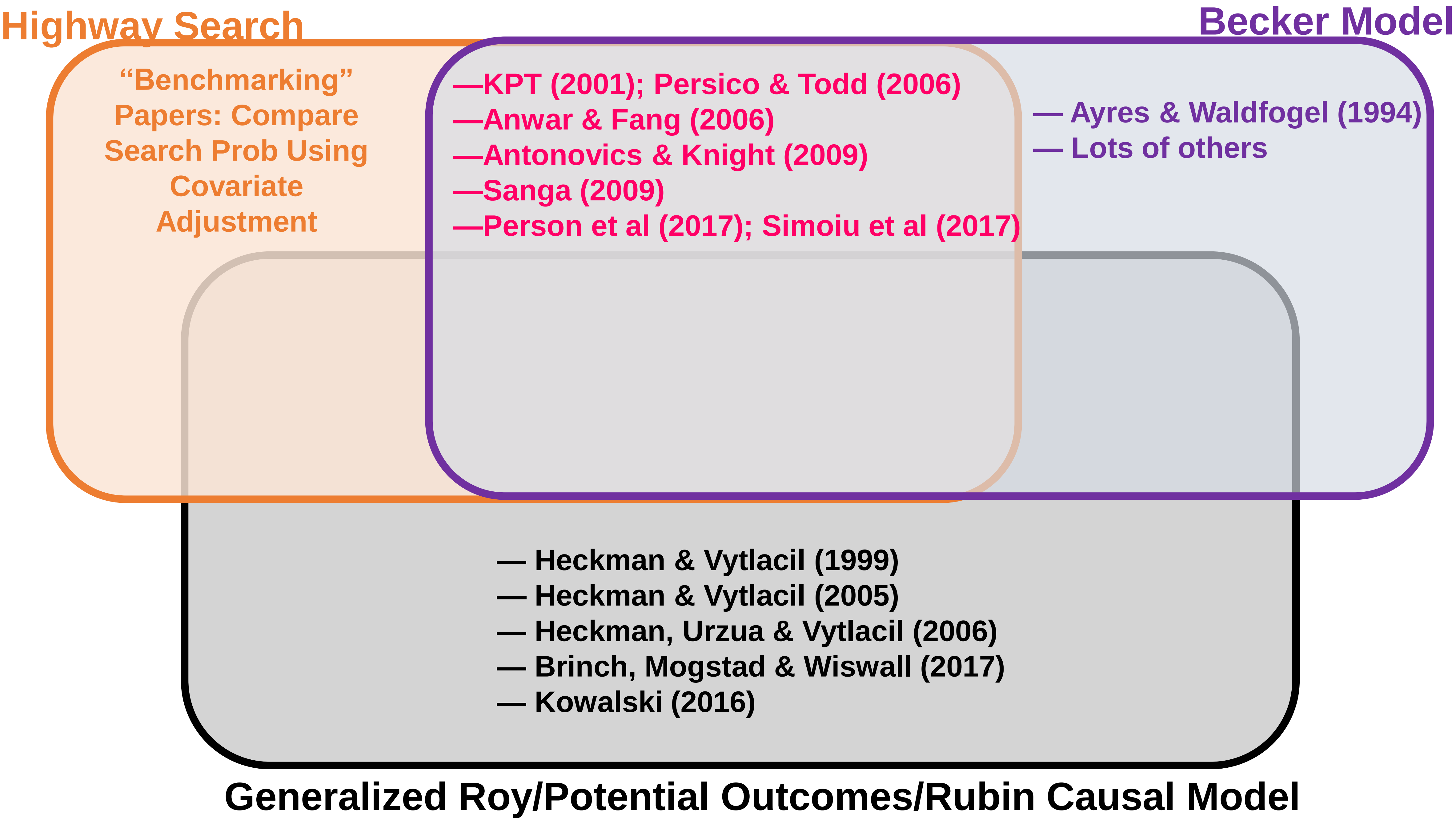
University of Pennsylvania Law School

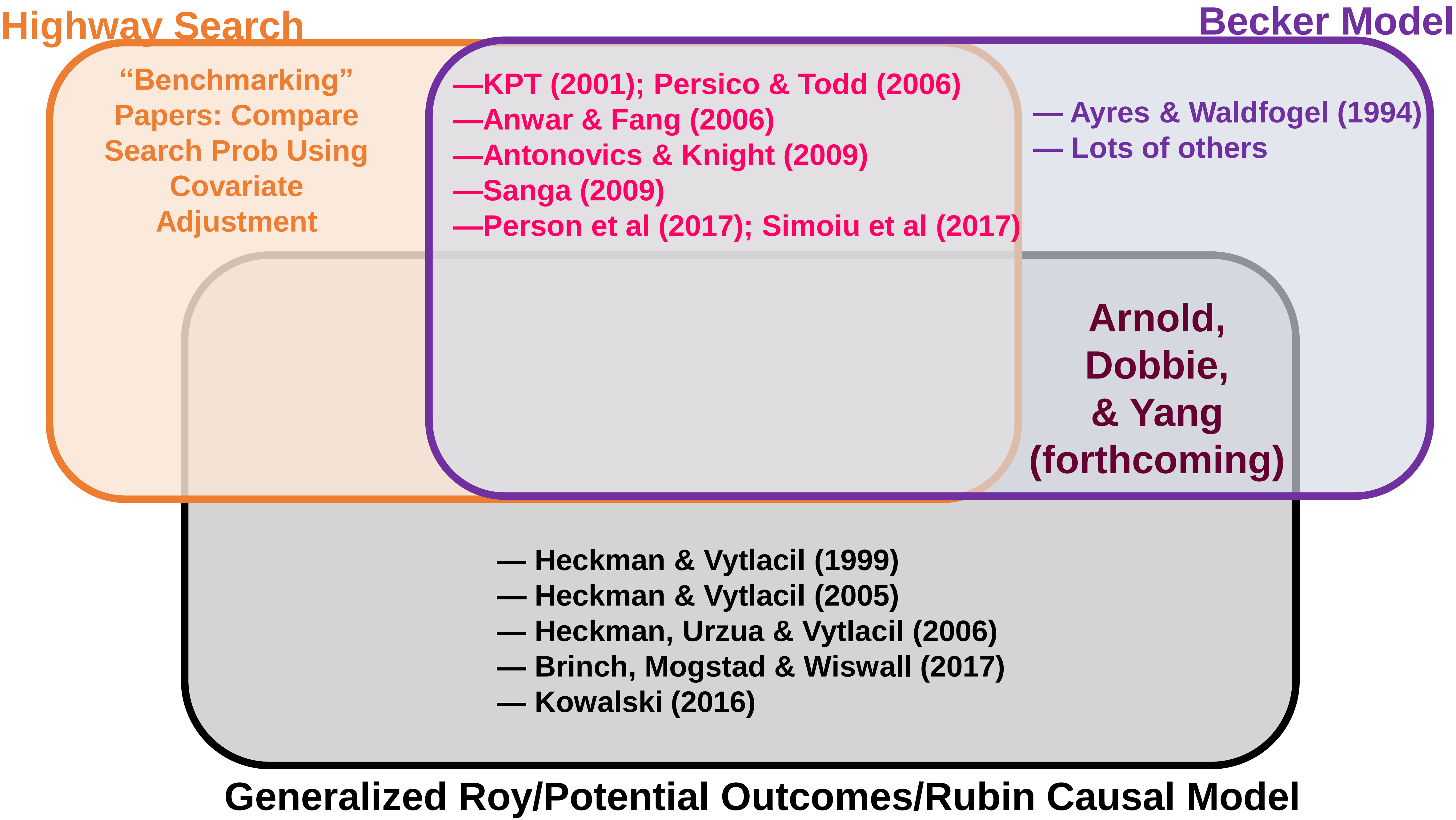
