

The Determinants of International Migration Accounting for Self-selection

Simone Bertoli (EUI)

Jesús Fernández-Huertas Moraga (IAE-CSIC)

Francesc Ortega (UPF)

Second International Conference on Migration and Development
Washington DC, 11th September 2009

Questions in International Migration

- Why do people migrate?
 - Follow better economic opportunities (i.e. higher wages) but:
 - How much? How responsive are individuals to income differentials?
 - Whose economic opportunities? Individuals are observed at their preferred location but need to infer how they would fare elsewhere
- Who migrates?
 - Migrants are selected: highly educated individuals tend to migrate more in general

Challenge

- Logical inconsistency of international migration models trying to estimate elasticities of migration with respect to income
- They try to estimate *why* people migrate without taking into account *who* migrates
- Very strong assumptions are implicit:
 - No unobserved ability, defined as individual-specific unobserved factors that determine wages and migration simultaneously
- Migrant selection itself biases counterfactual income estimates required to estimate migration equations

Challenge (II)

- Disconnect with literature on internal migration flows, which has shown this is an important source of bias: Dahl (2002), Kennan and Walker (2003), Bayer et al. (2008)
- Of course, some international migration studies have recognized this problem but have not been able to tackle it:
 - Need of individual-level data
 - But only country-level data generally available for international migration flows

Our contribution

- Estimation of a model of international migration controlling for selection using state-of-the-art techniques from the internal migration literature
- Use of micro evidence to test the main theories of scale, selection and sorting of migration flows, following the macro level studies of Grogger and Hanson (2008) and Belot and Hatton (2008)
- Study a great episode in international migration: Ecuador (1999-2005)

Ecuadorian migration

- Large migration flows due to severe economic crisis in 1999-2000
- Vast majority of flows to only two destinations: US and Spain (80-90 per cent per year)
- We merge three individual-level datasets that provide information on Ecuadorians in Ecuador (ENEMDU 2005), in the US (ACS 2007) and in Spain (ENI 2007)

Preliminary results

- Significant selection bias
- The two main models of migration are supported by the data: migration decisions are sensitive to income differentials
- Migration costs seem to be a more important determinant of migration than income differentials. In our case:
 - Language, culture and **migration policies** favor the Spanish destination
 - Networks favor the US destination

Relevant literature

- International migration: Grogger and Hanson (2008); Belot and Hatton (2008); Clark, Hatton and Williamson (2007); Mayda (2008); Clemens, Montenegro and Pritchett (2008); Ortega and Peri (2009)
- Internal migration: Falaris (1987, 1988); Bayer, Keohane and Timmins (2008); Borjas, Bronars and Trejo (1992); Nakosteen and Zimmer (1980)
- Selection methods: Dahl (2002); Bourguignon, Fournier and Gurgand (2007); Bayer, Khan and Timmins (2008); Dubin and McFadden (1984); Lee (1983); Heckman (1979)
- Ecuador: Jokisch and Pribilsky (2002); Bertoli (2008); Gratton (2007); Jacomé (2004); OPI (2007)

Model

- Assume individuals locate wherever they maximize their expected utility
- Two equations for each individual (i) and potential destination (j): discrete choice migration equation and wage equation
- Unobserved individual component (i.e. ability), affecting both wages and migration
- Equations:

$$U_{ij} = \alpha w_{ij} + \beta x_i + \gamma_j + \delta \rho_{ij} + \eta_i + \zeta_j + \epsilon_{ij}^m$$

$$w_{ij} = \alpha z_i + \beta_j + \delta \rho_{ij} + \eta_i + \zeta_j + \epsilon_{ij}^w$$

Estimation

- Classical method: estimate wage equation (2) (i.e. OLS Mincer regression) and use predicted wages to estimate migration equation (1) with a conditional logit, assuming error terms are i.i.d. across destinations (j) and follow a Extreme Value Type I distribution (IIA assumption).
- Unobserved individual-specific factors introduce two challenges:
 - Challenge 1: Self-selection problem in wage equation (predicted wages are biased)
 - Challenge 2: Correlated shocks across alternatives for a given individual (IIA assumption is violated)


Estimation (II)

- What are unobservable individual-specific factors constant across locations?
 - Ability: more capable individuals can earn both higher wages and be better able to migrate (lower migration costs)
 - Low risk aversion: increases propensity to migrate to all destinations and increases wages (risk premium)
 - English fluency (in principle observable, in practice not in the data): increases probability to migrate mostly to the US and increases wages in all three locations
 - ...

