A Nonparametric Finite Mixture Approach to Difference-in-Difference Estimation, with an Application to Professional Training and Wages

Oliver Cassagneau-Francis[1] Robert Gary-Bobo[2] Julie Pernaudet[3]
Jean-Marc Robin[4]

[1]Sciences Po, [2]CREST, ENSAE, [3]University of Chicago, [4]Sciences Po & UCL

February 2020

1. Introduction

What we are doing in this paper

1. We develop a finite-mixture framework for nonparametric difference-in-difference analysis with
    1. unobserved heterogeneity correlating treatment and outcome,
    2. an instrumental variable for the treatment,
    3. no common trend restriction,
    4. Markovian outcome.
2. We apply this framework to an evaluation of the effect of on-the-job/professional (re)training on wages.

Literature

Parallel trends conditional on observed covariates

- Matching: Heckman et al. (1997, 1998), Smith & Todd (2005)
- Nonlinear diff-in-diff: Athey & Imbens (2006), Bonhomme & Sanders (2011), Callaway & Tong (2019)
- Semiparametric: Abadie (2005)
  - Recent work: Li & Li (2019), Sant'Anna & Zhao (2018), Zimmert (2018)
- Empirical likelihood: Qin & Zhang (2008)
- Multiple periods: de Chaisemartin & D'Haultfoeuille (2017), Callaway & Sant'Anna (2019)
- Hansen, Shapiro, Fredholm (2018)

Theoretical contribution

- Replace parallel trends by instrument
- Nonparametric identification proof.
  - Builds on finite mixture models: Hall & Zhou (2003), Hu (2008), Henry et al. (2014), Levine et al. (2011), Kasahara & Shimotsu (2009), Hu & Schennach (2008), Shiu & Hu (2013), Hu and Shum (2012), Sasaki (2015), Bonhomme, Jochmans, Robin (2016a,b, 2017)

Empirical application

- Panel of workers covering three years, 2013-15, for whom we observe the following variables.
- **Treatment**: occurrence of training in 2014; $D_i = 1, 0$ if trained/untrained
- **Instrument**: training advertisement by the employer; $z_i = 1$ if the worker reports receiving information through any of the following channels: hierarchy, training or HR manager, coworkers, or staff representatives
- **Outcome**: log wages w_{it}, $t = 2013, 14, 15$ before and after the treatment.

2. The model

- Identification
- Treatment effects

Model

- Workers can be clustered into K different groups: $k \in \{1, ..., K\}$.
- $\pi(k, z, d)$ is the joint probability of type k, a binary instrument $z \in \{0, 1\}$, and treatment $d \in \{0, 1, ...\}$ (possibly multivalued).
- $f_1(w_1|k)$ is the distribution of pre-treatment outcome w_1 in $t = 1$ given type k. Independent of both treatment and instrument.
- $f_{2|1}(w_2|w_1, k, d)$ and $f_{3|2}(w_3|w_2, k, d)$ are the distributions of outcome w_t given w_{t-1} in $t = 2, 3$ given type k and treatment d.
  - One single post-treatment outcome observation is sufficient if wages are iid given heterogeneity and treatment.
  - Two for first-order Markov
  - Note the non stationarity.

Roy model

- Possible rationale: Roy model (Heckman and Vytlacil (2005); Carneiro et al. (2010, 2011)):

$$y = y(k,0) + [y(k,1) - y(k,0)]\, D$$
$$D = 1 \text{ if } \mathsf{E}[y(1) - y(0)|k] \geq c(k,z),$$

  where
  - k is individual heterogeneity (different social backgrounds, as measured/influenced by controls variables such as education, gender, etc, produce different social types $k = 1, ..., K$)
  - z is the instrument, ie an environmental variable affecting treatment decision (eg training offer or information)
  - $y(k,0), y(k,1)$ are treatment-specific outcome variables (random given k and independent of z)
  - $c(k,z)$ is training cost (random given k, z)
- Difference-in-difference version: condition on pre-treatment wage.
- Important difference with Heckman & Vytlacil: k and z may be correlated.

2.1. Identification

Complete likelihood

- Probability of instrument z, treatment d, and three wages w_1, w_2, w_3:

$$p(z,d,w_1,w_2,w_3) = \sum_k \pi(k,z,d)\, f_1(w_1|k)\, f_{2|1}(w_2|w_1,k,d)\, f_{3|2}(w_3|w_2,k,d)$$

$$= \sum_k \pi(k,z,d)\, \frac{f_1(w_1|k)\, f_{2|1}(w_2|w_1,k,d)}{f_2(w_2|k,d)}\, f_2(w_2|k,d)\, f_{3|2}(w_3|w_2,k,d)$$

$$= \sum_k \pi(k,z,d)\, f_{1|2}(w_1|w_2,k,d)\, f_2(w_2|k,d)\, f_{3|2}(w_3|w_2,k,d)$$

- Where

$$f_2(w_2|k,d) = \int f_1(w_1|k)\, f_{2|1}(w_2|w_1,k,d)\, dw_1$$

and

$$f_{1|2}(w_1|w_2,k,d) = \frac{f_1(w_1|k)\, f_{2|1}(w_2|w_1,k,d)}{f_2(w_2|k,d)}$$

Matrix notation

$$p(z, d, w_1, w_2, w_3) = \sum_k \left[f_{1|2}(w_1|w_2, k, d)\right] \left[\pi(k, z, d)\, f_2(w_2|k, d)\right] \left[f_{3|2}(w_3|w_2, k, d)\right]$$

- Assume discrete wages (N points) and construct the matrices

$$\underset{N\times N}{P(z, d, w_2)} = \left[p(z, d, w_1, w_2, w_3)\right]_{w_1 \times w_3}$$

and

$$\underset{N\times K}{F_1(d, w_2)} = \left[f_{1|2}(w_1|w_2, k, d)\right]_{w_1 \times k} \qquad \underset{N\times K}{F_2(d, w_2)} = \left[f_{3|2}(w_3|w_2, k, d)\right]_{w_3 \times k}$$

$$\underset{K\times K}{D(z, d, w_2)} = \text{diag}\left[\pi(k, z, d)\, f_2(w_2|k, d)\right]_k$$

- We then have, for all d, w_2,

$$P(z, d, w_2) = F_1(d, w_2)\, D(z, d, w_2)\, F_2(d, w_2)^\top$$

Assumptions

- Social types must produce sufficient variation in treatment decisions and outcomes.
- For all treatment values d,

1. $\pi(k, z, d) \neq 0$: all treatments ($d = 0, 1$) are possible for all k and z
2. $\frac{\pi(k,1,d)}{\pi(k,0,d)} \neq \frac{\pi(k',1,d)}{\pi(k',0,d)}$ for all k, k': sufficient richness of interaction between type and instrument in treatment probabilities
3. $\{f_{t|2}(w_t|w_2, k, d), k = 1, ..., K\}$, $t = 1, 3$, are two linearly independent systems: types create different wages distributions

1. SVD

- Fix (d, w_2) and omit it from $P(z, d, w_2) \equiv P(z)$ for the moment.
- Assumptions 1 and 3 imply that $P(0) = F_1 \, D(0) \, F_2^\top$ has rank K.
- SVD: $P(0) = U \Lambda V^\top$, $U^\top U = I_N$, $V^\top V = I_N$, Λ diagonal
- For simplicity, set $N = K$ (same number of wages than worker types). Assumption 3 implies $N > K$.

