RA work for “Migration, Specialization, and Trade: Evidence from the Brazilian March to the West”

Brian Cevallos Fujiy

May 4, 2020

1 Overview of the models

This is a collection of models related to the paper “Migration, Specialization, and Trade: Evidence from the Brazilian March to the West” version June 2019 by Heitor Pellegrina and Sebastian Sotelo. We have 8 different model specifications. Each model first simulates data from Matlab script files, and then run regressions with Stata do-files. All models will require to previously run “masterfile.do”, which changes the directory where you are working and creates shortcuts for the folders you need to work. Section 2 is describes the model that generates aggregated data. Section 2.3 is the baseline model. For this model first run “masterfile.m” for simulations, and then “regs_simdata_sample.do” and “regs_simdata_aggreg.do” for regressions. Section 2.4 is the model with fixed geography and $\beta = 0$. For this model first run “masterfile_mod1.m” for simulations, and then “regs_simdata_aggreg_mod1.do” for regressions. Models so far have simulated aggregated data, and ran regressions on these. All the remaining sections try to include actual sampling procedures when data is collected by the Brazilian government, such that we introduce sampling error and regressions don’t match perfectly anymore. That is, these model samples and run regressions at the individual level, no longer at the location level.

Section 3.4.1 relies on the base model in Section 2, but introducing sampling error on the regressions. For this model first run “masterfile.m” for simulations, and then “regs_simdata_sample.do” for regressions. Section 3.4.2 is a model with an alternative definition of earnings that goes into regression analyses. For this model first run “masterfile_mod2.m” for simulations, and then “regs_simdata_sample_mod2.do” for regressions. Section 3.4.3 is a model where total welfare also includes a productivity draw. For this model first run “masterfile_mod3.m” for simulations, and then “regs_simdata_sample_mod3.do” for regressions. Section 3.4.4 is a model with an alternative definition of earnings plus migration costs $\mu_{ij,kt}$

with higher variance. For this model first run “masterfile_mod4.m” for simulations, and then “regs_simdata_sample_mod4.do” for regressions. Section 3.4.5 is a model with welfare including a productivity draw plus migration costs $\mu_{ij,kt}$ with higher variance. For this model first run “masterfile_mod5.m” for simulations, and then “regs_simdata_sample_mod5.do” for regressions. Section 3.4.6 is a model with welfare including a productivity draw plus the productivity draw $\nu_{ij,kt}$ with higher variance. For this model first run “masterfile_mod6.m” for simulations, and then “regs_simdata_sample_mod6.do” for regressions. Section 3.4.7 is a model with welfare including a productivity draw plus the productivity draw $\nu_{ij,kt}$ with higher variance plus lagged labor allocations $L_{i,kt-1}$ truncated from below by its median. For this model first run “masterfile_mod7.m” for simulations, and then “regs_simdata_sample_mod7.do” for regressions.

2 Base model

This model corresponds to the base model for version June 2019 of the paper. The

2.1 Primitives

We first lay out the primitives of the model. We consider 50 regions with Brazil, so $I = J = 50$. The dataset allows to identify up to 15 different types of crop, so $K = 10$. Finally, we consider $T = 5$. We choose arbitrary $L_{i,k0}$ such that $\sum_i \sum_k L_{i,k0} = L_0 = 1$. We then lay out an arbitrary geography in the model $\left\{A_{j,kt}, \bar{H_{j,t}}, \tau_{ij,kt}, \mu_{ij,kt}\right\}$. We model natural advantages $A_{i,kt}$ as a Frechet distribution, total land endowments $\overline{H}_{i,t}$ which is a unity vector, bilateral trade costs $\tau_{ij,kt}$ are drawn from a uniform distribution $U[1, 1.5]$ and bilateral migration costs $\mu_{ij,kt}$ are drawn from $LogNormal(1, 0.1)$ with location 1. In the model, workers' knowledge to grow each crop $s_{i,kt}$ depends on previous employment in the origin region $L_{i,kt-1}$ and on $\overline{s}_{i,kt}$. We consider $\overline{s}_{i,kt} = \tilde{s_k}$. Households have preference shifters $a_{k,t}$ which we assume to be a unit vector. In terms of the parameters, we follow the numbers in Table 4. For the algorithm that determines equilibrium wages and land rental rates, we consider a dampening factor $v = 0.001$. For the algorithm that determines equilibrium lambdas, we consider the averaging weight $\phi = 0.5$. Finally, we consider a a tolerance of 10^{-5} for tractability, which we will describe further below. Throught the code we denote $x^{(i)}$ as a vector of guesses that we feed in a loop, and $x^{(o)}$ as a vector of outputs that come out of the loop.