Estimation (III)

- Solutions:
 - Challenge 1: estimation of the wage equation correcting for self-selection into migration. We use Dahl's (2002) method and use predicted wages in the second step
 - Challenge 2: estimation of the migration equation when the IIA assumption is violated. We use:
 - Nested logit model: assume the error term is GEV and establish a nest for stayers (degenerate) and a nest for migrants (assuming the error term of migrant destinations is correlated)
 - Alternative specific multinomial probit model: assume the error term follows a multinomial normal

The Ecuador case

- Population: 12 million in 2001 (last census)
- GDP pc in PPP terms: \$7,000 in 2006
 - US: \$44,000
 - Spain: \$29,000
- Pre-crisis (before 1999) Ecuadorian diaspora:
 - US: 272,000 (source: 2000 US Census)
 - Spain: 76,000 (source: 2001 Spain Census, only 8,000 were registered according to the Padrón) 
- Post-crisis (2005) Ecuadorian diaspora:
 - US: 394,000 (source: ACS 2007)
 - Spain: 457,000 (source: Padrón 2006)

Why Ecuador?

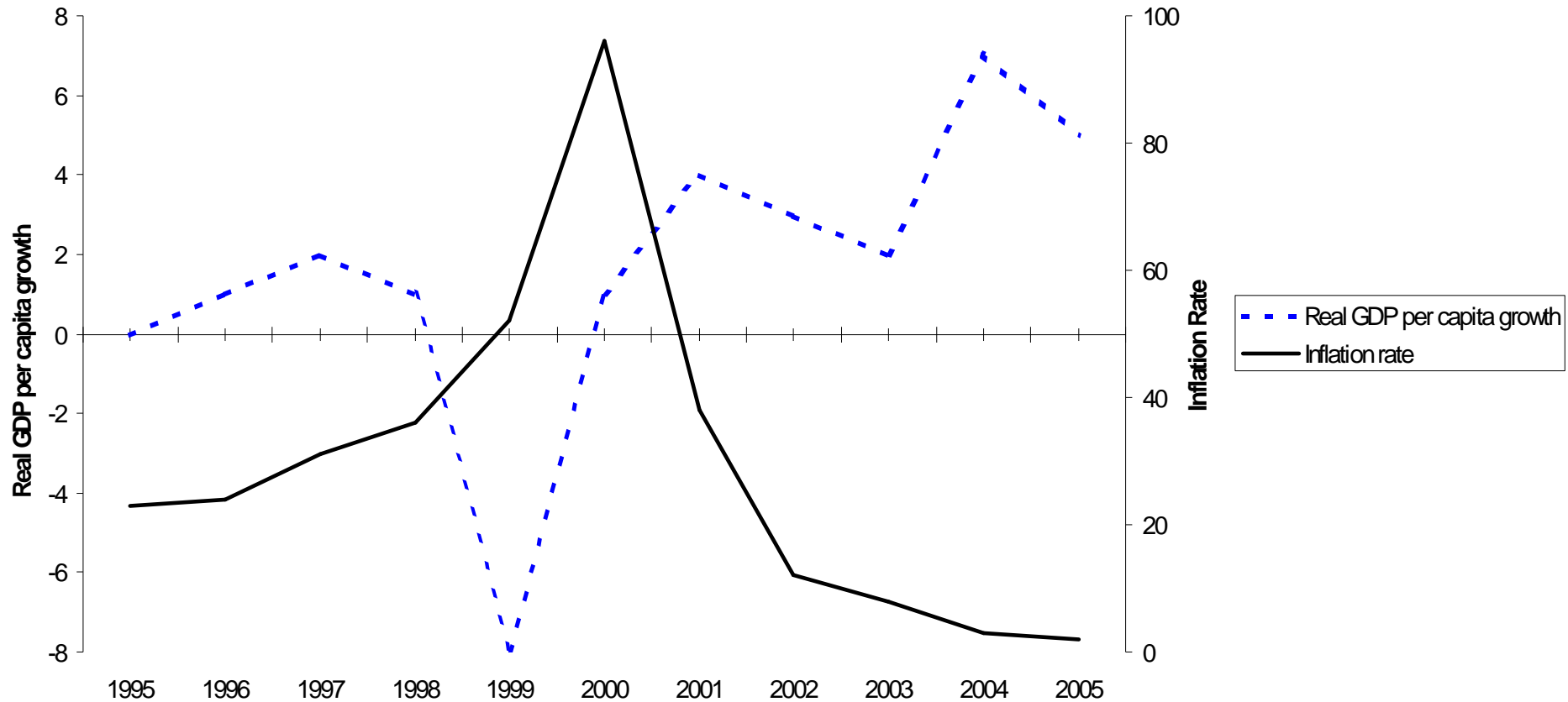
- Large migration episode: approximately 600,000 emigrants left in 1999-2005
- Only two relevant destinations: Spain and the US (80-90 per cent of the flows)
- Large income differences between the three countries
- If an income maximization theory of migration can fail, it has to fail in this case: most of the migrants went to the lowest income destination (Spain over US was preferred)
- Data availability