2. "Whitening"

- SVD $P(0) = U\Lambda V^\top$ implies that

$$\Lambda^{-1} U^\top P(0) V = I_K$$

$$\iff \underbrace{\Lambda^{-1} U^\top F_1}_{=W \text{ (say)}} \times \underbrace{D(0) F_2^\top V}_{=W^{-1}} = I_K$$

- It follows that, for $z = 1$,

$$\begin{aligned}
\Lambda^{-1} U^\top P(1) V &= \Lambda^{-1} U^\top F_1 D(1) F_2^\top V \\
&= \Lambda^{-1} U^\top F_1 \, D(1) D(0)^{-1} \, D(0) F_2^\top V \\
&= W \, D(1) D(0)^{-1} \, W^{-1}.
\end{aligned}$$

- The instrument creates variation giving algebraic structure to identifying restrictions.

3. Group labels given treatment, across wages w_2

- The diagonal entries of

$$D(1)D(0)^{-1} = \text{diag}\left[\frac{\pi(k,1,d)}{\pi(k,0,d)}\right]_k$$

  are uniquely determined as the eigenvalues of the matrix $\Lambda^{-1}U^\top P(1)V$.
- They are independent of w_2. So, for each d, we can reorder groups consistently across different wages w_2.

4. Diagonalization

- Because eigenvalues are distinct, eigenspaces are unidimensional.
- Yet, eigenvectors are still determined only up to a multiplicative constant. One can show that this indetermination is resolved by the fact that the rows of F_1 sum to one (each column is a probability distribution).
- Hence, $W = \Lambda^{-1} U^\top F_1$ is identified.
- Hence, F_1 is identified.
- We can obtain $D(0)$ and F_2 similarly from W^{-1}.
- Finally, $D(1)$ is identified from $D(1)D(0)^{-1}$.

5. Densities

- $D(z, d, w_2) = \text{diag}\left[\pi(k, z, d)\, f_2(w_2|k, d)\right]_k$
- Summing over w_2 (**only possible because we have aligned labeling across** w_2) identifies $\pi(k, z, d)$.
- Hence $f_2(w_2|k, d)$ is identified.
- Finally, $f_1(w_1|k)$ and $f_{2|1}(w_2|w_1, k, d)$ can be recovered from the joint density

$$f_{1|2}(w_1|w_2, k, d)\, f_2(w_2|k, d) = f_1(w_1|k)\, f_{2|1}(w_2|w_1, k, d)$$

6. Group labels across treatments

- Having identified $f_1(w_1|k)$ for each d, we use that fact that wage distributions in the first period are independent of treatment to align the group labels across treatments.
- This identification argument applies to any number of treatments.

2.2. Treatment effects

ATE, ATT

- Define an outcome variable $y = y(d)$ ($y = w_2$ or w_3).
- ATE:

$$ATE(k) = \mathrm{E}\left[y(1)|k\right]) - \mathrm{E}\left[y(0)|k\right] = \mu(k,1) - \mu(k,0) \quad \text{(say)}$$
$$ATE = \sum_k \pi(k)\, ATE(k)$$

  with $\pi(k) = \sum_{z,d} \pi(k,z,d)$
- ATT:

$$ATT(k) = ATE(k)$$
$$ATT = \sum_{k,z} \pi(k,z|d=1)\, ATE(k)$$

  with $\pi(k,z|d=1) = \pi(k,z,1) \Big/ \sum_{k,z} \pi(k,z,1)$.

OLS

- Regress y on $D = 1$:

$$
\begin{aligned}
b_{OLS} &= \frac{\text{Cov}(y, D)}{\text{Var}(D)} = \text{E}[y(1)|D = 1] - \text{E}[y(0)|D = 0] \\
&= \sum_{k,z} \pi(k, z|d = 1)\, \mu(k, 1) - \sum_{k,z} \pi(k, z|d = 0)\, \mu(k, 0) \\
&= ATT + \sum_{k,z} \left[\pi(k, z|d = 1) - \pi(k, z|d = 0)\right] \mu(k, 0).
\end{aligned}
$$

- The blue term is not signed.

IV

$$b_{IV} = \frac{\text{Cov}(y,z)}{\text{Cov}(D,z)} = \frac{\text{E}(y|z=1) - \text{E}(y|z=0)}{\text{E}(D|z=1) - \text{E}(D|z=0)}$$

- Let
$$\pi(k,d|z) = \frac{\pi(k,z,d)}{\sum_{k,d}\pi(k,z,d)}, \quad \pi(k|z) = \sum_{d}\pi(k,d|z).$$
- Denominator:
$$\text{E}(D|z=1) - \text{E}(D|z=0) = \sum_{k}\left[\pi(k,d=1|z=1) - \pi(k,d=1|z=0)\right].$$
- Monotonicity: $\pi(k,d|z=1) \geq \pi(k,d|z=0)$

IV $\neq$ LATE

- Numerator:

$$\begin{aligned}
\mathsf{E}(y|z=1) - \mathsf{E}(y|z=0) &= \sum_k \left[\sum_d \pi(k,d|z=1)\,\mu(k,d) \right] \\
&\quad - \sum_k \left[\sum_d \pi(k,d|z=0)\,\mu(k,d) \right] \\
&= \sum_k \left[\pi(k,1|z=1) - \pi(k,1|z=0) \right] \mathit{ATE}(k) \\
&\quad + \sum_k \left[\pi(k|z=1) - \pi(k|z=0) \right] \mu(k,0)
\end{aligned}$$

- The blue term does not vanish if k and z are correlated.

3. The data

Data

- Panel of workers covering three years, 2013-15, for whom we observe the following variables.
- **Treatment**: occurrence of training in 2014; $D_i = 1, 0$ if trained/untrained
- **Instrument**: training advertisement by the employer; $z_i = 1$ if the worker reports receiving information through any of the following channels: hierarchy, training or HR manager, coworkers, or staff representatives
- **Outcome**: log wages w_{it}, $t = 2013, 14, 15$ before and after the treatment.

OLS and IV

- Regress log wages in 2013, 2014 and 2015 on treatment, controlling for many individual and employer characteristics

	GLS	(no controls)	3SLS	(no controls)
2013	0.043	0.184	0.086	0.272
	(0.006)	(0.008)	(0.046)	(0.049)
2014	0.048	0.191	0.156	0.326
	(0.006)	(0.008)	(0.047)	(0.050)
2015	0.047	0.189	0.147	0.324
	(0.006)	(0.008)	(0.047)	(0.050)
N	9571	10043	9571	10043

- Instrumentation (and controls) renders effect of treatment on initial wage not significant.

DiD

	FE, OLS	FE, IV	FD, IV
Treatment	0.007	0.053	0.055
	(0.003)	(0.019)	(0.021)
Year 2014	0.028	0.006	0.005
	(0.002)	(0.009)	(0.010)
Year 2015	0.054	0.033	0.032
	(0.002)	(0.009)	(0.010)
N	30129	30129	20086

- DiD significant only when the treatment is instrumented
- Note: no controls here as they are not time-varying

Take away

- Some evidence of endogenous treatment even after exhaustive control
- Standard within-group estimation (DiD) does not work
- What about nonlinear and heterogeneous treatments?

4. Estimation

Estimation procedure

- Wages:

$$w_1 = \mu_1(k) + u_1, \quad u_1 \sim N\left(0, \sigma_1^2(k)\right)$$
$$w_t = \mu_t(k,d) + u_t, \quad u_t \sim N\left(\rho u_{t-1}, \sigma_t^2(k,d)\right), \quad t = 2,3$$

- Given (ρ, K), we use the EM algorithm to estimate the discrete mixture.
  - E-step: calculate posterior probabilities of all individuals' types
  - M-step: 1) estimate μ's and σ's by empirical means and variances weighted by posterior probas; 2) estimate π by averaging posterior probas.
- We arbitrarily label groups by increasing $\mu_1(k)$.