2.2 Algorithm

1. Guess $\lambda^{(i)}_{ij,k1}$ such that $\sum_j \sum_k \lambda^{(i)}_{ij,k1} = 1, \forall i$

2. Given L_{i0}, compute migration flows $L_{ij,k1} = \lambda^{(i)}_{ij,k1} L_{i0}$

3. Compute worker's knowledge $s_{i,k1} = \overline{s}_{i,k1} L^{\beta}_{i,k0}$, where $\overline{s}_{i,k1}$ is described in the first paragraph of this section

4. Compute effective labor supply $E_{ij,k1} = s_{i,k1} L_{ij,k1}$, and then $E_{j,k1} = \sum_i E_{ij,k1}$

5. Given our guess $\lambda^{(i)}_{ij,k1}$ and the variables that we constructed given our guess, we know make guesses for workers' wages and land rental rate $\left\{ w^{(i)}_{i,k1}, r^{(i)}_{i,1} \right\}$ as unit vectors and run the following inner loop:

- Calculate labor supply $H_{i,1} = \frac{{r^{(i)}_{i,1}}^{\zeta}}{b_{i,1} + {r^{(i)}_{i,1}}^{\zeta}} \overline{H}_{i,1}$

- Calculate cost shares $c_{i,k1} = \frac{{w^{(i)}_{i,k1}}^{(1-\gamma_k)} \; {r^{(i)}_{i,1}}^{\gamma_k}}{\kappa_{\pi}}$

- Calculate prices $P_{ij,k1} = \frac{c_{i,k1} \tau_{ij,k1}}{A_{i,k1}}$

- Given preferences, we can obtain aggregated price indexes $P_{j,k1} = (\sum_i P^{1-\eta}_{ij,k1})^{\frac{1}{1-\eta}}$ and $P_{j,1} = (\sum_k a_{k,1} P^{1-\sigma}_{j,k1})^{\frac{1}{1-\sigma}}$

- Total nominal expenditure in region j reflects payments to factors $X_{j,1} = \sum_k w_{j,k1} E_{j,k1} + r_{j,1} H_{j,1}$

- Total real expenditure in region j: $Q_{j,1} = \frac{X_{j,1}}{P_{j,1}}$

- The nested CES structure of preferences yields the real demand $Q_{ij,k1} = a_{k,1} (\frac{P_{ij,k1}}{P_{j,k1}})^{-\eta} (\frac{P_{j,k1}}{P_{j,1}})^{-\sigma} Q_{j,1}$

- Nominal demands are then $X_{ij,k1} = P_{ij,k1} Q_{ij,k1}$ and $X_{j,k1} = \sum_i X_{ij,k1}$

- Expenditure share of region j in sector $k$produced by region i is $\pi_{ij,k1} = \frac{X_{ij,k1}}{\sum_{i'} X_{i'j,k1}}$

- Define $\tilde{r}_{i,1} = log(r_{i,1})$. Using the land market clearing condition, we obtain a new guess for the log of the rental rate $\tilde{r}^{(o)}_{i,1} = log(\sum_k \gamma_k \sum_j \pi_{ij,k1} X_{j,k1}) - log(H_{i,1})$

- Define $\tilde{w}_{i,k1} = log(w_{i,k1})$. Using the land market clearing condition, we obtain a new guess for the log of the rental rate $\tilde{w}^{(o)}_{i,k1} = log((1 - \gamma_k) \sum_j \pi_{ij,k1} X_{j,k1}) - log(E_{i,k1})$

- Consider the numeraire price $\tilde{w}_{1,1t}$. We then normalize wages and rental rates to the numeraire

- We calculate a critical value for wages $w^{crit} = max_{i,k} | exp(\tilde{w}^{(o)}_{i,k1}) - w^{(i)}_{i,k1} |$ and land rental rates $r^{crit} = max_i | exp(\tilde{r}^{(o)}_{i,1}) - r^{(i)}_{i,1} |$. We repeat this inner loop as long as the critical values are above the tolerance level we previously defined

- If the loop continues, we update our guesses using $r_{i,1}^{(i)} = r_{i,1}^{(i)} exp(v(\tilde{r}_{i,1}^{(o)} - log(r_{i,1}^{(i)})))$ and $w_{i,k1}^{(i)} = w_{i,k1}^{(i)} exp(v(\tilde{w}_{i,k1}^{(o)} - log(w_{i,k1}^{(i)})))$

- Keep running the loop until we find equilibrium $\left\{w_{i,k1}^{*}, r_{i,1}^{*}\right\}$

1. [THIS SHOULD SAY BULLET 6] Given equilibrium input prices $\left\{w_{i,k1}^{*}, r_{i,1}^{*}\right\}$, we construct welfare $W_{ij,k1} = \frac{w_{j,k1}^{*} s_{i,k1}}{\mu_{ij,k1} P_{j,1}}$ and obtain the new guess $\lambda_{ij,k1}^{(o)} = \frac{W_{ij,k1}^{\kappa}}{\sum_j \sum_k W_{ij,k1}^{\kappa}}$

2. Calculate critical values $\lambda^{crit} = max_{ij,k} |\lambda_{ij,k1}^{(o)} - \lambda_{ij,k1}^{(i)}|$. We repeat this inner loop as long as the critical values are above the tolerance level we previously defined

3. If the loop continues, we update our guesses using $\lambda_{ij,k1}^{(i)} = \phi \lambda_{ij,k1}^{(o)} + (1 - \phi) \lambda_{ij,k1}^{(i)}$

4. Keep running the loop until we find equilibrium $\lambda_{ij,k1}^{*}$

5. Using $\lambda_{ij,k1}^{*}$ and $L_{i,0}$, we can compute equilibrium migration flows $L_{ij,k1}^{*}$

6. We repeat this whole procedure so far for $2 \leq t \leq T$. The only difference now in each round is that we use $\lambda_{ij,kt-1}^{*}$ as a guess for $\lambda_{ij,kt}^{(i)}$

2.3 Results of base model

The Matlab codes generate data, and then we use that data to run regressions (10) and (11) in the paper in Stata. Regressions are able to recover exactly the assumed values for β and κ.

