Validation of IPUMS International Industry and Education Data

Lutz Hendricks
UNC at Chapel Hill

March 1, 2010

Abstract

This document collects validation information for select samples of the IPUMS International dataset. The focus is on industry and education data.

1 Introduction

The IPUMS International project (Minnesota Population Center, 2009) collects and harmonizes international micro census data. The database currently covers 44 countries, 130 censuses, and more than 270 million persons. The purpose of this document is to investigate the quality of this data. I focus on education and industry data. I only examine samples that include potentially useful data for both variables.

2 Industry Employment Shares

This section compares industry employment shares with the World Development Indicators 2009 (World Bank, 2009) and with the ILO's Labosta database (International Labour Office, 2009). I find that the following samples contain major problems: COL2005, PHL2000, and VNM1989. ROM has problems within the services category.

2.1 Comparison with World Development Indicators 2009

The World Development Indicators (WDI) report the fraction of a country's employment in agriculture, services, and industry. The underlying data come from the ILO. Below I compare the IPUMS data directly with ILO Laborsta data, which have finer industry detail, but cover fewer countries.

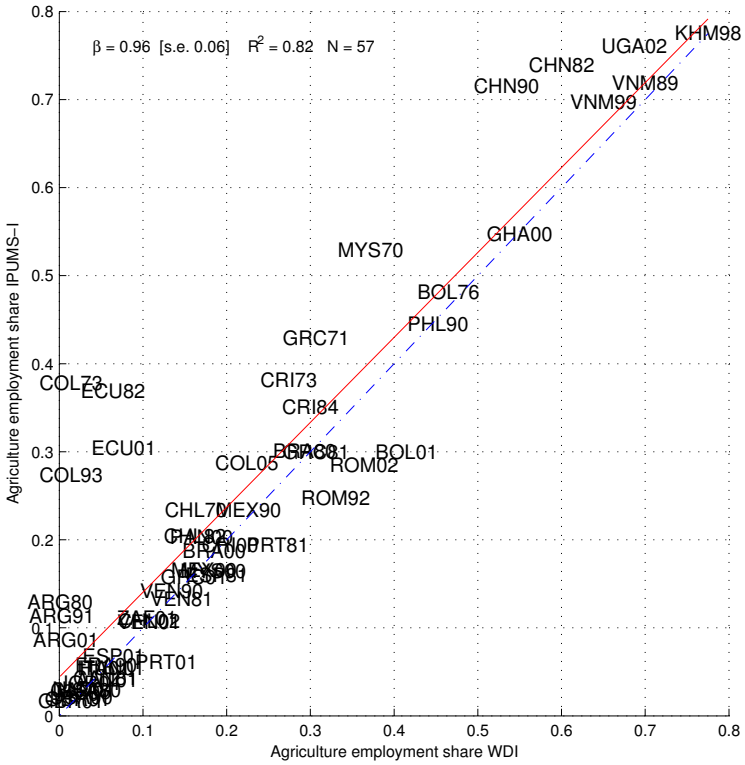

Figure 1: Employment shares in agriculture

Figures 1 through 3 compare the WDI data with employment shares calculated from IPUMS.[1] Details of the IPUMS calculations are as follows:

- The industry variable is INDGEN, which roughly corresponds to 2 digit ISIC-3 codes.
- Agriculture is code 10. Services are codes 60 through 120. Industry contains codes 20 through 50.
- Employment is the total weight of persons reporting each industry.
- For each IPUMS sample the WDI sample with the closest year within a 10 year window is used.

For most countries, the IPUMS employment shares lines up well with WDI. The main exceptions are ARG, COL and ECU with implausibly low agriculture shares in WDI, which are offset by higher shares in services. The likely reason is that the ILO collects urban samples for these countries. The WDI database is based on ILO data.

CHN and MYS also have higher agriculture shares in IPUMS. The reason is unclear.

[1]Most figures also show 45 degree lines (dashed) and OLS regression lines (solid).