The Ecuadorian Crisis: Facts

- 1997: El Niño floods in the coastal region (great damages to agriculture)
- 1997-1998: Sharp drop in global oil prices (Ecuador's main export)
- 1997-1998: Emerging markets instability
- March 1999: Fears of bank runs, bank holidays, deposits frozen, etc. (salvataje). 40 per cent of deposit base in failed banks.
- January 2000: Dollarization is announced

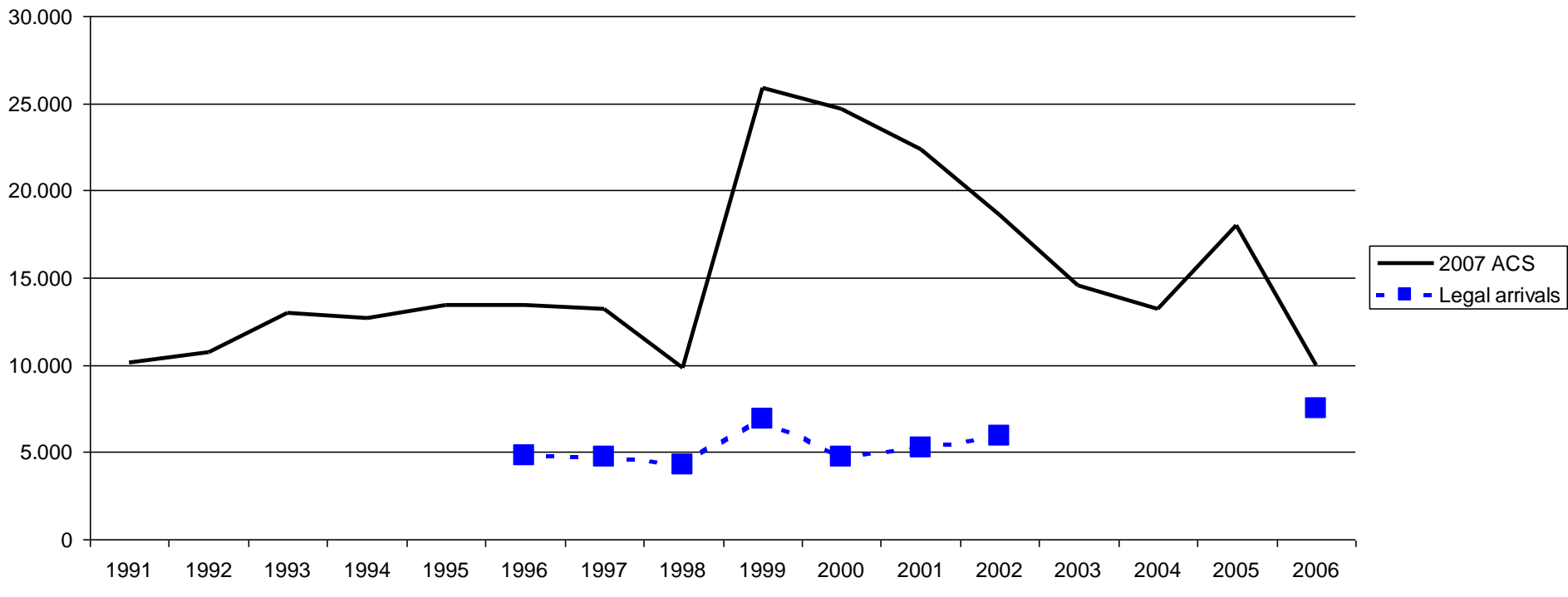
The Ecuadorian Crisis: Macro

Macroeconomic Conditions in Ecuador (1995-2005); WDI 2008



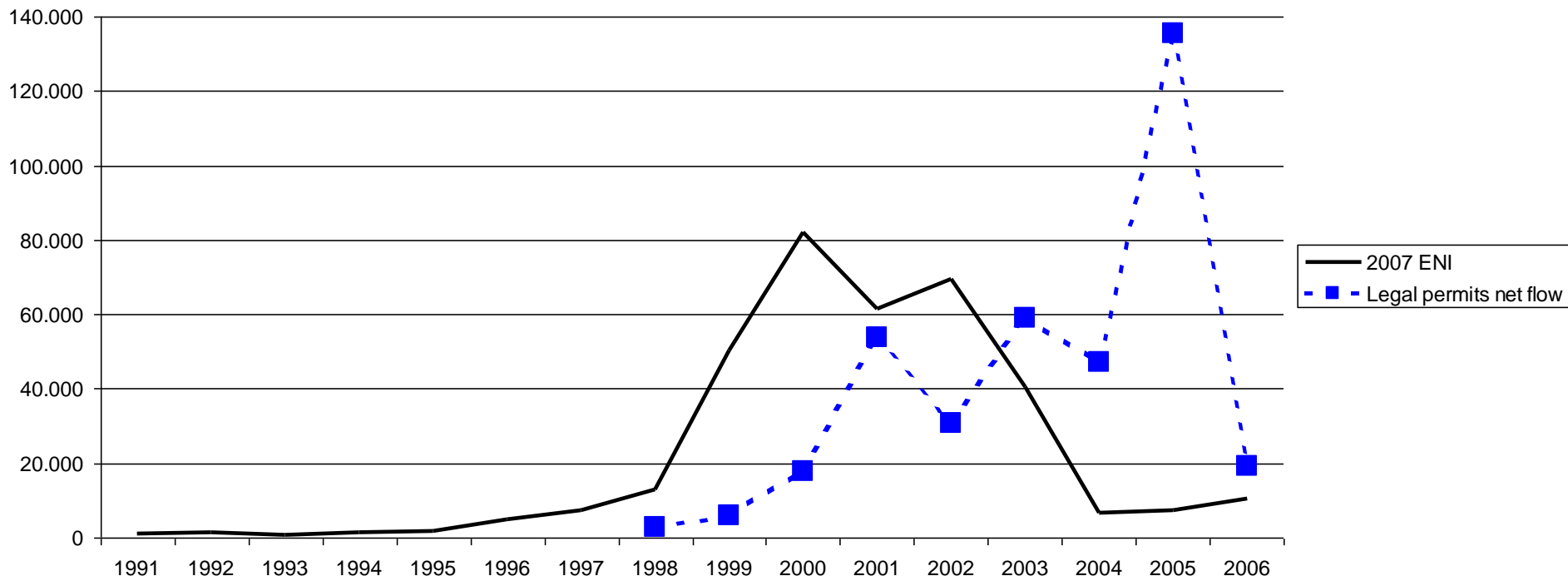
The Ecuadorian Crisis: Migration to the US

Arrivals of Ecuadorians in the US according to the ACS 2007



The Ecuadorian Crisis: Migration to Spain

Arrivals of Ecuadorians in Spain according to the ENI 2007



Sample selection

- Individuals born in Ecuador living in Ecuador in 1998 **DESCRIPTIVES**
- Aged 16 to 49 in 1998 (born between 1949 and 1982; 70 per cent of the US and 77 per cent of the Spanish sample older than 16). Robustness checks with age group 25-49 in 1998 (born in 1949-1973; 35 per cent of the US and 38 per cent of the Spanish sample).
- Surveys:
 - ENEMDU 2005: 28,122 non-migrants
 - ACS 2007: 509 migrated in 1999-2005
 - ENI 2007: 915 migrated in 1999-2005

Migrant selection and sorting

- Probits on being a college graduate (marginal effects reported)
Males aged 16 to 49 in 1998 (born in 1949-1982)

Selection to the United States	-0,002 [0.022]	0,011 [0.024]	0,01 [0.024]	0,01 [0.024]	