2.4 Results of model with fixed geography and $\beta = 0$

These are the estimators for regressions (10) and (11) in the paper at the aggregated level, assuming $\beta = 0$ and fixed geography. By fixed geography we mean that we draw $\{A_{j,k1}, \bar{H}_{j,1}, \tau_{ij,k1}, \mu_{ij,k1}, b_{j,1}\}$ and we keep them fixed $\forall t$. The intention of this exercise is to study whether we can actually recover $\beta = 0$ in the regressions, so we make sure that the regressions are not capturing some weird underlying correlations. Also, this time we assume $T = 30$ and we burn the first 25 periods, such that we only include the last 5 periods in the regressions. In Tables 1 and 2 we see that we can indeed recover $\beta = 0$.

3 Sampling algorithm

The objective of this section is to generate data that allows for sampling error such that regressions do not match exactly.

Table 1: Estimators of log-earnings on loglagged-labor allocations assuming $\beta = 0$, OLS and PPML

	(1)	(2)
L_iktlag_log	-2.07e-49	0
	(.)	(.)
Observations	125000	125000
Pseudo R-sq		0.312

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Estimators of log-migration flows on loglagged-labor allocations assuming $\beta = 0$, OLS and PPML

	(1)	(2)
L_iktlag_log	0.00107	-0.0000330
	(0.00)	(0.00)
Observations	125000	125000
Pseudo R-sq		0.0372

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.1 Primitives

Brazilian population in 1980 was of 120 mln, so we consider total population $M = 120,000$ for every period. We want to analyze how results change for different sample sizes, so we consider $N_s = [5\%, 10\%, 25\%, 50\%, 100\%]$. The higher the sample size, the closer the estimators should get towards their population counterparts. In the algorithn, we refer X^n to a measure of some variable X that corresponds to an individual n.

3.2 Algorithm

1. Choose a sample size $x = N_S$
2. Choose period $t = 1$
3. Generate a random vector of origin regions of size M using L_{i0} as probability weights for the randomization. Recall that labor allocations are normalized such that that mass of people each period sums up to 1, so their values automatically become weights
4. Consider all pairs $\{j, k\}$ of potential destination region and sector to work on for each person. Given that each person already has an origin $i$associated to it, we can compute

their corresponding welfare $W^n_{ij,k1}$. If we then consider drawn taste shocks $\epsilon^n_{i,k1}$, we can then calculate total welfare $W^n_{ij,k1}\epsilon^n_{i,k1}$

5. Then, each person chooses the pair $\{j^*, k^*\}$ that maximizes its total welfare

6. We now know for each person their origin i, destination j, sector k, and the period is chosen (in this case is $t = 0$). We can then input for each person their wage $w_{j,k1}$, knowledge $s_{i,k1}$, and the population in their origin region in the previous period $L_{i,k0}$

7. We now proceed to do the sampling procedure. For each person we generate a random realization from a uniform distribution $U[0, 1]$, and keep only the people whose realization is lower than the chosen sample size x

8. With our sampled dataset, we can then calculate statistics we need for the regression. We first calculate $L_{ij,k0}$ by a simple counting procedure. Second, we calculate the sample version of $L_{i,k0}$ by assuming $L_{j,kt} = L_{i,kt-1}$ and then do the counting. Notice that by doing this, we lose the first period of our database. Then, with wages $w_{j,k1}$ and knowledge $s_{i,k1}$ we can then calculate their earnings. An important caveat is that we assume missing earnings for regions for which there was no migration (i.e. $L_{ij,kt} = 0$)

9. We have now calculated all variables that we need for regressions, we export them to Stata, and run regressions (10) and (11) from the paper

3.3 Comparing migration flows of the model VS the sampling ones

In this section, we compare L_{ijkt} that come from the sampling procedure vs the one that come from the model, both in levels and in logs. In Figure 1 we see that these two variables in levels are highly correlated, which complies with what we expect.

In Figure 2 we observe the same result for the variables in logs.