Highway Search

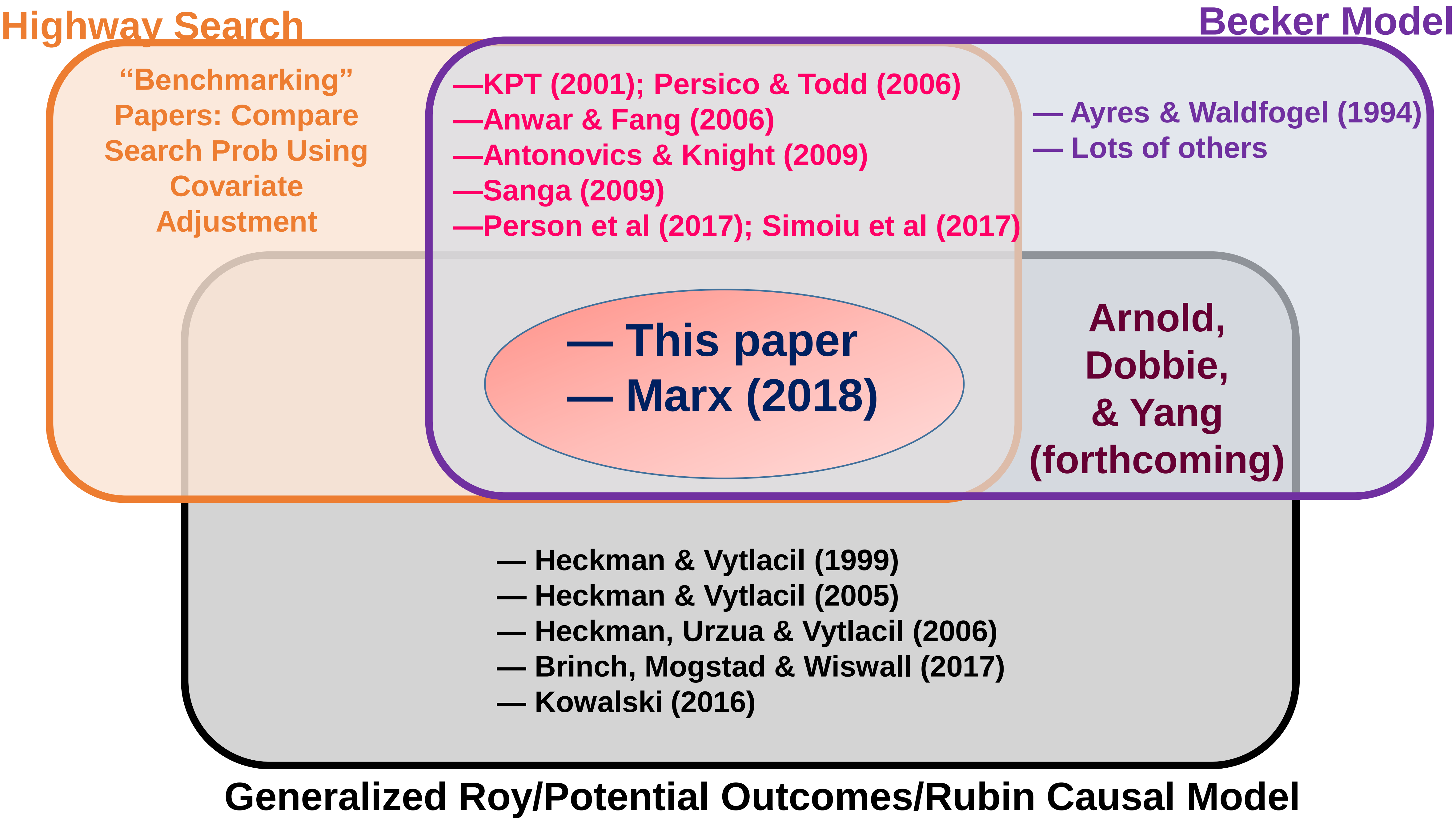
“Benchmarking” Papers: Compare Search Prob Using Covariate Adjustment