E-step

- Complete individual likelihood:

$$\ell_i(k|\beta) = q(x_i|k, z_i, d_i)\, \pi(k, z_i, d_i)\, f_1(w_{i1}|k)\, f_{2|1}(w_{2i}|w_{1i}, k, d_i)\, f_{3|2}(w_{3i}|w_{2i}, k, d_i)$$

  where $x = (x^1, ..., x^H)$ a vector of control dummy variables (female, low education, manufacturing, etc.) satisfying the conditional independence assumption:

$$q(x_i|k, z_i, d_i) = q_1(x_i^1|k) \times ... \times q_H(x_i^H|k).$$

- For a given value $\beta^{(m)}$ of the parameter, the posterior probability of worker i to be of type k (also called *responsibility*) is

$$p_i^{(m)}(k) \equiv \frac{\ell_i(k|\beta^{(m)})}{\sum_k \ell_i(k|\beta^{(m)})}$$

M-step (1)

Estimate μ's and σ's by empirical means and variances weighted by posterior probas:

$$\mu_1^{(m+1)}(k) = \frac{\sum_i p_i^{(m)}(k) w_{i1}}{\sum_i p_i^{(m)}(k)}, \quad \sigma_1^{(m+1)}(k)^2 = \frac{\sum_i p_i^{(m)}(k) u_{i1}^{(m+1)}(k)^2}{\sum_i p_i^{(m)}(k)}$$

with $u_{i1}^{(m+1)}(k) = w_{i1} - \mu_1^{(m+1)}(k)$, and for $t = 2, 3$

$$\mu_t^{(m+1)}(k,d) = \frac{\sum_i p_i^{(m)}(k) D_{di} \left[w_{it} - \rho u_{i,t-1}^{(m+1)}(k,d) \right]}{\sum_i p_i^{(m)}(k) D_{di}}$$

$$\sigma_t^{(m+1)}(k,d)^2 = \frac{\sum_i p_i^{(m)}(k) D_{di} \left[u_t^{(m+1)}(k,d) - \rho u_{i,t-1}^{(m+1)}(k,d) \right]^2}{\sum_i p_i^{(m)}(k) D_{di}}$$

with $u_{it}^{(m+1)}(k,d) = w_{it} - \mu_t^{(m+1)}(k,d)$.

M-step (2)

Estimate

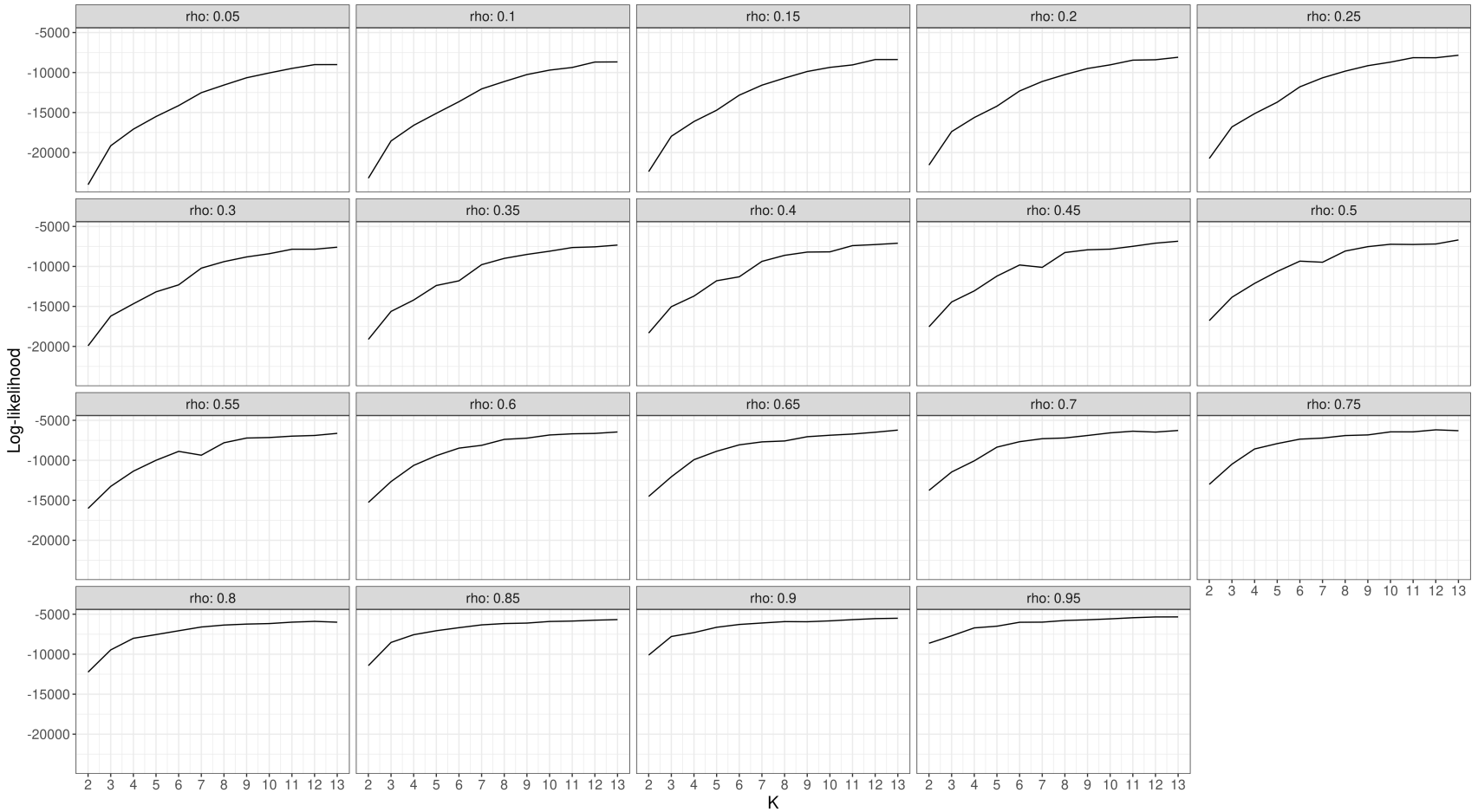
$$\pi^{(m+1)}(k, z_i, d_i) = \frac{1}{N} \sum_{i:z_i=z, d_i=d} p_i^{(m)}(k)$$