3.4 Results

All tables include destination-crop-year fixed effects and origin-destination-year fixed effects, and standard errors are clustered to the destination-crop-year level. Each column of each table represents different sample sizes described in the first paragraph of this section. We group tables according to different models we are trying out. Each group of tables contain four tables, where the first one corresponds to the regression of sampled log-earnings on loglagged-labor allocations; the second one, of sampled log-earnings on loglagged-labor

Figure 1: Migration flows from the model and from the sampling procedure, in levels

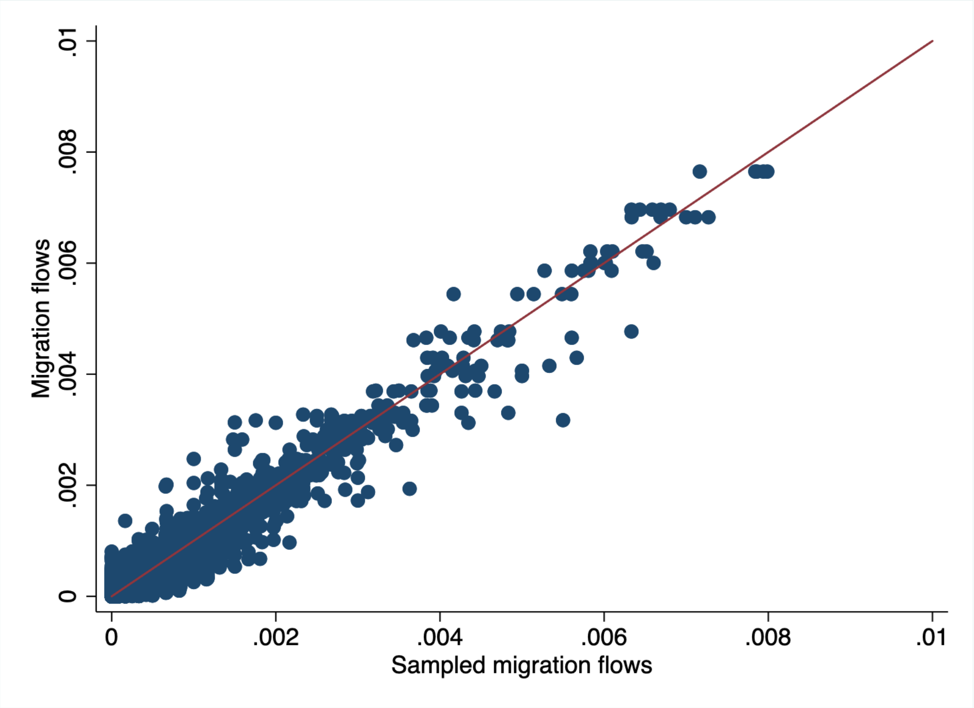

Figure 2: Migration flows from the model and from the sampling procedure, in logs

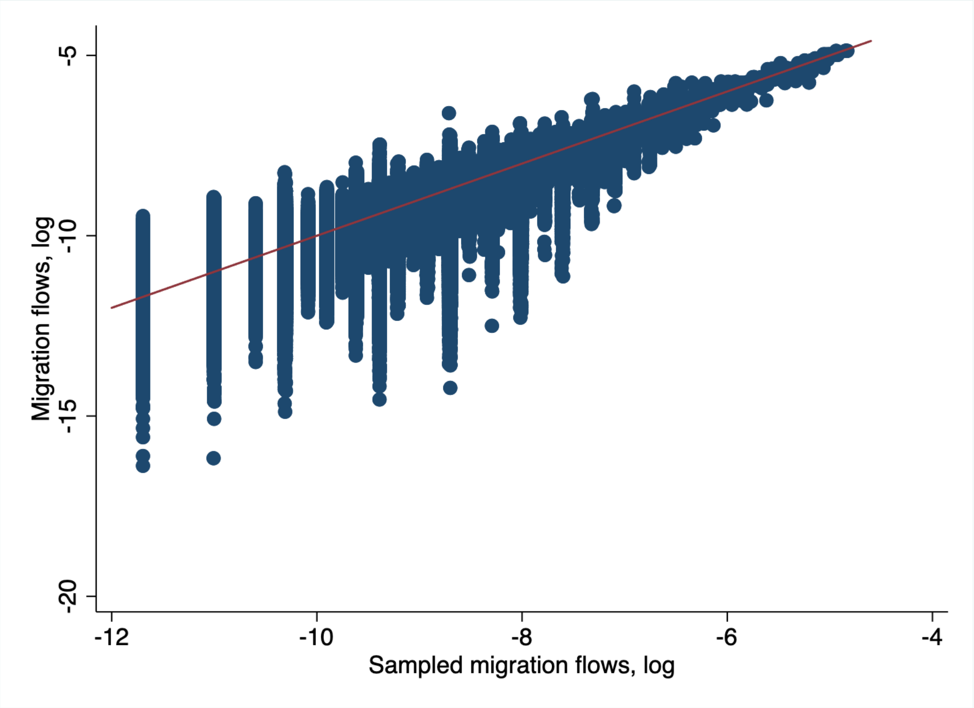

Table 3: Estimators of sampled log-earnings on sampled loglagged-labor allocations, baseline model, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0853***	0.0906***	0.0941***	0.0964***	0.0989***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Observations	12300	24827	46163	63479	78922
R-sq	0.999	0.999	1.000	1.000	1.000

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0860***	0.0904***	0.0940***	0.0964***	0.0987***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Observations	12300	24827	46163	63479	78922
Pseudo R-sq	0.397	0.364	0.324	0.302	0.292

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

allocations; the third one, of sampled log-migration flows on sampled loglagged-labor allocations; the fourth one, of sampled log-migration flows on loglagged-labor allocations. Each table contain results for OLS and PPML estimators.