Highway Search
Becker Model
"Benchmarking" Papers: Compare Search Prob Using Covariate Adjustment
— Ayres & Waldfogel (1994)
— Lots of others


Highway Search
Becker Model
"Benchmarking" Papers: Compare Search Prob Using Covariate Adjustment
— Ayres & Waldfogel (1994)
— Lots of others
— Heckman & Vytlacil (1999)
— Heckman & Vytlacil (2005)
— Heckman, Urzua & Vytlacil (2006)
— Brinch, Mogstad & Wiswall (2017)
— Kowalski (2016)
Generalized Roy/Potential Outcomes/Rubin Causal Model


Highway Search
Becker Model
“Benchmarking” Papers: Compare Search Prob Using Covariate Adjustment
—KPT (2001); Persico & Todd (2006)
—Anwar & Fang (2006)
—Antonovics & Knight (2009)
—Sanga (2009)
—Person et al (2017); Simoiu et al (2017)
— Ayres & Waldfogel (1994)
— Lots of others
— Heckman & Vytlacil (1999)
— Heckman & Vytlacil (2005)
— Heckman, Urzua & Vytlacil (2006)
— Brinch, Mogstad & Wiswall (2017)
— Kowalski (2016)
Generalized Roy/Potential Outcomes/Rubin Causal Model


Highway Search
Becker Model
"Benchmarking" Papers: Compare Search Prob Using Covariate Adjustment
—KPT (2001); Persico & Todd (2006)
—Anwar & Fang (2006)
—Antonovics & Knight (2009)
—Sanga (2009)
—Person et al (2017); Simoiu et al (2017)
— Ayres & Waldfogel (1994)
— Lots of others
Arnold, Dobbie, & Yang (forthcoming)
— Heckman & Vytlacil (1999)
— Heckman & Vytlacil (2005)
— Heckman, Urzua & Vytlacil (2006)
— Brinch, Mogstad & Wiswall (2017)
— Kowalski (2016)
Generalized Roy/Potential Outcomes/Rubin Causal Model


Highway Search
Becker Model
"Benchmarking" Papers: Compare Search Prob Using Covariate Adjustment
—KPT (2001); Persico & Todd (2006)
—Anwar & Fang (2006)
—Antonovics & Knight (2009)
—Sanga (2009)
—Person et al (2017); Simoiu et al (2017)
— Ayres & Waldfogel (1994)
— Lots of others
— This paper
— Marx (2018)
Arnold, Dobbie, & Yang (forthcoming)
— Heckman & Vytlacil (1999)
— Heckman & Vytlacil (2005)
— Heckman, Urzua & Vytlacil (2006)
— Brinch, Mogstad & Wiswall (2017)
— Kowalski (2016)
Generalized Roy/Potential Outcomes/Rubin Causal Model

The Model

Highway stops by officer *j* occur exogenously

Highway stops by officer j occur exogenously

Produce driver with misconduct probability M

Highway stops by officer j occur exogenously

Produce driver with misconduct probability M

Officer has search distaste δ

Search occurs if $M > \delta$

Highway stops by officer j occur exogenously

Produce driver with misconduct probability M

If race is r & driver/stop characteristics $X = x$:

Search occurs if $M > \delta_{r,j}(x)$

Measure of stop-level discrimination:

$$\Delta_j(\ r_2, r_1) \equiv \delta_{r_2,j} - \delta_{r_1,j}$$

Same officer

Measure of stop-level discrimination:

Same driver/stop characteristics

$$\Delta_j(x; r_2, r_1) \equiv \delta_{r_2,j}(x) - \delta_{r_1,j}(x)$$

Same officer

Measure of stop-level discrimination:

$$\Delta_j(x; r_2, r_1) \equiv \delta_{r_2,j}(x) - \delta_{r_1,j}(x)$$

We can average this over X, over j, or both

So the ideal object of estimation is

$$\delta_{r,j}(x)$$

Assume

$$\boldsymbol{\delta_{r,j}(x) = \delta_r(x, \tilde{z}_j)},$$

where $\tilde{z}_j$ is an instrumental variable.