Selection to Spain	-0,066 [0.015]***	-0,057 [0.016]***	-0,059 [0.015]***	-0,054 [0.016]***	-0,052 [0.016]***
Age groups	No	Yes	No	Yes	Yes
Year of birth dummy	No	No	Yes	No	No
Marital status	No	No	No	Yes	Yes
Province of origin	No	No	No	No	Yes
Observations	15.294	15.294	15.294	15.294	15.035

Females aged 16 to 49 in 1998 (born in 1949-1982)

Selection to the United States	0,083 [0.030]***	0,081 [0.030]***	0,084 [0.030]***	0,093 [0.031]***	
Selection to Spain	0,019 [0.021]	0,012 [0.021]	0,012 [0.021]	0,02 [0.021]	0,018 [0.021]
Age groups	No	Yes	No	Yes	Yes
Year of birth dummy	No	No	Yes	No	No
Marital status	No	No	No	Yes	Yes
Province of origin	No	No	No	No	Yes
Observations	16.320	16.320	16.320	16.320	16.047

Earnings and Employment

EMPLOYMENT RATES

	Ecuador	US	Spain
NOCO			
men	0,92	0,90	0,90
women	0,80	0,63	0,79
COG			
men	0,94	0,92	0,92
women	0,54	0,63	0,84

MEDIAN EARNINGS IN 2005 US DOLLARS

	Ecuador	US	Spain
NOCO			
men	2.280	21.440	15.431
women	1.560	14.865	10.521
COG			
men	6.000	30.492	15.431
women	4.344	20.582	10.942

RETURNS TO SKILL

Results (I)

- Sample selection: the Dahl polynomial always matters for the US and also for Ecuador in some specifications. Heckman's rho is negatively significant for both the US and Ecuador and statistically zero for Spain. After the correction:
 - Predicted US income falls by 1 per cent
 - Predicted Spanish income increases by 1 per cent
 - Predicted Ecuador income drops by 1 per cent
- Correlation of shocks across destinations. We run IIA tests on the conditional logit model by dropping one destination at a time: IIA fails