and for $h = 1, ..., H$,

$$q_h^{(m+1)} = \sum_{i:x_i^h=1} p_i^{(m)}(k) \Big/ \sum_i p_i^{(m)}(k)$$

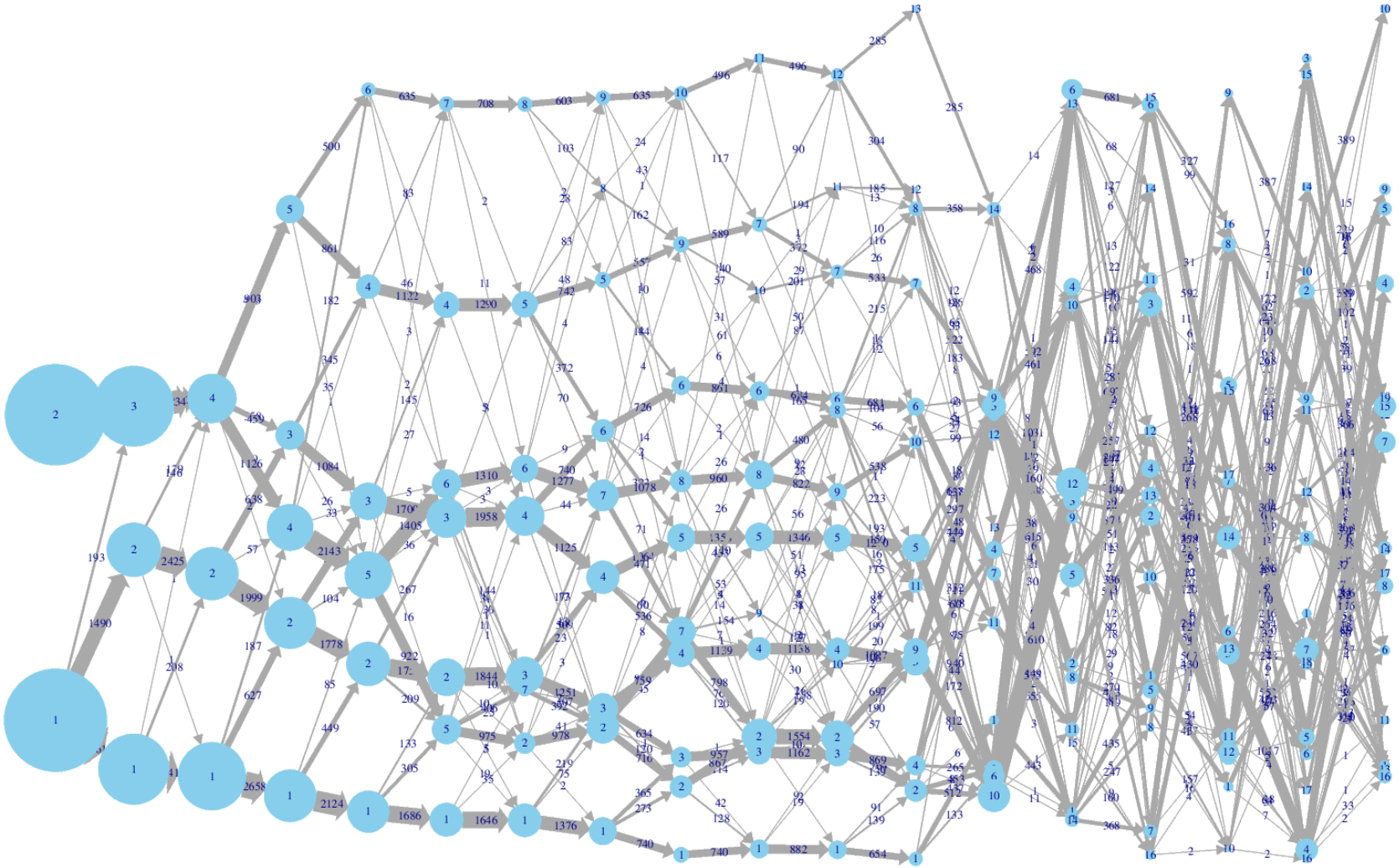
Log-likelihood

- Likelihood increases in ρ and K. However, for greater ρ, smaller K is enough.



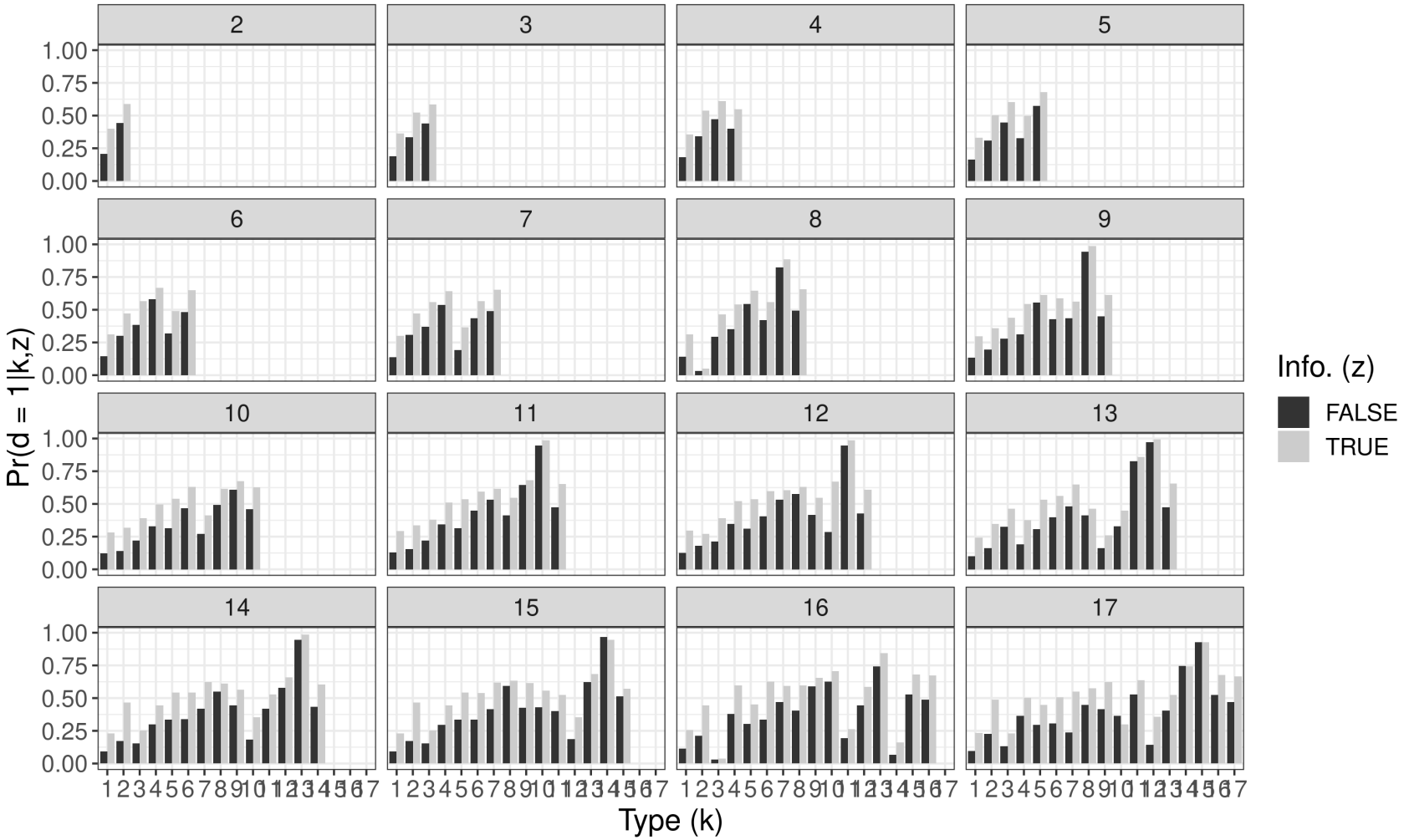
Assignment plot ($\rho = 0.7$)

- Assign most probable type to workers and join Ks
- Messy for $K \geq 14$. Similar graph for different ρ