3.4.1 Baseline model

Tables 3-6 report baseline results. In Table 3 we see that both OLS and PPML estimators exhibit a downward bias on β but that bias dampens as sample size increases. As expected, in Table 4 we exactly recover $\beta = 0.1$. In Table 5 we also observe that downward bias on the coefficient $\kappa\beta$ according ti regression (11) of the paper, this bias also dampens with sample size, and the PPML seem to be much less affected by this bias in comparison to the OLS estimator. As expected, estimators in Table 6 are closer to its theoretical value $\kappa\beta = 1.25$ in contrast to Table 5 due to lesser sampling errors.

3.4.2 Results with a different definition of earnings

Tables 7-10 report results considering earnings to be $\frac{w_{j,kt}s_{i,kt}}{\mu_{ij,kt}}$ instead of $w_{j,kt}s_{i,kt}$ as in regression (10) of the paper. The main difference with previous results is that in Table 8 we start observing a downward bias that we didn't observe in Table 4. Again, these biases also dampen with sample size.

Table 4: Estimators of sampled log-earnings on loglagged-labor allocations, baseline model, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.100***	0.100***	0.100***	0.100***	0.100***
	(0.000000162)	(9.91e-08)	(6.88e-08)	(5.36e-08)	(4.86e-08)
Observations	12760	25009	46212	63510	78922
R-sq	1.000	1.000	1.000	1.000	1.000

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.100***	0.100***	0.100***	0.100***	0.100***
	(0.000000161)	(0.000000110)	(7.94e-08)	(6.58e-08)	(6.01e-08)
Observations	12760	25009	46212	63510	78922
Pseudo R-sq	0.396	0.365	0.324	0.302	0.292

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Estimators of sampled log-migration flows on sampled loglagged-labor allocations, baseline model, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0160***	0.0301***	0.0534***	0.0775***	0.0959***
	(0.00457)	(0.00331)	(0.00269)	(0.00230)	(0.00209)
Observations	12300	24827	46163	63479	78922
R-sq	0.688	0.653	0.666	0.701	0.753

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.108***	0.110***	0.118***	0.124***	0.122***
	(0.00702)	(0.00489)	(0.00324)	(0.00245)	(0.00177)
Observations	72445	90026	96794	97637	97853
Pseudo R-sq	0.0847	0.0857	0.0843	0.0831	0.0819

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Estimators of sampled log-migration flows on loglagged-labor allocations, baseline model, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0192***	0.0332***	0.0571***	0.0802***	0.0965***
	(0.00457)	(0.00340)	(0.00280)	(0.00233)	(0.00210)
Observations	12760	25009	46212	63510	78922
R-sq	0.684	0.653	0.666	0.701	0.753

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.129***	0.120***	0.124***	0.128***	0.123***
	(0.00706)	(0.00499)	(0.00334)	(0.00248)	(0.00178)
Observations	75692	90674	96949	97686	97853
Pseudo R-sq	0.0855	0.0858	0.0843	0.0831	0.0819

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Estimators of sampled log-earnings on sampled loglagged-labor allocations, model 2, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0865***	0.0879***	0.0942***	0.0950***	0.0977***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Observations	12017	24658	46173	63703	79045
R-sq	0.986	0.980	0.975	0.972	0.970

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0860***	0.0884***	0.0944***	0.0951***	0.0980***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Observations	12017	24658	46173	63703	79045
Pseudo R-sq	0.331	0.286	0.254	0.236	0.228

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Estimators of sampled log-earnings on loglagged-labor allocations, model 2, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0980***	0.0990***	0.0996***	0.0992***	0.0995***
	(0.00170)	(0.00112)	(0.000795)	(0.000661)	(0.000582)
Observations	12542	24939	46318	63724	79078
R-sq	0.986	0.981	0.975	0.972	0.970

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0985***	0.0982***	0.0994***	0.0993***	0.0999***
	(0.00186)	(0.00113)	(0.000810)	(0.000687)	(0.000636)
Observations	12542	24939	46318	63724	79078
Pseudo R-sq	0.332	0.288	0.254	0.237	0.229

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9: Estimators of sampled log-migration flows on sampled loglagged-labor allocations, model 2, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0233***	0.0322***	0.0579***	0.0771***	0.0961***
	(0.00485)	(0.00339)	(0.00268)	(0.00233)	(0.00215)
Observations	12017	24658	46173	63703	79045
R-sq	0.695	0.656	0.668	0.704	0.750

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.112***	0.117***	0.123***	0.123***	0.122***
	(0.00733)	(0.00512)	(0.00328)	(0.00230)	(0.00177)
Observations	71534	88804	96554	97607	97902
Pseudo R-sq	0.0838	0.0867	0.0843	0.0825	0.0816