Migration theories

- Testing migration theories: introduce income in the discrete choice model in two different ways
 - Linearly. A positive coefficient on income validates the linear utility model: Grogger and Hanson (2008)
 - Logarithm. A positive coefficient on log income validates the log utility model: Belot and Hatton (2008)

Results (II)

Nested Logit Model	MAIN (LOG)	LINEAR	CLOGIT	TAX-ADJUSTED
Linear income		0.308***		
Log income	0.821***		0.885***	0.8008***
United States				
College	0.015	-1.093	0.152	0.0153
Female	0.216**	-3.424	0.194*	0.1894**
Constant	-5.230***	-54.422	-4.664***	-5.1956***
Spain				
College	-0.011	-0.081	-0.015	-0.0212
Female	0.345***	2.844	0.395***	0.2983***
Constant	-6.282***	-39.735	-7.059***	-6.0958***
Observations	29546	29546	29546	29546
Dissimilarity coefficient	0.492***	64.363	1	0.4531***
US-Spain correlation	0.76	-	0	0.795
Log-likelihood	-1257407	-1238554	-1258053	-1258001

- Includes controls for Age, Age2, Married, Household size
- Normalization Ecuador controls equal to zero.
- Implied correlation for the linear model is greater than 1 since the dissimilarity coefficient is too large

Results (III)

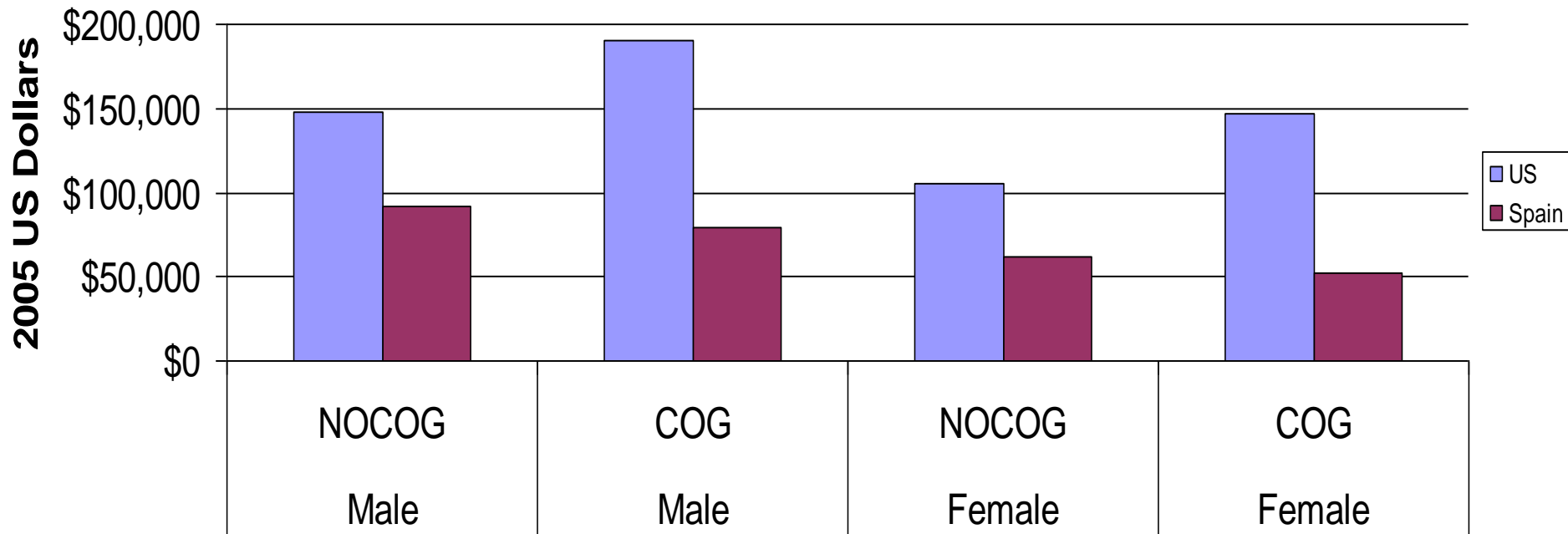
- The conditional logit model is rejected by the data (no IIA): unobserved heterogeneity correlated across destinations
- Income is a significant determinant of migration decisions in all specifications:
 - The log utility model is supported by both the nested logit and multinomial probit framework
 - The linear utility model works well in the multinomial probit framework but its nested logit results are weaker (the model cannot be rejected though)
- Magnitudes: sensitivity of migration to income higher than in macro papers but low to explain observed patterns

Migration costs (I)