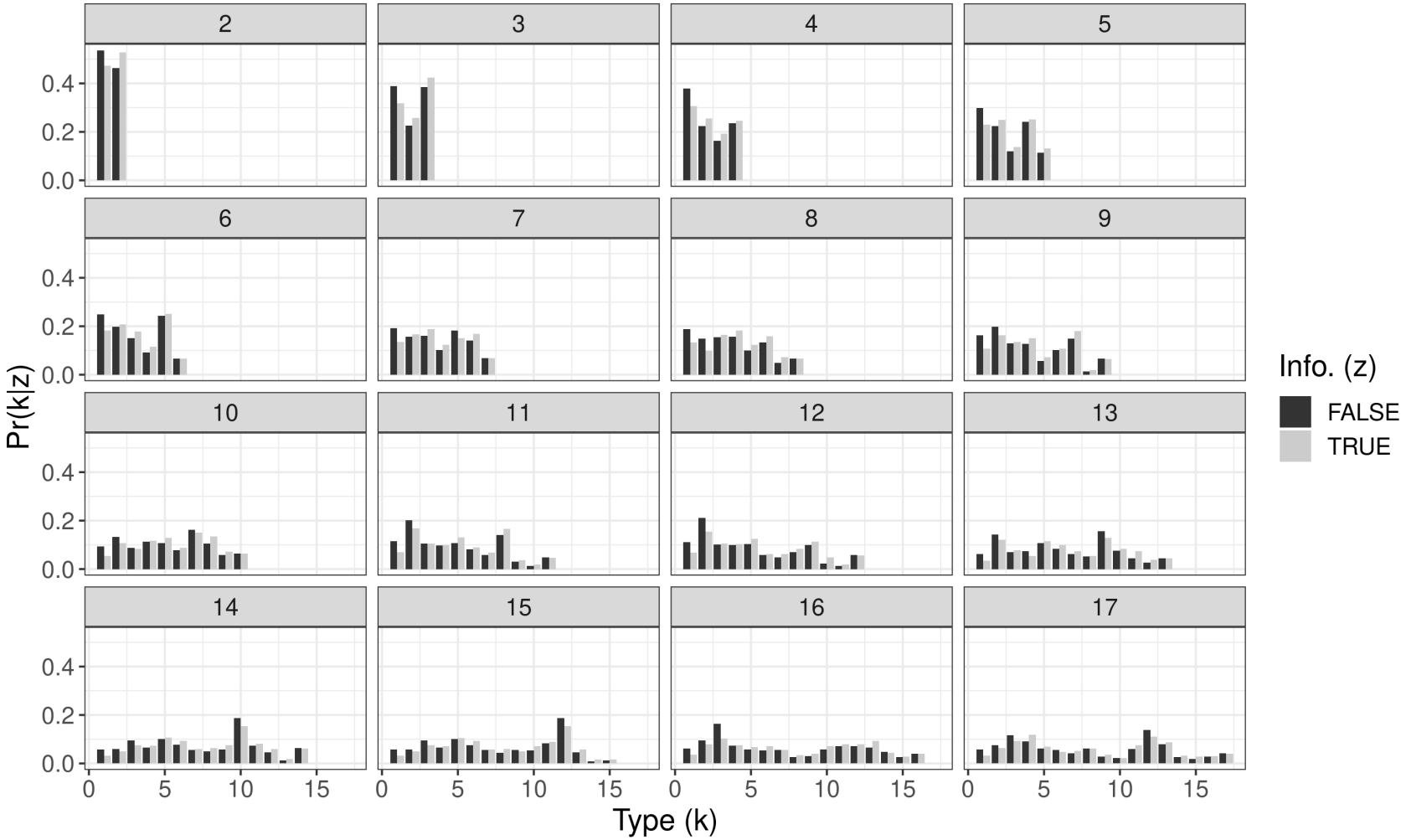
Treatment probability, $\pi(d = 1|k, z)$

- Monotonicity holds
- Good types train more.



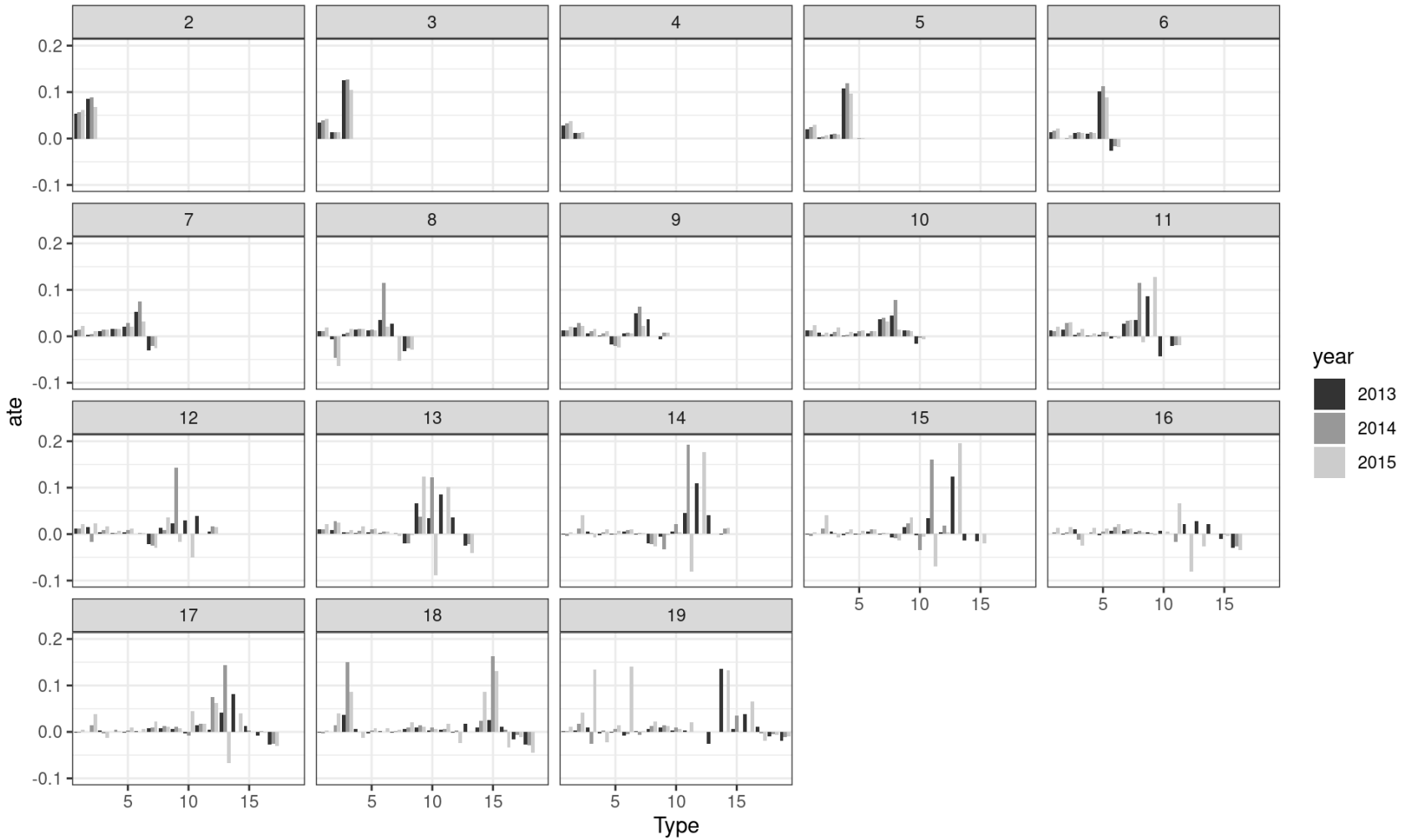
Composition, $\pi(k|z)$

- Weak positive link between k and z (black bars higher than grey at low k: small positive correlations)
- Low k's more often offered training