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10: Estimators of sampled log-migration flows on loglagged-labor allocations, model 2, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0263***	0.0362***	0.0603***	0.0793***	0.0973***
	(0.00464)	(0.00348)	(0.00272)	(0.00239)	(0.00215)
Observations	12542	24939	46318	63724	79078
R-sq	0.693	0.655	0.668	0.704	0.750

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.131***	0.130***	0.128***	0.127***	0.124***
	(0.00717)	(0.00510)	(0.00326)	(0.00234)	(0.00177)
Observations	75071	90035	96897	97656	97951
Pseudo R-sq	0.0842	0.0870	0.0843	0.0825	0.0817

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.4.3 Results when total utility also has a productivity draw

Tables 11-14 report results considering total welfare to be $W_{ij,kt}\epsilon_{i,kt}\nu_{ij,kt}$, where the productivity draw $\nu_{ij,kt}$ distributes $LogNormal(1,1)$. Notice that this also changes the definition of earnings, so we now have $earn_{ij,kt} = w_{j,kt}s_{i,kt}\nu_{ij,kt}$. In this case, it seems like the downward bias for tables 1 and 2 are not disappearing as sample size increases. These estimators exhibit the same downward biases as in the previous subsection.

3.4.4 Results when a different definition of earnings + higher variance of $\mu_{ij,kt}$

Tables 15-18 show results also for our alternative definition of earnings, but we now consider a higher variance for migration costs such that $\mu_{ij,kt} \sim LogNormal(1,2)$ with location 1. There aren't really marked differences with respect to Tables 7-10.

3.4.5 Results with the productivity draw + higher variance of $\mu_{ij,kt}$

Tables 19-22 show results also making total welfare to be a function of a productivity draw with the same higher variance for migration costs as in the previous subsubsection. In comparison to Tables 11-14, estimators for the earnings regressions seem to exhibit a more persistent bias that does not dissipate with sample size, and estimators for the migration flow regression are more biased downwards.

Table 11: Estimators of sampled log-earnings on sampled loglagged-labor allocations, model 3, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0656***	0.0667***	0.0754***	0.0766***	0.0809***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)
Observations	9724	19396	36590	50835	64907
R-sq	0.760	0.702	0.632	0.588	0.563

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0687***	0.0737***	0.0813***	0.0863***	0.0907***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)
Observations	9724	19396	36590	50835	64907
Pseudo R-sq	0.437	0.397	0.353	0.323	0.308

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 12: Estimators of sampled log-earnings on loglagged-labor allocations, model 3, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0793***	0.0768***	0.0842***	0.0819***	0.0859***
	(0.0104)	(0.00674)	(0.00453)	(0.00373)	(0.00334)
Observations	10129	19630	36679	50835	64940
R-sq	0.759	0.702	0.632	0.589	0.564

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0748***	0.0799***	0.0900***	0.0927***	0.0966***
	(0.0103)	(0.00704)	(0.00510)	(0.00488)	(0.00456)
Observations	10129	19630	36679	50835	64940
Pseudo R-sq	0.440	0.397	0.353	0.323	0.308

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 13: Estimators of sampled log-migration flows on sampled loglagged-labor allocations, model 3, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0185*	0.0325***	0.0485***	0.0620***	0.0753***
	(0.00732)	(0.00538)	(0.00429)	(0.00391)	(0.00384)
Observations	9724	19396	36590	50835	64907
R-sq	0.628	0.559	0.496	0.467	0.467

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.129***	0.130***	0.119***	0.127***	0.124***
	(0.00932)	(0.00797)	(0.00717)	(0.00674)	(0.00649)
Observations	67988	85046	95786	97565	97872
Pseudo R-sq	0.0929	0.0977	0.0987	0.0979	0.0971

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 14: Estimators of sampled log-migration flows on loglagged-labor allocations, model 3, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0216**	0.0310***	0.0541***	0.0651***	0.0808***
	(0.00708)	(0.00548)	(0.00450)	(0.00408)	(0.00389)
Observations	10129	19630	36679	50835	64940
R-sq	0.624	0.558	0.496	0.467	0.467

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.143***	0.139***	0.131***	0.134***	0.130***
	(0.00947)	(0.00847)	(0.00733)	(0.00687)	(0.00654)
Observations	70869	86452	96029	97565	97921
Pseudo R-sq	0.0930	0.0976	0.0986	0.0980	0.0971

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 15: Estimators of sampled log-earnings on sampled loglagged-labor allocations, model 4, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0816***	0.0898***	0.0920***	0.0921***	0.0947***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Observations	11943	23953	44689	61146	76228
R-sq	0.937	0.913	0.880	0.856	0.837

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0792***	0.0911***	0.0921***	0.0935***	0.0969***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Observations	11943	23953	44689	61146	76228
Pseudo R-sq	0.345	0.306	0.270	0.252	0.244

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 16: Estimators of sampled log-earnings on loglagged-labor allocations, model 4, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0945***	0.0977***	0.0970***	0.0955***	0.0958***
	(0.00363)	(0.00269)	(0.00192)	(0.00167)	(0.00152)
Observations	12442	24190	44763	61208	76263
R-sq	0.938	0.914	0.880	0.856	0.837