- All controls in the migration decision equation can be interpreted as group-specific migration costs (individual-specific migration costs remain in the error term)
- We can thus recover implied migration costs by destination, gender and education
- Compare with physical migration costs:
 - Spain: \$2,500-\$3,000 (source: ENI 2007)
 - US: \$7,000-\$9,000 paid to smugglers (popular press)

Migration costs (II)

Structure of migration costs from Ecuador to the United States and Spain (migrants aged 16-49 in 1998 who left between 1999 and 2005; log model)



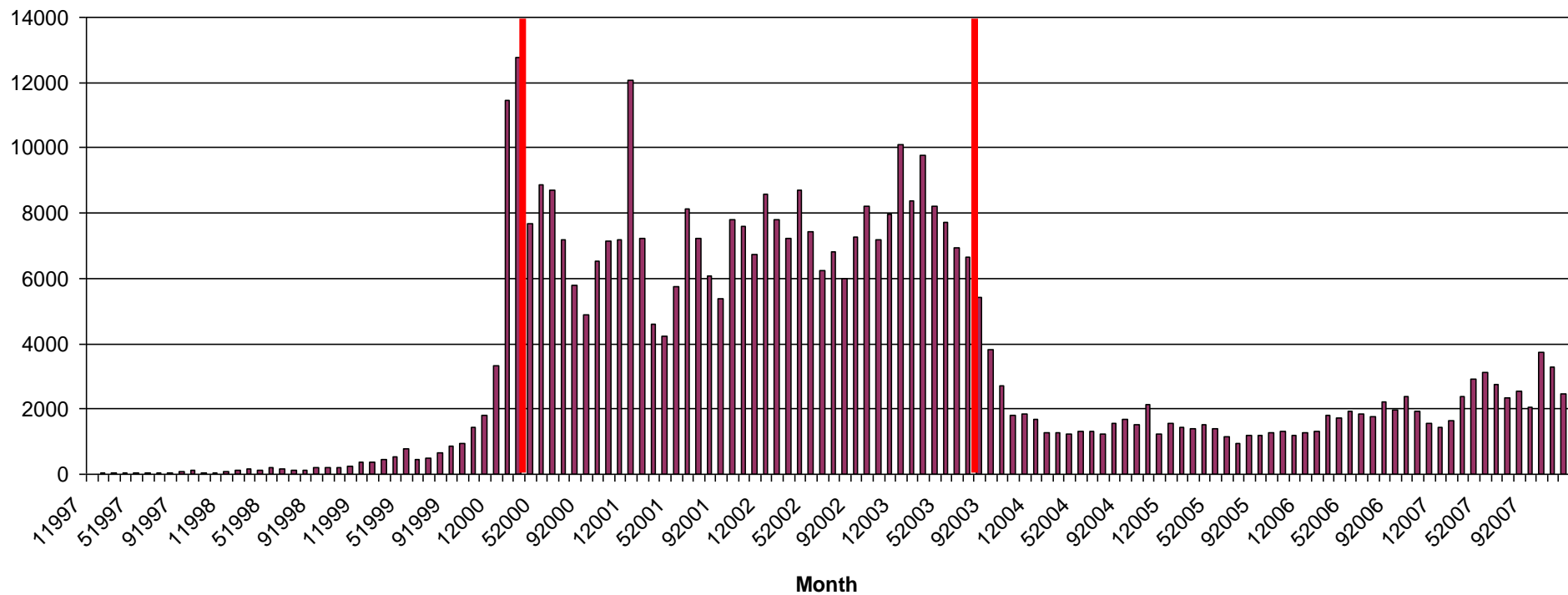
Discussion

- Migration costs substantially larger than real costs:
 - Larger to the US than to Spain
 - Larger for males than for females
 - Larger for unskilled than skilled in Spain; larger for skilled than unskilled in the US
- Why is the cost of migrating larger to the US than to Spain?

Migration policies in Spain and Monthly Immigration

- April 2000: registration implies access to the health and educational system
- August 2003: Ecuadorians need a visa to enter Spain