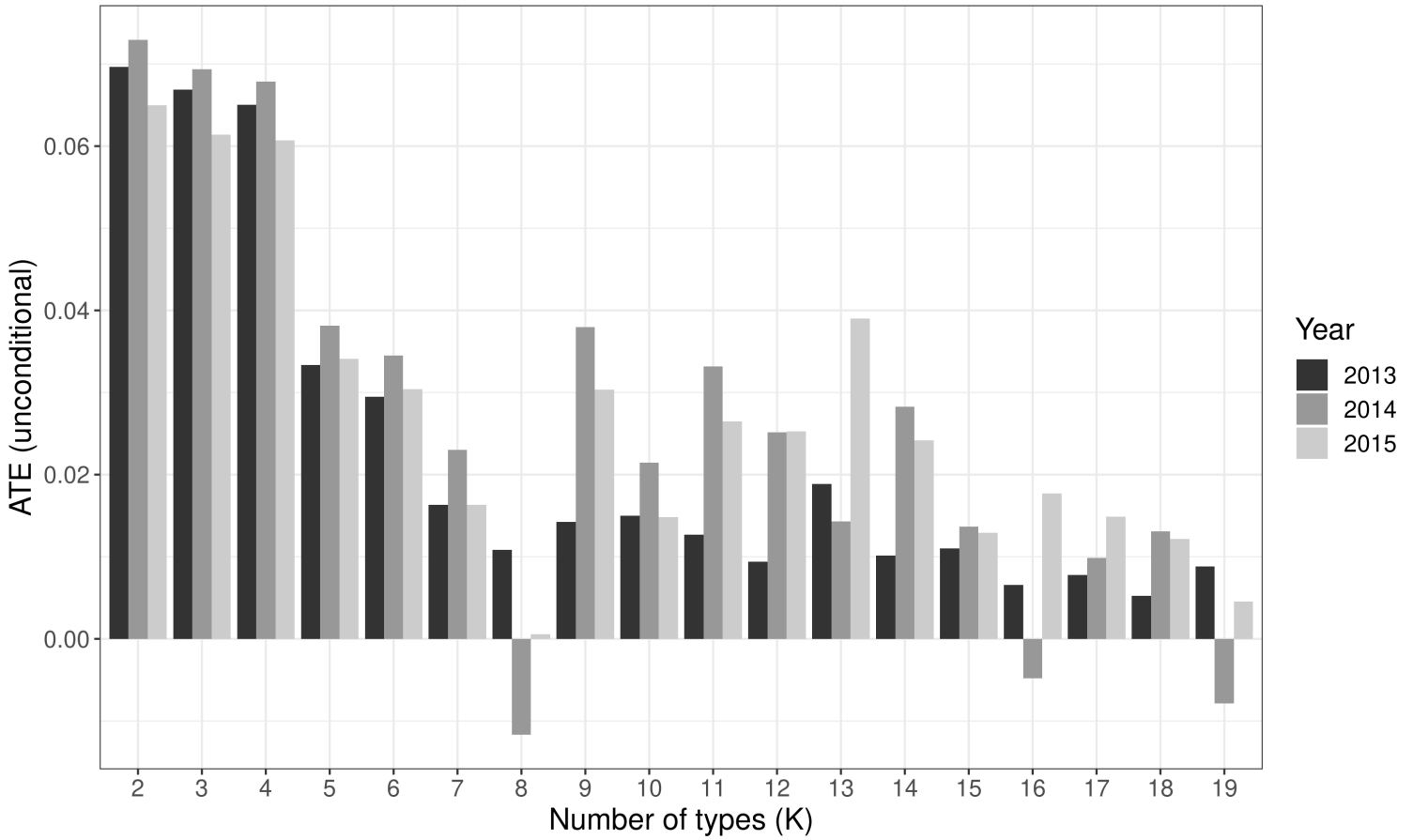


$ATE(k) = \mu(k,1) - \mu(k,0)$

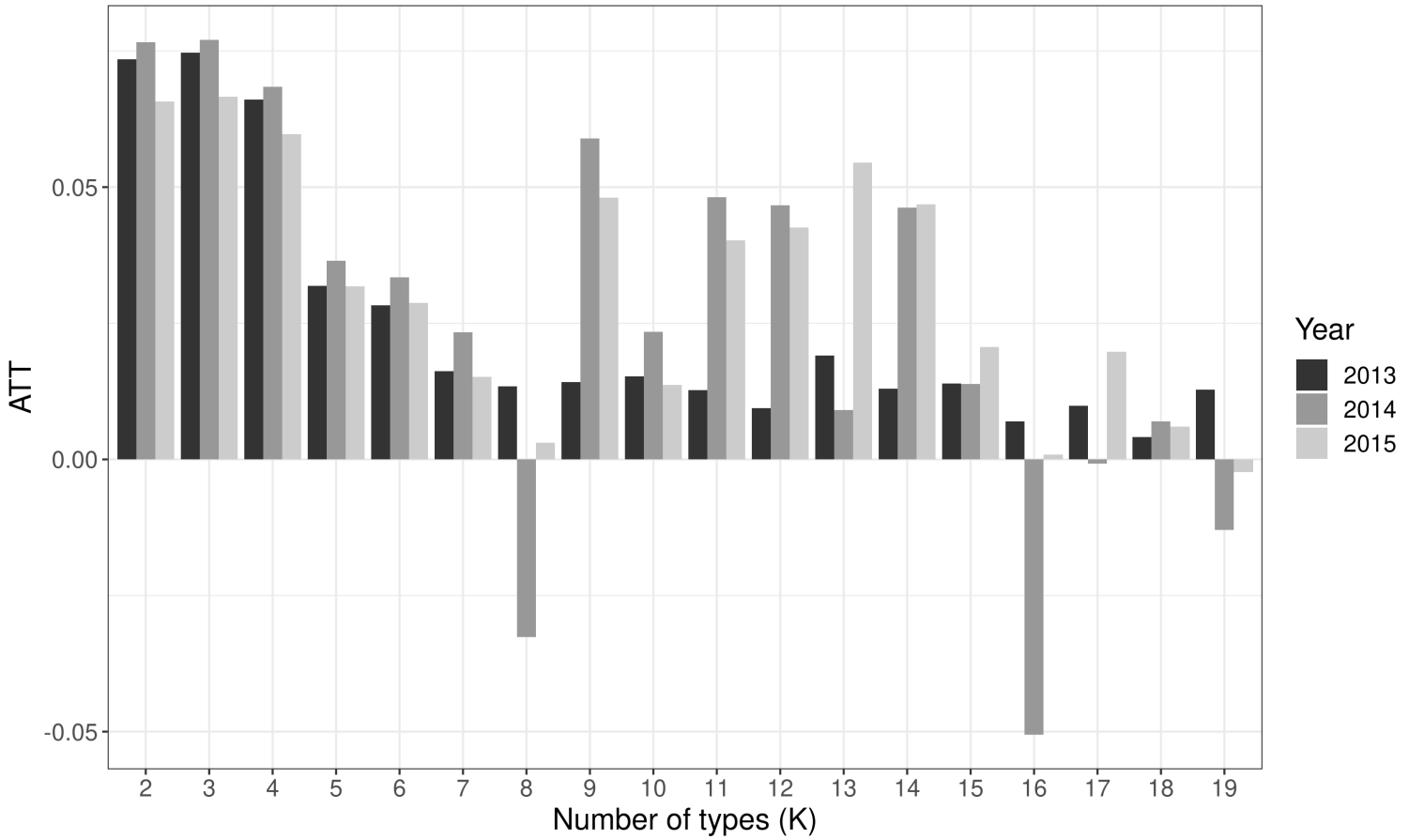
- After estimating the model assuming w_1 independent of d, calculate conditional 2013 means given future treatment. Counterfactual at low K
- $ATE(k)$ higher for high k's