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0933***	0.0990***	0.0978***	0.0965***	0.0983***
	(0.00277)	(0.00212)	(0.00160)	(0.00151)	(0.00141)
Observations	12442	24190	44763	61208	76263
Pseudo R-sq	0.345	0.306	0.270	0.252	0.244

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 17: Estimators of sampled log-migration flows on sampled loglagged-labor allocations, model 4, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0222***	0.0293***	0.0534***	0.0758***	0.0897***
	(0.00493)	(0.00363)	(0.00287)	(0.00269)	(0.00252)
Observations	11943	23953	44689	61146	76228
R-sq	0.682	0.644	0.632	0.652	0.682

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.114***	0.111***	0.119***	0.124***	0.122***
	(0.00763)	(0.00506)	(0.00371)	(0.00298)	(0.00261)
Observations	71949	89113	96755	97510	97755
Pseudo R-sq	0.0863	0.0876	0.0861	0.0849	0.0840

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 18: Estimators of sampled log-migration flows on loglagged-labor allocations, model 4, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0255***	0.0321***	0.0563***	0.0779***	0.0906***
	(0.00492)	(0.00384)	(0.00295)	(0.00272)	(0.00255)
Observations	12442	24190	44763	61208	76263
R-sq	0.681	0.644	0.632	0.652	0.682

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.137***	0.119***	0.125***	0.127***	0.123***
	(0.00749)	(0.00523)	(0.00380)	(0.00302)	(0.00263)
Observations	75737	90163	96949	97608	97804
Pseudo R-sq	0.0866	0.0876	0.0861	0.0849	0.0840

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 19: Estimators of sampled log-earnings on sampled loglagged-labor allocations, model 5, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0631***	0.0828***	0.0723***	0.0738***	0.0794***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)
Observations	9196	19298	35709	49496	63144
R-sq	0.772	0.693	0.628	0.590	0.564

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0599***	0.0768***	0.0734***	0.0764***	0.0832***
	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Observations	9196	19298	35709	49496	63144
Pseudo R-sq	0.418	0.389	0.341	0.318	0.303

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 20: Estimators of sampled log-earnings on loglagged-labor allocations, model 5, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0831***	0.0982***	0.0806***	0.0792***	0.0848***
	(0.0107)	(0.00728)	(0.00482)	(0.00394)	(0.00339)
Observations	9675	19516	35753	49521	63176
R-sq	0.769	0.693	0.629	0.590	0.564

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0761***	0.0967***	0.0847***	0.0847***	0.0902***
	(0.0111)	(0.00801)	(0.00573)	(0.00499)	(0.00458)
Observations	9675	19516	35753	49521	63176
Pseudo R-sq	0.420	0.390	0.342	0.319	0.303

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 21: Estimators of sampled log-migration flows on sampled loglagged-labor allocations, model 5, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0237**	0.0359***	0.0448***	0.0580***	0.0723***
	(0.00799)	(0.00573)	(0.00446)	(0.00410)	(0.00402)
Observations	9196	19298	35709	49496	63144
R-sq	0.613	0.537	0.480	0.452	0.440

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0968***	0.109***	0.119***	0.121***	0.124***
	(0.0113)	(0.00911)	(0.00842)	(0.00814)	(0.00811)
Observations	65869	85010	95742	97447	97843
Pseudo R-sq	0.0930	0.0950	0.0975	0.0969	0.0959

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 22: Estimators of sampled log-migration flows on loglagged-labor allocations, model 5, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0280***	0.0455***	0.0502***	0.0629***	0.0769***
	(0.00769)	(0.00585)	(0.00468)	(0.00426)	(0.00410)
Observations	9675	19516	35753	49521	63176
R-sq	0.609	0.536	0.481	0.452	0.440

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.127***	0.132***	0.132***	0.131***	0.132***
	(0.0111)	(0.00904)	(0.00879)	(0.00842)	(0.00837)
Observations	69718	86213	95897	97496	97892
Pseudo R-sq	0.0930	0.0952	0.0975	0.0969	0.0960

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 23: Estimators of sampled log-earnings on sampled loglagged-labor allocations, model 6, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0201	0.0405**	0.0332***	0.0340***	0.0440***
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	4862	10893	21721	31041	41810
R-sq	0.714	0.613	0.525	0.463	0.414

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	-0.00838	0.0505**	0.0347**	0.0317**	0.0538***
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
Observations	4862	10893	21721	31041	41810
Pseudo R-sq	0.616	0.581	0.494	0.447	0.450

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.4.6 Results with the productivity draw + higher variance of $\nu_{ij,kt}$

Tables 23-26 report results with the same productivity draw as before but with a higher variance such that $\nu_{ij,kt} \sim LogNormal(1, 10)$. We can now see clear downward biases all across the board. Moreover, with small sample sizes a lot of the estimators are not even significant. Finally, is not even clear that the biases are actually disappearing with sample sizes.