Assume

$$\delta_{r,j}(x) = \delta_r\left(x, \tilde{z}_j\right),$$

For stop i, use

$$\tilde{z}_j \equiv \frac{1}{N_j - 1} \sum_{k \neq i} D_k$$

Driver misconduct probability distribution

$$\boldsymbol{P(M \leq m) \quad = \quad F_M(m)}$$

Driver misconduct probability distribution

Allowing differences by r and X:

Driver misconduct probability distribution

Allowing differences by r and X:

$$P(M \leq m | r, X = x) = F_{M|X,r}(m | r, X = x)$$

Driver misconduct probability distribution

I will often drop the r subscript and focus on

$$P(M \leq m|X = x) = \boxed{F_{M|X}(m|X = x)}$$

Driver misconduct probability distribution

I will often drop the r subscript and focus on

$$P(M \leq m|X = x) = \boxed{F_{M|X}(m|X = x)}$$

1. Notice that $\tilde{z}$ does not appear.

Driver misconduct probability distribution

I will often drop the r subscript and focus on

$$P(M \leq m|X = x) = \boxed{F_{M|X}(m|X = x)}$$

2. KPT: This is degenerate—all mass at $\mathbf{m} = \bar{h}$

Driver misconduct probability distribution

I will often drop the r subscript and focus on

$$P(M \leq m|X = x) = \boxed{F_{M|X}(m|X = x)}$$

Time to define search & hit rates…

The search rate

$$s(\delta) = 1 - F_M(\delta)$$

The search rate and the hit rate

$$s(\delta) = 1 - F_M(\delta)$$

$$h(\delta) = \frac{\int_{\delta}^{1} m f_m(m)\,dm}{s(\delta)}$$

The unconditional hit rate

$$\int_{\delta}^{1} m f_m(m)\,dm$$

$$v(\delta) \quad = \quad s(\delta)h(\delta)$$

Two key facts

1. The hit rate and search rate are negatively related
2. The slope of the unconditional hit rate identifies δ

$$\frac{dw}{d\delta} = -\delta f(\delta), \quad \frac{ds}{d\delta} = -f(\delta) \Rightarrow \frac{dw}{ds} = \delta$$

Two key facts

1. The hit rate and search rate are negatively related

2. The slope of the unconditional hit rate identifies δ:

$$\frac{dh}{d\delta} = -\delta f(\delta) \qquad \frac{ds}{d\delta} = -f(\delta) \quad \Rightarrow \quad \frac{dh}{ds} = \delta$$