Unconditional *ATE*



Unconditional *ATT*

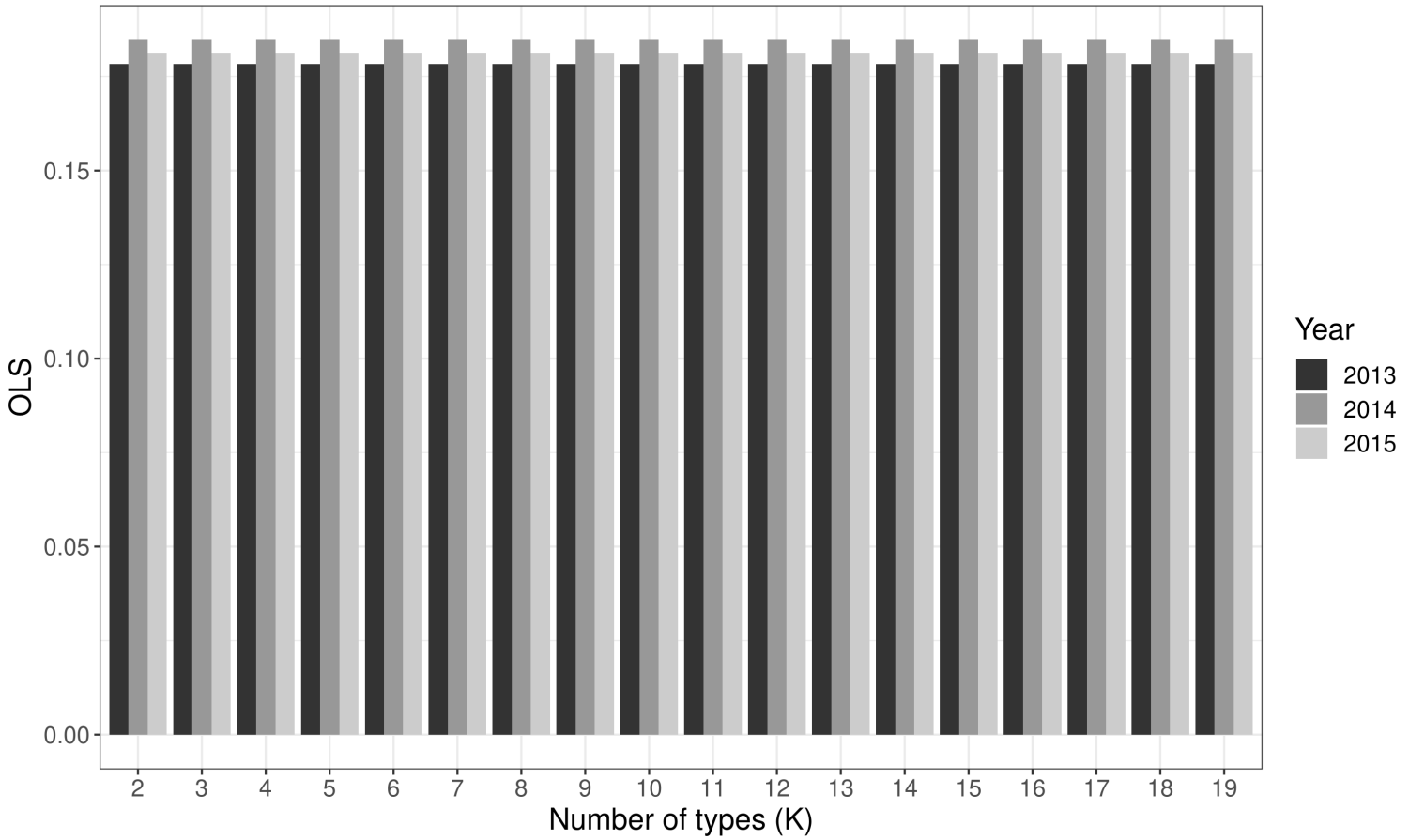


$ATE < ATT$

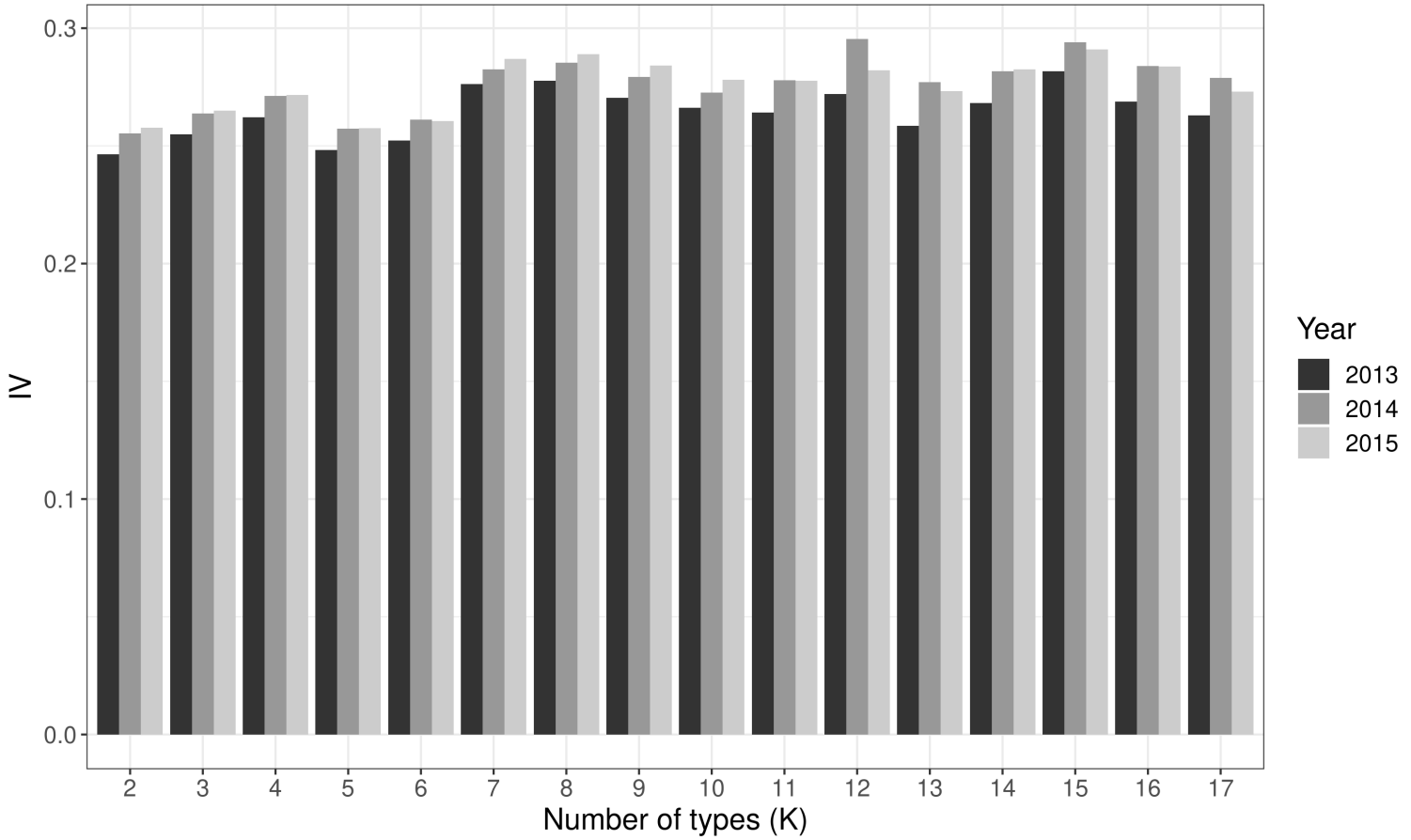
- ATE half of DiD; $ATT > ATE$

	ATE			ATT		
K	2013	2014	2015	2013	2014	2015
2	0.070	0.073	0.065	0.073	0.077	0.066
3	0.067	0.069	0.061	0.075	0.077	0.067
4	0.065	0.068	0.061	0.066	0.068	0.060
5	0.033	0.038	0.034	0.032	0.037	0.032
6	0.029	0.035	0.030	0.028	0.033	0.029
7	0.016	0.023	0.016	0.016	0.023	0.015
8	0.011	-0.012	0.001	0.013	-0.033	0.003
9	0.014	0.038	0.030	0.014	0.059	0.048
10	0.015	0.021	0.015	0.015	0.023	0.014
11	0.013	0.033	0.026	0.013	0.048	0.040
12	0.009	0.025	0.025	0.009	0.047	0.043
13	0.019	0.014	0.039	0.019	0.009	0.055
14	0.010	0.028	0.024	0.013	0.046	0.047
15	0.011	0.014	0.013	0.014	0.014	0.021
16	0.007	-0.005	0.018	0.007	-0.051	0.001
17	0.008	0.010	0.015	0.010	-0.001	0.020
18	0.005	0.013	0.012	0.004	0.007	0.006
19	0.009	-0.008	0.005	0.013	-0.013	-0.002