3.4.7 Results with the productivity draw + higher variance of $\nu_{ij,kt}$ + truncated $L_{i,kt-1}$ from below by its median

Tables 27-30 show results including the productivity draw $\nu_{ij,kt} \sim LogNormal(1, 10)$ and the right-hand side variable in all equations $L_{i,kt-1}$ truncated from below by its median by period. We are doing this to check if part of the downward bias in the migration flow regressions come from errors being drawn asymmetrically at $L_{ij,kt} = 0$. We confirm this by comparing these results with Tables 25 and 26, where now the downward bias is dampened. Finally, note that Table 30 has one column less in comparison to all previous ones. This is because the PPML command did not converge for a sample size of 5%, so we excluded that first column.

Table 24: Estimators of sampled log-earnings on loglagged-labor allocations, model 6, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.0351	0.0458**	0.0466***	0.0520***	0.0559***
	(0.0245)	(0.0152)	(0.0102)	(0.00825)	(0.00689)
Observations	5261	11112	21782	31050	41810
R-sq	0.703	0.611	0.525	0.463	0.414

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.00102	0.0570***	0.0514***	0.0470***	0.0630***
	(0.0248)	(0.0171)	(0.0127)	(0.0119)	(0.0103)
Observations	5261	11112	21782	31050	41810
Pseudo R-sq	0.610	0.579	0.494	0.448	0.450

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 25: Estimators of sampled log-migration flows on sampled loglagged-labor allocations, model 6, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	-0.00630	0.0241*	0.0163*	0.0230***	0.0363***
	(0.0143)	(0.00969)	(0.00765)	(0.00664)	(0.00614)
Observations	4862	10893	21721	31041	41810
R-sq	0.653	0.553	0.458	0.399	0.351

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0740***	0.0746***	0.0747***	0.0680***	0.0759***
	(0.0172)	(0.0146)	(0.0143)	(0.0141)	(0.0141)
Observations	52515	71801	88938	94701	97111
Pseudo R-sq	0.122	0.131	0.137	0.141	0.142

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 26: Estimators of sampled log-migration flows on loglagged-labor allocations, model 6, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	-0.0107	0.0293**	0.0295***	0.0340***	0.0462***
	(0.0150)	(0.0106)	(0.00860)	(0.00762)	(0.00680)
Observations	5261	11112	21782	31050	41810
R-sq	0.641	0.552	0.458	0.400	0.352

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	0.103***	0.0967***	0.0987***	0.0896***	0.0933***
	(0.0184)	(0.0167)	(0.0163)	(0.0163)	(0.0161)
Observations	56675	73515	89183	94759	97111
Pseudo R-sq	0.122	0.131	0.137	0.141	0.142

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 27: Estimators of sampled log-earnings on sampled loglagged-labor allocations, model 7, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0666	0.0426	0.0285	0.0394*	0.0468**
	(0.08)	(0.05)	(0.03)	(0.02)	(0.02)
Observations	1057	3290	9205	13967	19908
R-sq	0.789	0.745	0.644	0.582	0.535

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	-0.0175	-0.00591	0.0335	0.0113	0.0428*
	(0.07)	(0.04)	(0.02)	(0.02)	(0.02)
Observations	1044	3488	9106	13937	19836
Pseudo R-sq	0.674	0.678	0.577	0.523	0.550

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 28: Estimators of sampled log-earnings on loglagged-labor allocations, model 7, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	-0.0114	-0.0106	0.0463	0.0776***	0.0726***
	(0.0826)	(0.0421)	(0.0243)	(0.0196)	(0.0159)
Observations	1048	3488	9106	13937	19836
R-sq	0.775	0.740	0.653	0.578	0.526

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	-0.0254	-0.0173	0.0825**	0.0401	0.0773***
	(0.0841)	(0.0469)	(0.0281)	(0.0236)	(0.0199)
Observations	1048	3488	9106	13937	19836
Pseudo R-sq	0.674	0.678	0.578	0.523	0.550

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 29: Estimators of sampled log-migration flows on sampled loglagged-labor allocations, model 7, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	-0.0307	0.00865	0.0256	0.0278	0.0465**
	(0.0552)	(0.0335)	(0.0228)	(0.0189)	(0.0161)
Observations	1057	3290	9205	13967	19908
R-sq	0.760	0.690	0.582	0.529	0.485

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
sample_L_iktlag_log	0.0707*	0.128***	0.0745*	0.0829**	0.0898**
	(0.0360)	(0.0297)	(0.0326)	(0.0313)	(0.0316)
Observations	18924	26748	37691	41752	45369
Pseudo R-sq	0.127	0.135	0.148	0.154	0.159

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 30: Estimators of sampled log-migration flows on loglagged-labor allocations, model 7, OLS and PPML

	(1)	(2)	(3)	(4)	(5)
L_iktlag_log	-0.116	-0.00240	0.0281	0.0450*	0.0592***
	(0.0612)	(0.0316)	(0.0215)	(0.0192)	(0.0167)
Observations	1048	3488	9106	13937	19836
R-sq	0.758	0.687	0.598	0.527	0.482

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)
L_iktlag_log	0.114***	0.103***	0.0998***	0.101***
	(0.0318)	(0.0306)	(0.0303)	(0.0305)
Observations	26986	37004	41965	45247
Pseudo R-sq	0.142	0.154	0.162	0.167

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$