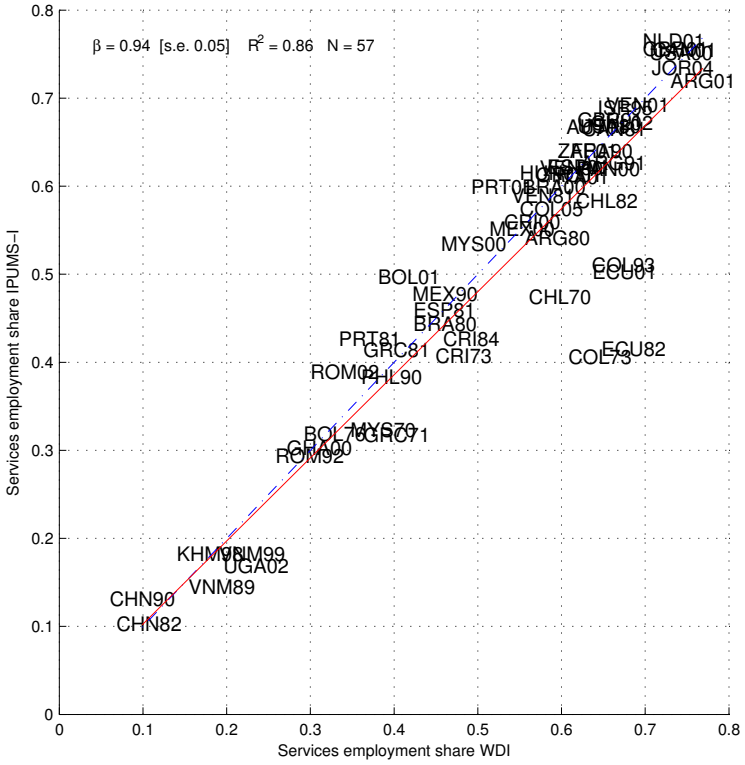

Figure 2: Employment shares in services

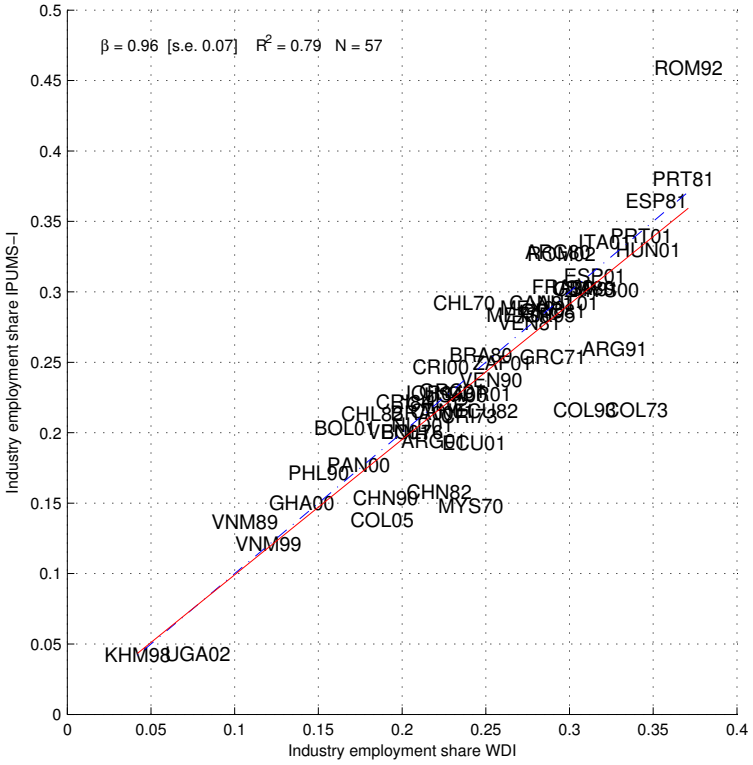

Figure 3: Employment shares in industry

2.2 Comparison with ILO Laborsta.

The ILO Laborsta database reports employment by ISIC industry for a large number of countries. The data cover the period 1960-2008. However, most low income countries do not have data prior to at least 1980.

Calculation of IPUMS employment shares:

- The industry variable is INDGEN.
- Employment is the total weight of all persons reporting a given INDGEN category.
- Only persons who report being employed (according to EMPSTAT or CLASSWKR) are counted.
- Fisheries are added to agriculture for consistency with the ISIC-3 classification.