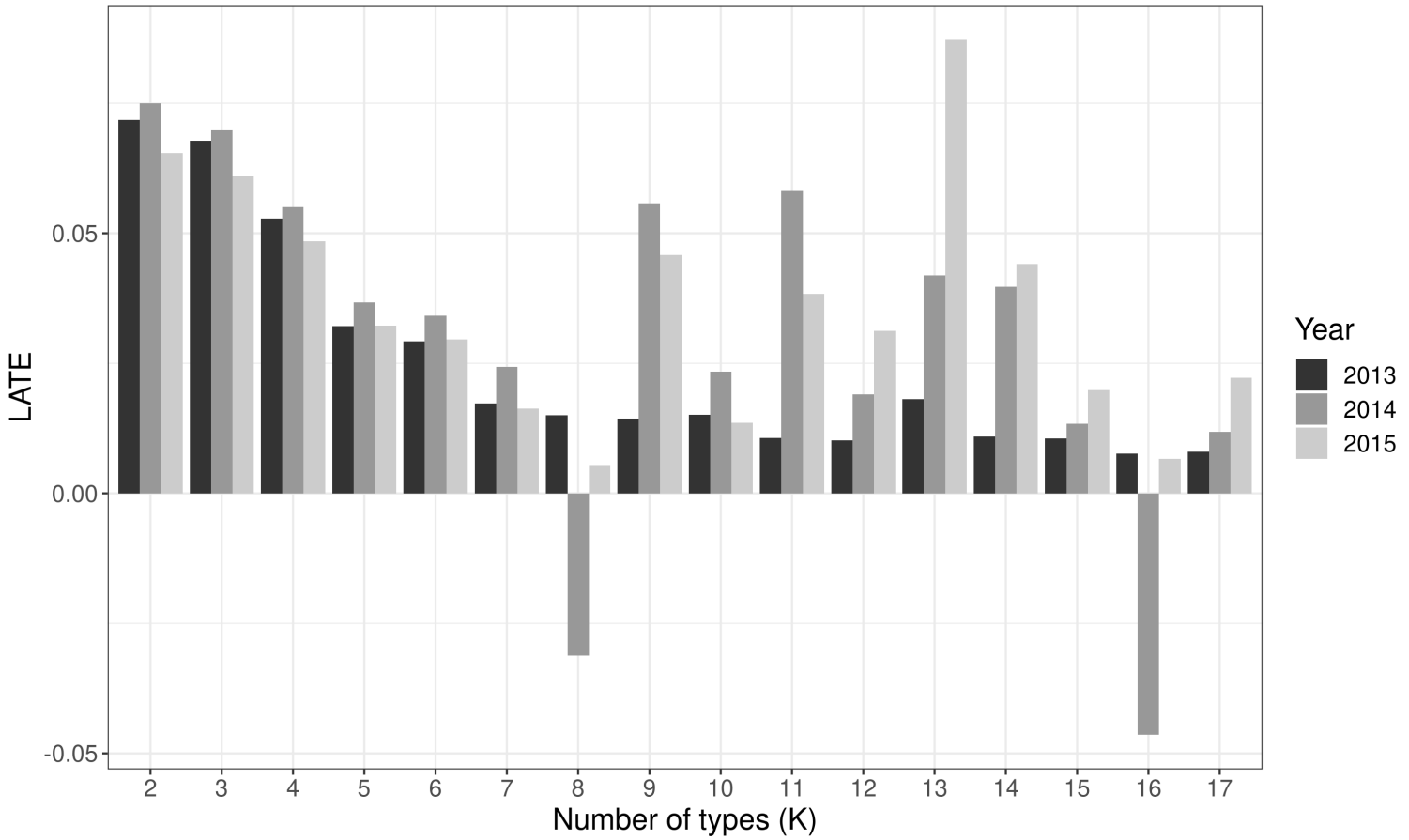
OLS


OLS
0.15
0.10
0.05
0.00
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Number of types (K)
Year
2013
2014
2015

IV


0.3
0.2
0.1
0.0
IV
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Number of types (K)
Year
2013
2014
2015

LATE


LATE
0.05
0.00
-0.05
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Number of types (K)
Year
2013
2014
2015

Observed worker characteristics by type ($K = 7$)

	1	2	3	4	5	6	7
Wage (2013)	9.04	11.59	15.24	21.94	10.48	17.94	31.63
Variance wage	0.82	1.12	1.92	2.74	2.01	7.66	5.99
Full-time	0.80	0.94	0.95	0.96	0.81	0.90	0.96
Open-ended contract	0.93	0.97	0.98	0.99	0.89	0.93	0.98
Unskilled manual	0.47	0.44	0.26	0.05	0.42	0.17	0.01
Skilled manual	0.41	0.21	0.11	0.03	0.32	0.12	0.02
Clerk	0.05	0.15	0.24	0.13	0.09	0.13	0.02
Foreman/Supervisor	0.06	0.16	0.19	0.11	0.13	0.15	0.04
Middle management	0.00	0.01	0.10	0.38	0.01	0.16	0.40
Management	0.01	0.02	0.08	0.25	0.02	0.24	0.43
Less than HS	0.58	0.49	0.35	0.15	0.50	0.26	0.09
HS gen. or voc.	0.23	0.22	0.18	0.13	0.20	0.16	0.09
HS or more	0.17	0.29	0.46	0.72	0.29	0.57	0.81
Partner	0.64	0.74	0.78	0.82	0.69	0.76	0.86
Children	0.47	0.56	0.61	0.65	0.53	0.59	0.69
French	0.94	0.97	0.98	0.98	0.95	0.97	0.97
Female	0.43	0.30	0.24	0.20	0.40	0.31	0.13
Less than 30	0.28	0.16	0.10	0.06	0.27	0.16	0.01
30-40	0.23	0.28	0.28	0.28	0.24	0.28	0.18
40-50	0.28	0.33	0.37	0.38	0.28	0.31	0.38
older than 50	0.21	0.23	0.25	0.29	0.21	0.26	0.43
Health issues (current)	0.12	0.10	0.07	0.04	0.18	0.12	0.02

Observed employer characteristics by type ($K = 7$)

	1	2	3	4	5	6	7
< 50	0.45	0.35	0.25	0.17	0.32	0.22	0.16
50-249	0.25	0.24	0.22	0.18	0.23	0.19	0.19
> 249	0.30	0.41	0.53	0.65	0.45	0.59	0.65
Manufacturing	0.19	0.36	0.41	0.39	0.27	0.29	0.34
Services	0.78	0.60	0.56	0.58	0.69	0.69	0.64
CDD at firm	12.32	9.31	7.24	5.58	10.49	7.59	7.85
Part-time at firm	19.56	9.69	7.52	8.55	16.69	10.43	9.28
Individual incentives	0.51	0.63	0.73	0.80	0.63	0.76	0.80
Collective incentives	0.57	0.72	0.79	0.86	0.70	0.80	0.83
Outsource	0.27	0.34	0.41	0.49	0.33	0.41	0.45
HR department	0.77	0.84	0.89	0.93	0.86	0.91	0.93

5. Conclusion

Summary

- We prove the nonparametric identification of a diff-in-diff model.
- The outcome variable can be Markovian and no parallel trend restriction is required.
- Identification rests on the existence of an instrument determining treatment but not the outcome.
- The estimation procedure uses the EM algorithm.
- We apply the model to an evaluation of on-the-job training on wages.
- ATE is estimated around .025-.03 and ATT around .04-.05.
- ToDo: Estimate a version of the model with unobserved AND observed heterogeneity

References

Carneiro, P., J. J. Heckman, and E. Vytlacil (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*, 78, 377–394.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–2781.

Heckman, J. J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738.