Two key facts

1. The hit rate and search rate are negatively related

2. The slope of the unconditional hit rate identifies δ:

$$\frac{dv}{d\delta} = -\delta f(\delta)$$

Two key facts

1. The hit rate and search rate are negatively related

2. The slope of the unconditional hit rate identifies δ:

$$\frac{dv}{d\delta} = -\delta f(\delta)$$

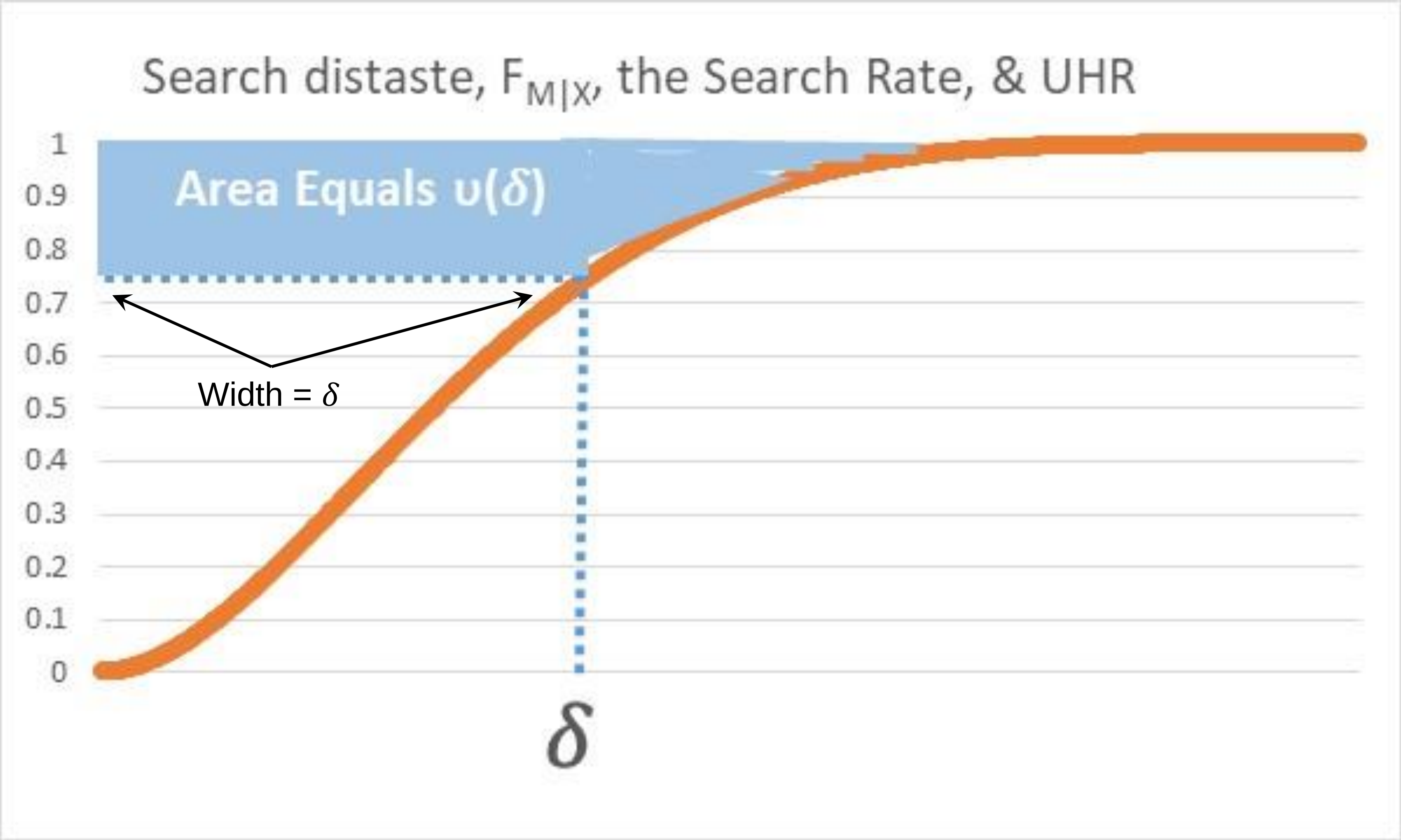
$$\frac{ds}{d\delta} = -f(\delta)$$

Two key facts

1. The hit rate and search rate are negatively related

2. The slope of the unconditional hit rate identifies δ:

$$\begin{aligned} \frac{dv}{d\delta} &= -\delta f(\delta) \\ \frac{ds}{d\delta} &= -f(\delta) \end{aligned} \Rightarrow \frac{dv}{ds} = \delta$$



Search distaste, $F_{M|X}$, the Search Rate, & UHR
Area Equals υ(δ)
Width = δ
δ
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Two key facts

1. The hit rate and search rate are negatively related

2. The slope of the unconditional hit rate identifies δ:

$$\frac{dv}{d\delta} = -\delta f(\delta), \quad \frac{ds}{d\delta} = -f(\delta) \Rightarrow \frac{dv}{ds} = \delta$$

This is empirically useful: $v(s)$ is E[Y|s]

Data

Florida data — From Anwar & Fang (2006)

— 906k stops from 2001

— 9k searches

— Data include lots of covariates

* Driver gender; time of day; location*
* Officer race & gender
* Out-of-state tags, number of passengers

Harris County data

— **From Stanford Open Policing Project**

— **600k stops from 2006-2015**

— **Also about 9k searches**

— **Fewer covariates available**

* **Driver gender; time of day; f.g. location**

Specification test: Are h & s negatively related?

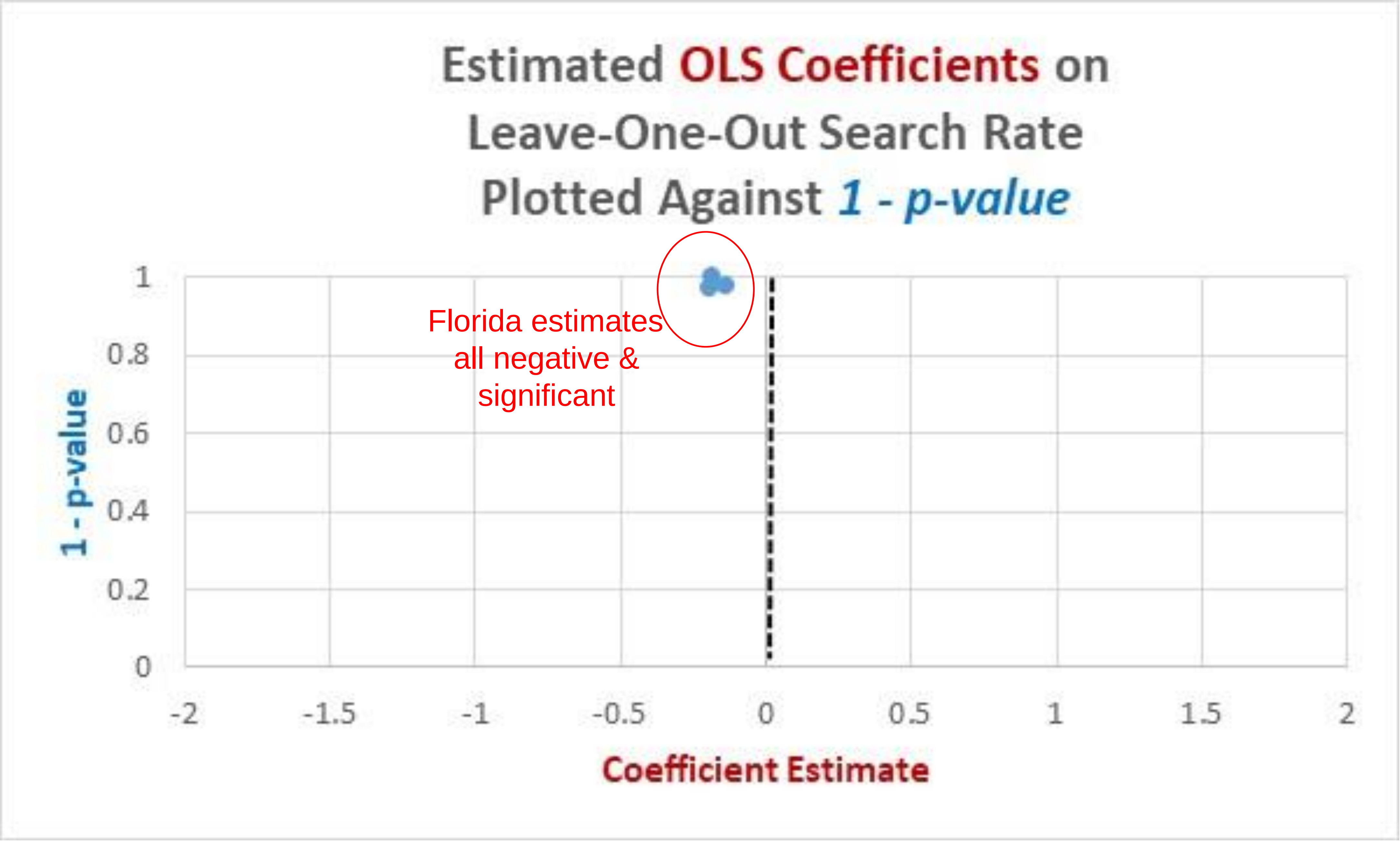
— Dependent variable is conditional hit dummy