Calculation of Laborsta employment shares:

- The data are taken from table 2B: "Total employment, by economic activity."
- The ISIC-3 industry classification is used, which maps directly into the INDGEN classification.
- For each IPUMS sample, the nearest year within a 10 year window is used.
- Some countries have data from multiple sources in Laborsta. Where available, the data are taken from labor force surveys or from official estimates. Only if neither is available do I use population census data. This is to avoid using potentially the same source as the data underlying IPUMS.
- For KHM1998 Laborsta reports data that differ greatly from WDI and IPUMS. I drop KHM1998 for this reason.
- Laborsta also has implausible values for COL in some industries. It is kept in the comparison below.
- All countries with more than 15% missing values in Laborsta are deleted.
- Laborsta has urban samples for ARG, COL, ECU. This might account for some of the discrepancies relative to IPUMS.

For each IPUMS sample, figure 4 compares the employment share in agriculture with Laborsta data. The only major discrepancies occur for countries with urban samples in Laborsta. This is consistent with the WDI data shown in figure 1.

Figures 5 through 15 show the shares employed in other industries. These are expressed as fractions of non-agricultural employment. A discussion of major discrepancies by industry follows:

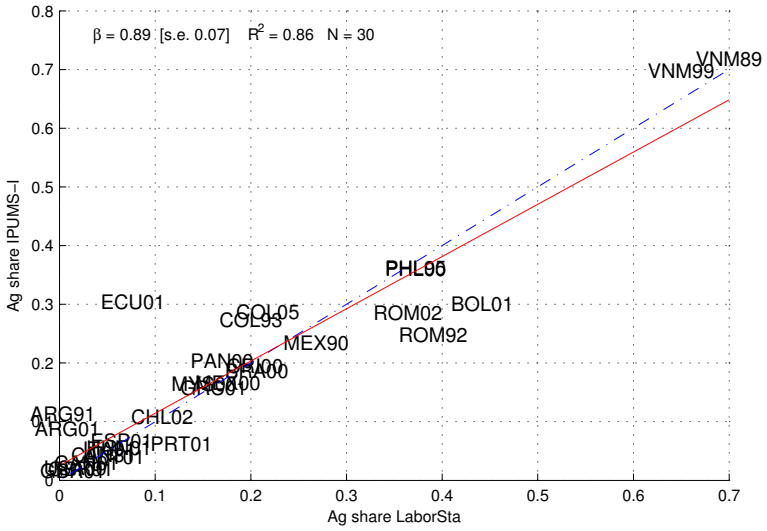

Figure 4: Employment shares in agriculture

1. Manufacturing: VNM1989 is implausible when compared with Laborsta and with VNM1998. The share for ROM1992 is high relative to Laborsta and ROM2002.

2. Electricity, gas, water: CRI2000 is the only major discrepancy, but with a very small share.

3. Construction: The share for PHL2000 is implausibly high compared with Laborsta and PHL1995. The correlation between the two datasets is not very high, but the absolute deviations in employment shares are tolerable.

4. Wholesale/retail: employment shares are generally somewhat lower in IPUMS. The share for COL2005 is implausible relative to Laborsta and COL1993. Other discrepancies are PHL, VNM, MEX2000, USA2000. It is not clear which dataset is more accurate.

5. Hotels/restaurants: There are some outliers, but the employment shares generally small. The share for the USA is low. This is also true in other years, except 2005 (not shown). The reason, according to IPUMS, is that restaurants are not counted in this category. The share for VNM1989 is implausibly low relative to Laborsta and VNM1998. Discrepancies with unknown reasons: PHL, ECU2001, MEX1990, ROM2002.

6. Transport/comm: PHL2000 and ROM2002 are clearly incorrect in IPUMS. COL shares differ from Laborsta. However, Laborsta contains implausible shares for some Colombian industries.

7. Financial services: no major discrepancies. CAN is higher in IPUMS, perhaps because real estate services are included.