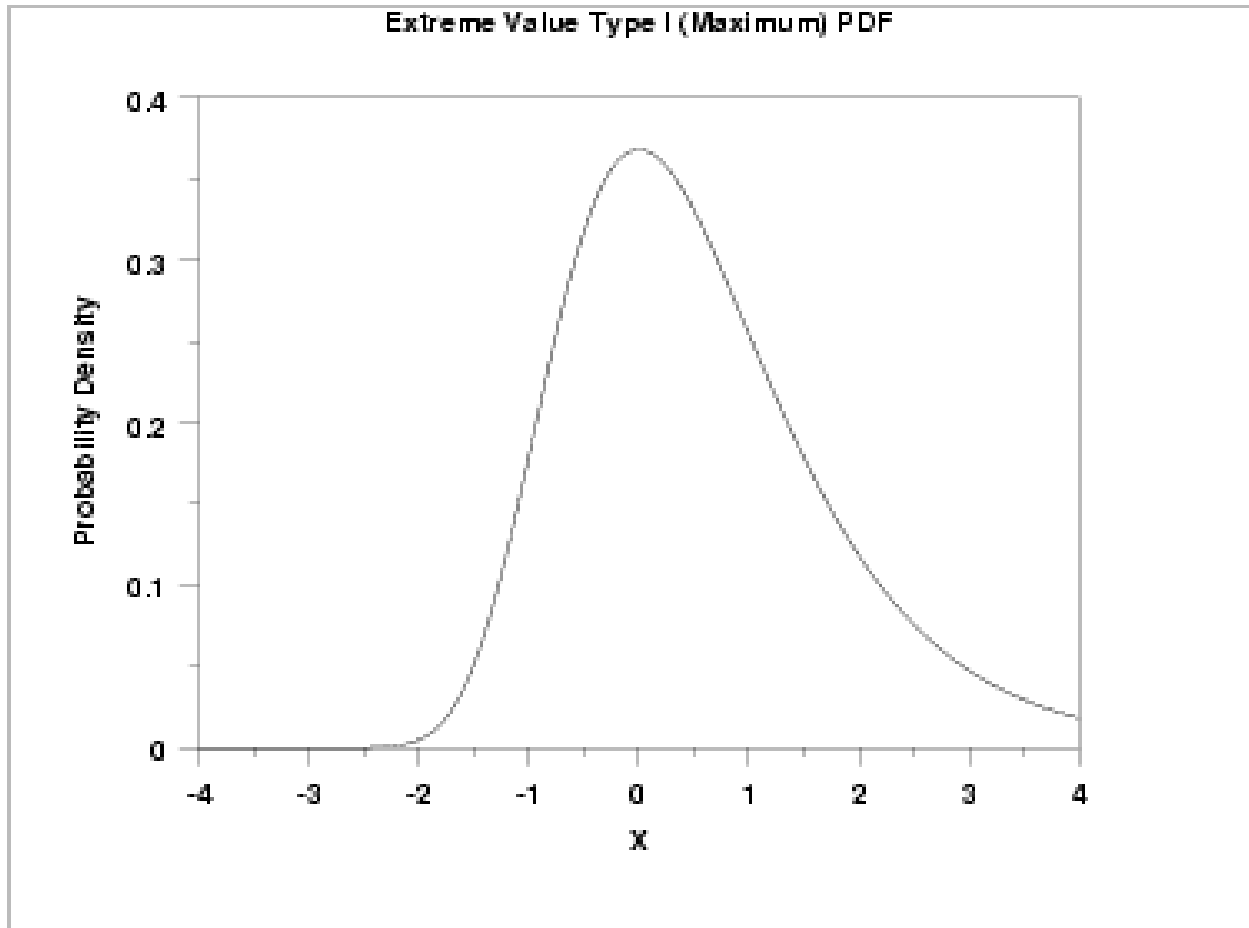
Inflows of Ecuadorians by initial registry



Conclusion

- Logically consistent estimation of an international migration model controlling for selection
- In the worst possible scenario, income still guides migration decisions but...
- Given enough degrees of freedom on the structure of migration costs, any theory based on income maximization can be validated by the data
- Future tests of theories must be based on migration costs

Gumbel Distribution



$$F(x) = e^{-e^{-x}}$$



Implementation (Step 1)

1. Correct the selection bias: use Dahl's (2002) method:
 - Divide the population in mutually exclusive cells
 - Stayers: 48 cells (2 education, 2 gender, 3 age groups, 2 marital status and 2 household size groups)
 - Migrants: 8 cells (2 education, 2 gender and 2 household size groups)
 - Compute proportion of individuals to each destination for each cell
 - Build a correction polynomial out of these proportions: use a second order polynomial in the retention probability for stayers and a second order polynomial in the retention and first-best probability for migrants

Implementation (Step 2)

2. Predict wages. To avoid the loss of non-employed observations and prevent further biases, apply Heckman selection model
 - Add Dahl's correction polynomial to the employment probit in Heckman's method:
 - Employment regression variables: education, age and its square, marital status, household size to proxy for children and Dahl's polynomial
 - Wage regression: same variables except for household size and the correction polynomial but adding Heckman's selection term



Implementation (Step 3)

3. Discrete choice location model: use predicted wages plus all controls from previous equations (to identify how they relate to migration costs)
 - Problem: identification of the wage coefficient from controls used to estimate predicted wages. Two solutions:
 - Use non-linearities created by the selection procedure
 - Add controls to the wage equation (i.e. occupation)
 - Three error structures (unobserved individual heterogeneity):
 - Extreme Value Type I: conditional logit model (IIA)
 - Generalized Extreme Value: nested logit model. It allows for correlation across migrant destinations
 - Multivariate Normal: multinomial probit model. It also allows for heteroskedasticity