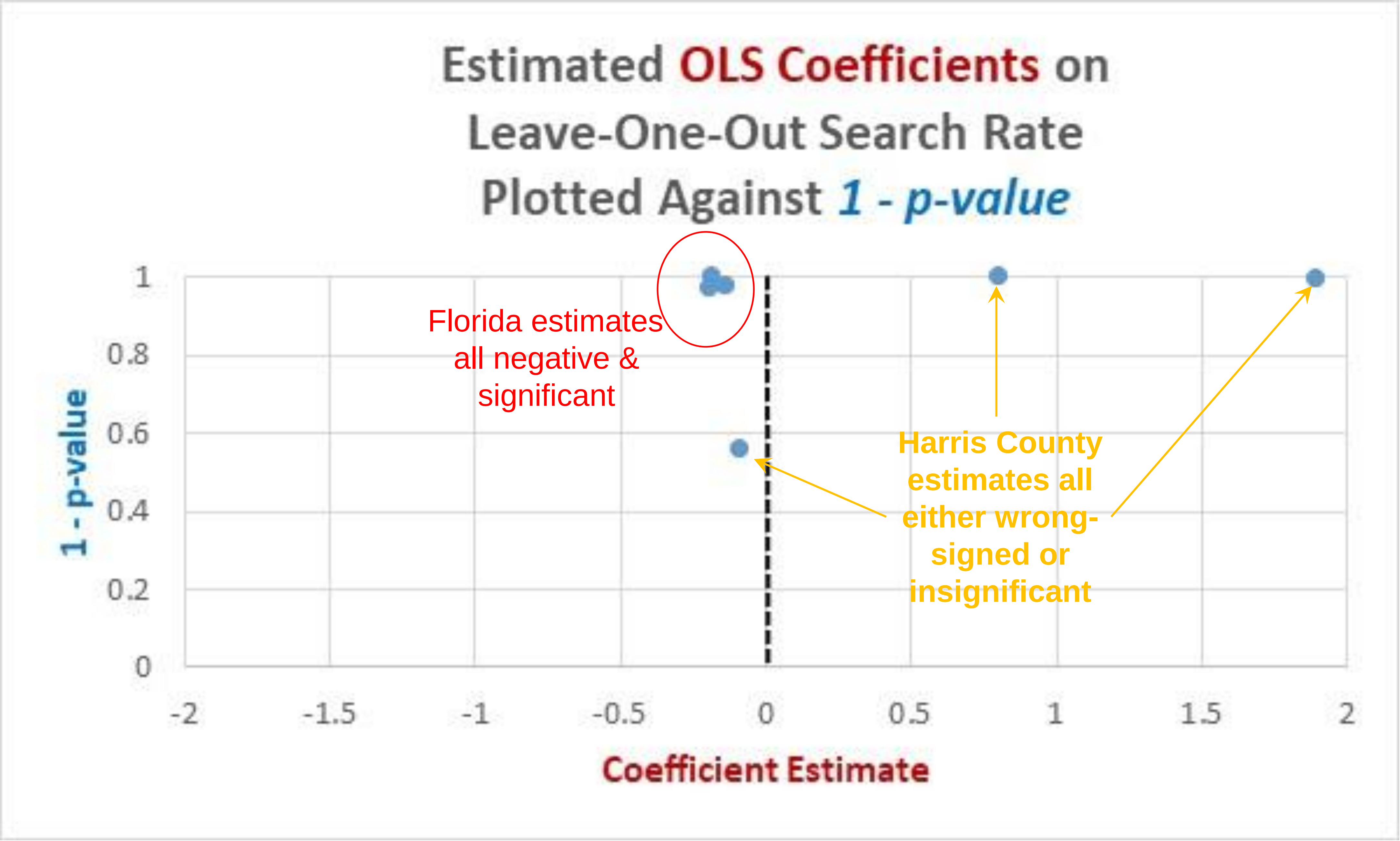
— OLS coefficient on officer search rate