8. Public administration: the correlation between datasets is low.

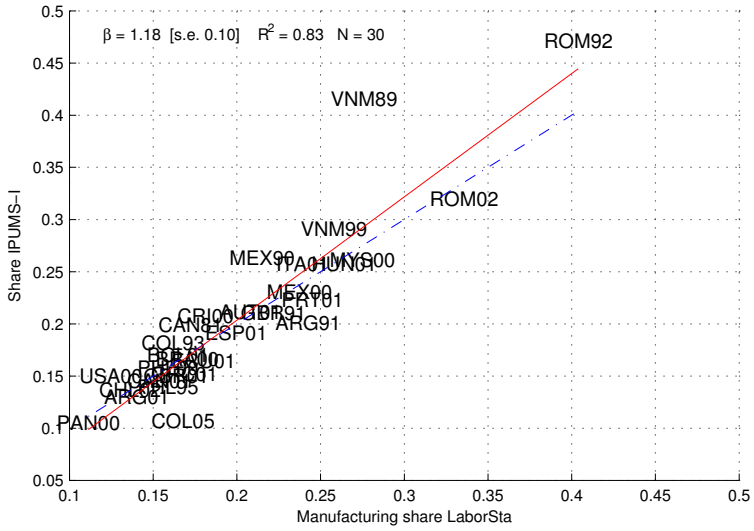

Figure 5: Employment shares in manufacturing

(a) The ARG1991 share is implausible in Laborsta compared with ARG2001.

(b) The share for PHL1995 likely includes Education. The share for PHL2000 is implausibly low.

(c) Discrepancies for unknown reasons can be seen for ROM1992, GRC2001. Coverage is a possible cause. Some samples in Laborsta exclude armed forces.

9. Real estate and business services:

   (a) The share for CAN is low in IPUMS, even though real estate is included.

   (b) CRI2000 has a near zero share in Laborsta. Likely misclassification.

   (c) MEX1990 shows a large fraction in other services and little in real estate, suggesting misclassification. IPUMS counts are consistent with the underlying variable.

   (d) The share for PHL1995 is high. A likely reason is that the category "other business services" in the original variable is broader than business services.

   (e) Also different for unknown reasons: ARG1991, USA2000.

10. Education: Discrepancies in the problem sample VNM1989. Also for unknown reasons: ARG2001, ROM1992, USA2000.

11. Health and social work: No major outliers, except COL, which is clearly wrong in Laborsta.

Time series changes in employment shares. An attempt to validate the IPUMS employment shares internally by looking for implausible changes over time did not uncover and problems other than those already noted.