Descriptive statistics

All individuals 16-64

	Stayers			United States			Spain		
	mean	std.dev.	s.e	mean	std.dev.	s.e	mean	std.dev.	s.e
female, share	0,51	0,50	0,00	0,45	0,50	0,02	0,54	0,50	0,02
Males									
Age at migration	35,02	13,61	0,12	26,92	9,51	0,56	27,70	8,95	0,54
years since migration				5,30	2,08	0,13	6,12	1,42	0,08
college grad., share	0,11	0,31	0,00	0,12	0,32	0,02	0,08	0,27	0,01
Labor income, 2005 USD	2.873	4.926	52	25.056	18.774	1.312	15.728	4.454	272
Females									
Age at migration	35,39	13,34	0,12	28,92	10,31	0,62	27,52	9,02	0,58
years since migration	0,00	0,00	0,00	5,34	1,96	0,13	5,82	1,44	0,10
college grad., share	0,10	0,30	0,00	0,19	0,39	0,03	0,13	0,34	0,02
Labor income, 2005 USD	1.213	2.815	30	17.071	12.074	1.206	10.567	3.630	222
obs		41.355			662			1124	
Source		ENEMDU			ACS			ENI	
Year		2005			2007			2007	



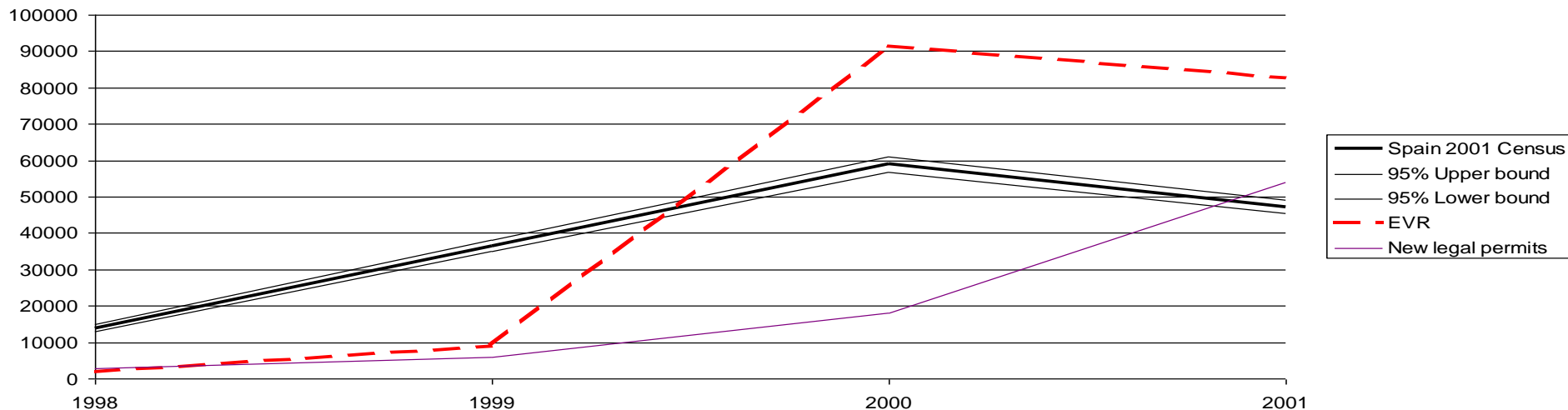
Returns to skill

	Ecuador		USA		Spain	
	men	women	men	women	men	women
cog	0,669 [0.0274]***	0,854 [0.0301]***	0,346 [0.109]***	0,332 [0.122]***	-0,062 [0.049]	-0,023 [0.049]
age	0,0439 [0.00880]***	0,0038 [0.0134]	-0,039 [0.049]	0,007 [0.049]	-0,008 [0.017]	0,003 [0.028]
age2	-0,000509 [0.000111]***	-4,18E-05 [0.000170]	0 [0.001]	0 [0.001]	0 [0.000]	0 [0.000]
ysm			0,053 [0.020]***	0,025 [0.029]	0,018 [0.010]*	0,003 [0.015]
rural	-0,412 [0.0153]***	-0,524 [0.0260]***				
Constant	7,219 [0.169]***	7,544 [0.256]***	10,458 [0.949]***	9,347 [0.886]***	9,678 [0.332]***	9,069 [0.500]***
Observations	10679	6353	188	136	378	373
R-squared	0,25	0,27	0,08	0,05	0,02	0,02

Ecuadorians in the Registry



The Ecuadorian Immigration Boom to Spain in different sources: Spain 2001 Census, New Registry Inscriptions (EVR) and legal residents



Inflows of Ecuadorians by initial registry