* Calculated using leave-one-out approach
* Search rate among others stopped by j

(other covariates included, too)

Specification Test Results


Estimated OLS Coefficients on
Leave-One-Out Search Rate
Plotted Against 1 - p-value
Florida estimates
all negative &
significant
1 - p-value
1
0.8
0.6
0.4
0.2
0
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
Coefficient Estimate



Estimated OLS Coefficients on Leave-One-Out Search Rate Plotted Against 1 - p-value
1 - p-value
1
0.8
0.6
0.4
0.2
0
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
Coefficient Estimate
Florida estimates all negative & significant
Harris County estimates all either wrong-signed or insignificant

Take-home points

1. **Florida data support non-degenerate $F_{M|X}$**

Take-home points

1. Florida data support non-degenerate $F_{M|X}$

2. Harris County data do <u>not</u>

   —But KPT model also not supported
   —T5 reports several significant coefficients
   —Maybe I lack enough X for Harris County?

Take-home points

1. Florida data support non-degenerate $F_{M|X}$

2. Harris County data do <u>not</u>

   —But KPT model also not supported
   —T5 reports several significant coefficients
   —Maybe I lack enough X for Harris County?

3. No support for estimating my model in Harris

Identification & Estimation

Identification of δ involves two key equations

$$P(D = 1 | x, \tilde{z}_j) \quad = \quad 1 - F_{M|X}\left(\delta\left(x, \tilde{z}_j\right)\right)$$

Propensity-score equation: At least identifies s

Identification of δ involves two key equations

$$P(D = 1|x, \tilde{z}_j) = 1 - F_{M|X}\left(\delta(x, \tilde{z}_j)\right)$$

$$\frac{dv}{ds} = \delta(x, \tilde{z}_j)$$

Slope of UHR

Object of interest

Identification of δ involves two key equations

$$P(D = 1|x, \tilde{z}_j) = 1 - F_{M|X}\left(\delta(x, \tilde{z}_j)\right)$$

$$\frac{dv}{ds} = \frac{dE[Y = 1|x, P = s]}{ds} = \delta(x, \tilde{z}_j)$$

Slope of UHR

Heckman & Vytlacil's Local IV Parameter

Object of interest

Two Examples In Which δ is Identified

Strategy #1: Identification via nonlinearity

Suppose the probability of search and δ satisfy

$$F_{M|X}(\delta|X = x) = F_0($$

Strategy #1: Identification via nonlinearity

Suppose the probability of search and δ satisfy

$$F_{M|X}(\delta|X = x) = F_0(F_0^{-1}(\delta))$$

δ-quantile of x-normalized misconduct distribution

Strategy #1: Identification via nonlinearity

Suppose the probability of search and δ satisfy

$$F_{M|X}(\delta|X=x) = F_0\left(F_0^{-1}(\delta) - x\alpha_0\right)$$

δ-quantile of x-normalized misconduct distribution

Location shift

Strategy #1: Identification via nonlinearity

Suppose the probability of search and δ satisfy

$$F_{M|X}(\delta|X = x) = F_0(F_0^{-1}(\delta) - x\alpha_0)$$

$$\delta(x, \tilde{z}) = g(x\alpha_1 + \tilde{z}\gamma).$$

Strategy #1: Identification via nonlinearity

Now suppose $F_0 = \Lambda$ (logistic) and $g = \Phi$

Strategy #1: Identification via nonlinearity

Now suppose $F_0 = \Lambda$ (logistic) and $g = \Phi$

Propensity score equation:

$$P(D = 1|x, \tilde{z}) = 1 - \Lambda\left(\ln\left(\frac{\Phi(\mathrm{x}\alpha_1+\tilde{z}\gamma)}{1-\Phi(\mathrm{x}\alpha_1+\tilde{z}\gamma)}\right) - x\alpha_0\right)$$

$$\overbrace{1 - \Lambda\left(\qquad\qquad\qquad\qquad \right)}^{F_{M|X}(\qquad |X=x)}$$

$$1 - \Lambda\left(\overbrace{\qquad\qquad\qquad\qquad\qquad\qquad}^{F_{M|X}(\delta(x,\tilde{z})|X=x)}\right)$$

$$1-\Lambda\left(\underbrace{\ln\left(\frac{\delta(x,\tilde{z})}{1-\delta(x,\tilde{z})}\right)}_{\mathrm{logit}(\delta(x,\tilde{z}))}\right)$$

$$1 - \Lambda\left(\underbrace{\ln\left(\frac{\Phi(\mathbf{x}\alpha_1 + \tilde{\mathbf{z}}\gamma)}{1 - \Phi(\mathbf{x}\alpha_1 + \tilde{\mathbf{z}}\gamma)}\right)}_{\text{logit}(\delta(x,\tilde{z}))}\right)$$

$$1 - \Lambda\left(\underbrace{\ln\left(\frac{\Phi(\mathrm{x}\alpha_1 + \tilde{z}\gamma)}{1 - \Phi(\mathrm{x}\alpha_1 + \tilde{z}\gamma)}\right)}_{\text{logit}(\delta(x,\tilde{z}))} \underbrace{-x\alpha_0}_{\text{Location shift}}\right)$$