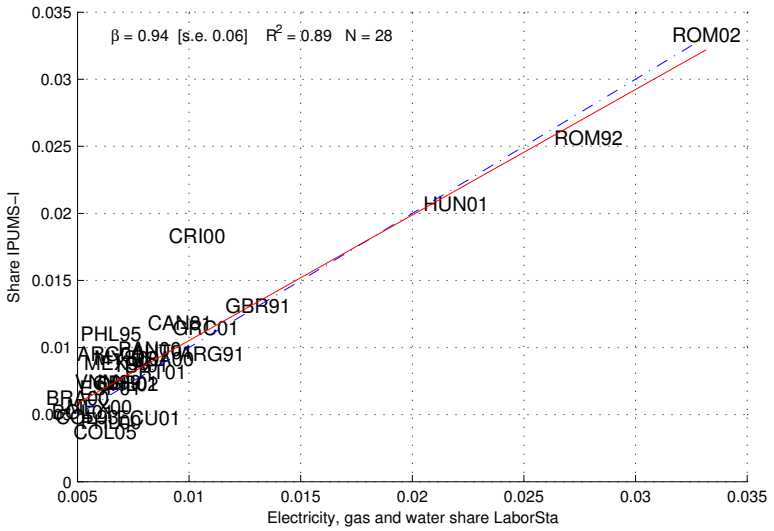

Figure 6: Employment shares in Electricity, Gas, Water

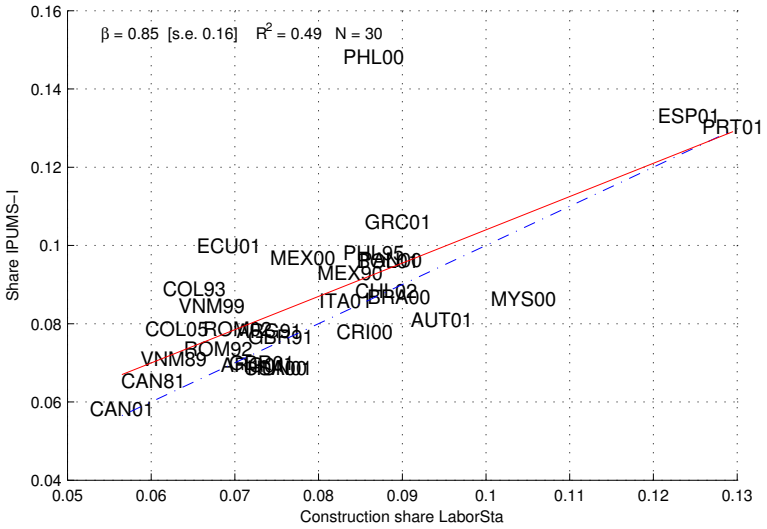

Figure 7: Employment shares in Construction

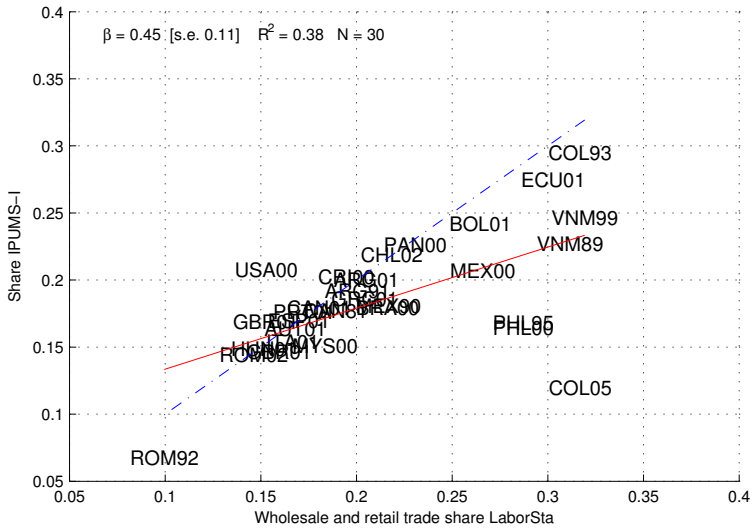

Figure 8: Employment shares in Wholesale / retail

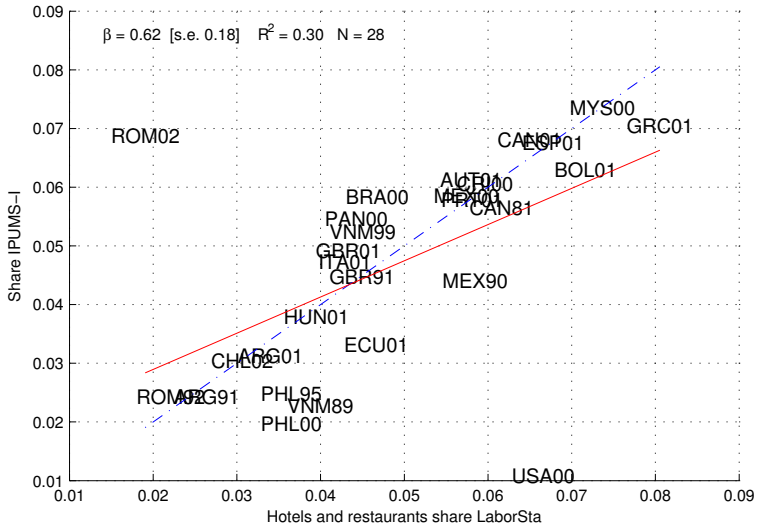

Figure 9: Employment shares in Hotels and Restaurants

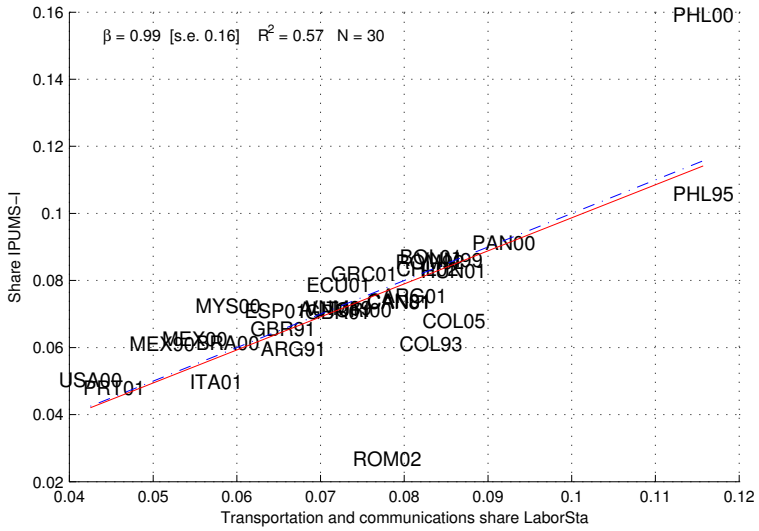

Figure 10: Employment shares in Transport and Communications

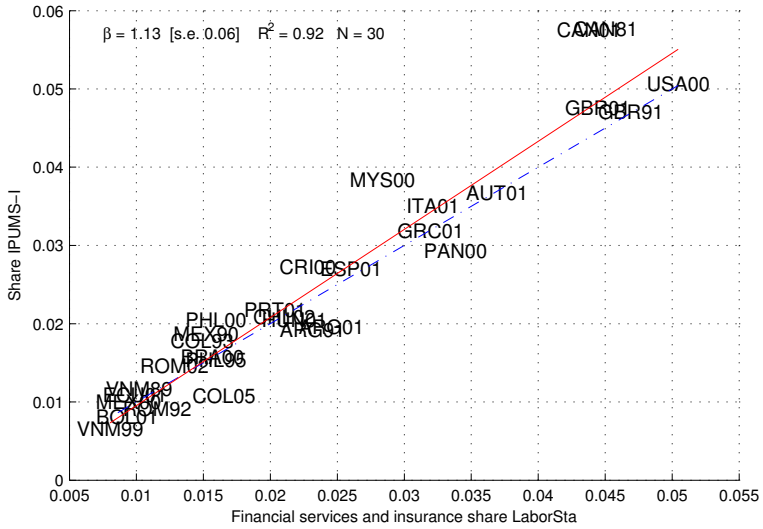

Figure 11: Employment shares in Financial Intermediation

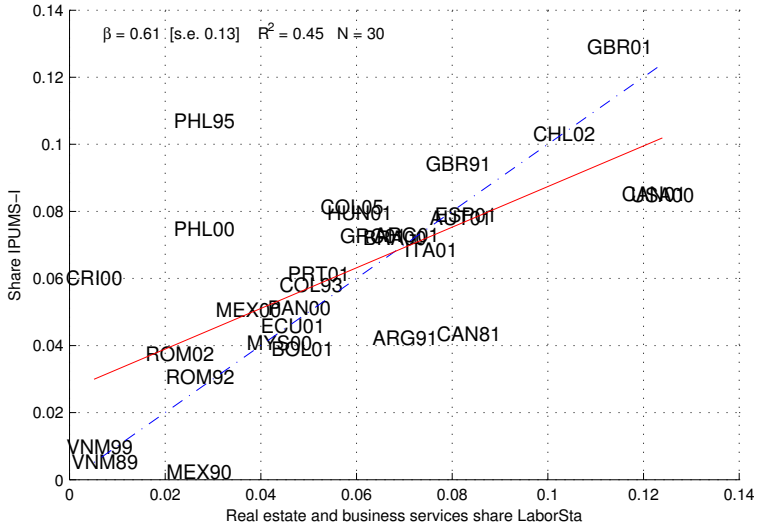

Figure 12: Employment shares in Real Estate, Renting and Business Activities

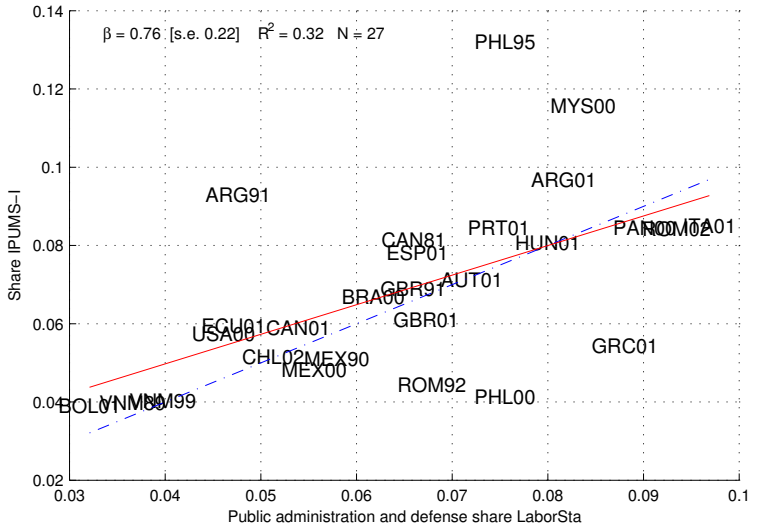

Figure 13: Employment shares in Public Administration and Defense; Compulsory Social Security