Strategy #1: Identification via nonlinearity

Now suppose $F_0 = \Lambda$ (logistic) and $g = \Phi$

Propensity score equation:

$$P(D = 1|x, \tilde{z}) = 1 - \Lambda\left(\ln\left(\frac{\Phi(x\alpha_1 + \tilde{z}\gamma)}{1 - \Phi(x\alpha_1 + \tilde{z}\gamma)}\right) - x\alpha_0\right)$$

Nonlinearity can be enough to identify α_1 & γ

Strategy #2: Identification/estimation via IV

Now suppose $F_0 = g$

Propensity score equation:

$$P(D = 1|x, \tilde{z}) = 1 - F_0(x(\alpha_1 - \alpha_0) + \tilde{z}\gamma)$$

Can't distinguish α_1 & α_0 using only P-score eq

Strategy #2: Identification/estimation via IV

Now suppose $F_0 = g$

Unconditional hit rate slope equation:

$$\frac{dE[Y|x,P(Z)=s]}{ds} = \mathbf{F_0}\left(\mathbf{x}\alpha_0 + \mathbf{F_0^{-1}}(\mathbf{1} - \mathbf{s})\right)$$

Strategy #2: Identification/estimation via IV

Now suppose $F_0 = g$

Unconditional hit rate slope equation:

$$\frac{dE[Y|x, P(Z)=s]}{ds} = \mathbf{F_0}\left(\mathbf{x}\alpha_0 + \mathbf{F_0^{-1}}(\mathbf{1} - \mathbf{s})\right)$$

Now we integrate…

Strategy #2: Identification/estimation via IV

Now suppose $F_0 = g$

Unconditional hit rate slope equation:

$$E[Y|x, P = s] = \mathbf{x}\alpha_2 + \int F_0\left(\mathbf{x}\alpha_0 + \mathbf{F_0^{-1}}(\mathbf{1} - \mathbf{s})\right)\mathbf{ds}$$

Note the presence of both x and s inside the $\int$

Strategy #2: Identification/estimation via IV

Now suppose $F_0 = g$ is linear.

Then:

$$E[Y|x, P(Z) = s] = \mathrm{x}\alpha_2 + (s\mathrm{x})\alpha_0 + Q(1 - s)$$

Strategy #2: Identification/estimation via IV

Now suppose $F_0 = g$ is linear.

$$\boldsymbol{E[Y|x, P(Z) = s] = x\alpha_2 + (sx)\alpha_0 + Q(1 - s)}$$

We can do

—Local regression: HUV (2006)

—Global poly: Kowalski (2016); Brinch et al (2017)

Generalized Roy Model Representation

Connecting to the Generalized Roy Model

Let Y_d be "found with contraband" if $D = d$

Connecting to the Generalized Roy Model

Let Y_d be "found with contraband" if $D = d$

Here is a GRM representation:

$$Y_0 = 0$$

$$Y_1 = 1[1 - U_1 \geq 0]$$

$$D = 1[U_D \leq \mu_D(Z)],\ U_D \sim \text{Unif}(0,1)$$

$$U_1 = F_{M|X}(U_1^*|X = x) + U_D,\ U_1^* \sim \text{Unif}(0,1)$$

$$\tilde{Z} \perp U_D|X,$$

2 More Specification Tests

(1) Leveraging the P-score equation

— $s = 1 - F_M(\delta) \Rightarrow ds = -f_M(\delta)d\delta < 0$

—So δ decreasing in search prob.

(1) Leveraging the P-score equation

—$s = 1 - F_M(\delta) \Rightarrow ds = -f_M(\delta)d\delta < 0$

—So δ decreasing in search prob.

(2) Leveraging inframarginality

—$E[Y|D = 1] \geq E[\delta]$

—So hit rate greater than average δ

Table 1: Means for variables used in analysis (Anwar & Fang Florida data)
(generating time: Mon Jul 23 11:49:05 2018 from file *summary-stats.ara*, table 1.)

	Full sample	Black	Hispanic	White
Search rate				
Search was conducted	0.010	0.013	0.013	0.008
Hit rates				
Among only those searched	0.210	0.209	0.115	0.251

What to Take Home

1. **Model identifies taste parameters under some restrictions**

What to Take Home

1. Model identifies taste parameters under some restrictions

2. Also makes testable predictions, which aren’t rejected in FL

What to Take Home

1. Model identifies taste parameters under some restrictions

2. Also makes testable predictions, which aren’t rejected in FL

3. Estimation is feasible

What to Take Home

1. Model identifies taste parameters under some restrictions
2. Also makes testable predictions, which aren't rejected in FL
3. Estimation is feasible