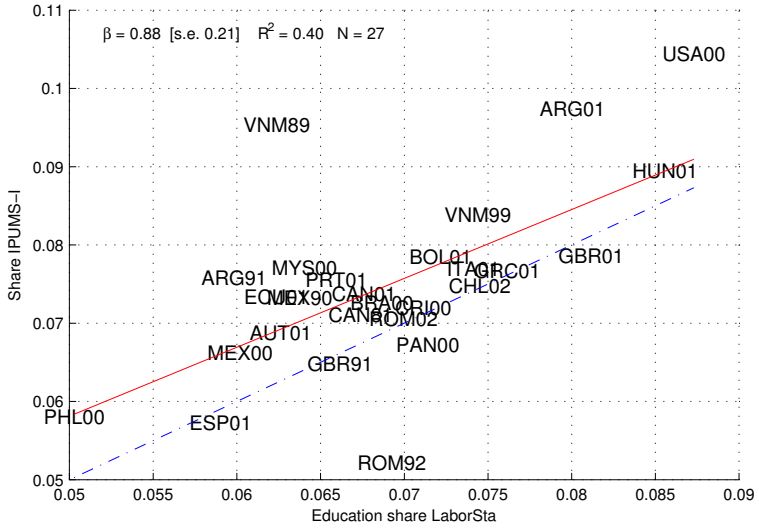

Figure 14: Employment shares in Education

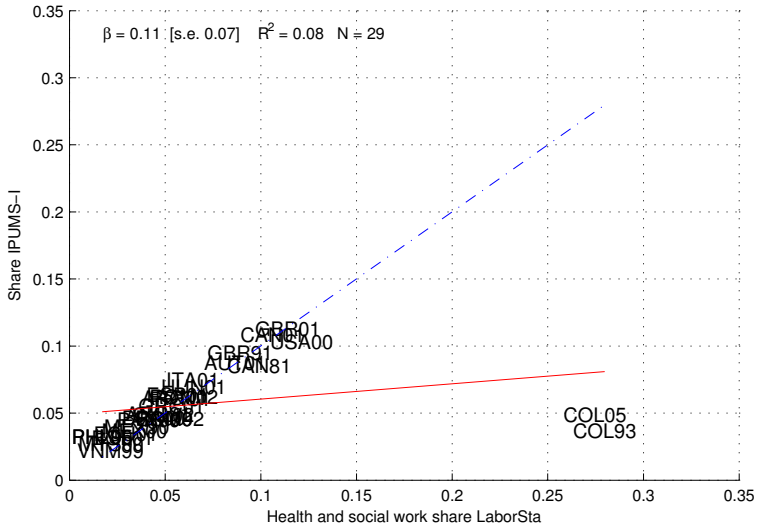

Figure 15: Employment shares in Health and Social Work

Missing values. Four samples contain more than 10% missing INDGEN values among those employed according to EMPSTAT: COL1993, COL2005, ISR1995, PHL1990. For COL2005, more than 50% of observations are missing.

Other issues.

- COL1973 shows almost no employment in some services, but high employment in domestic services. Possible misclassification.
- USA has low shares in hotels and restaurants.

3 Schooling

This section examines educational attainment data. I find that the following samples have major issues: AUT (all years), CHN1982, GBR (all years), HUN2001, IRQ1997, ISR (all years), NLD (all years), and VEN2001.

3.1 Country notes

For many countries, how the country specific schooling variables are coded into EDATTAND seems questionable. In these cases I recoded the country specific variables based on information about each country's school system.

- AUT: Schooling data for AUT are not usable. The original variable lacks detail and contains implausible case counts (e.g., fewer than 5% of persons holding a college degree in 2001).
- IRQ1997: The data contain 43% missing values.
- ISR: The data for ISR contain inconsistencies (confirmed by IPUMS).
- VEN2001: A large numbers of persons reports "post-secondary, 0 years completed." It is not clear how to classify these.
- Some samples do not distinguish no school / some primary (ARG1970, ARG1980, CAN2001, FRA (all years), GRC1971). This is not an important problem for my purposes. For PRT I treat "illiterate" as "no school."
- Countries that lack education detail are CHN1982, GBR, and NLD.

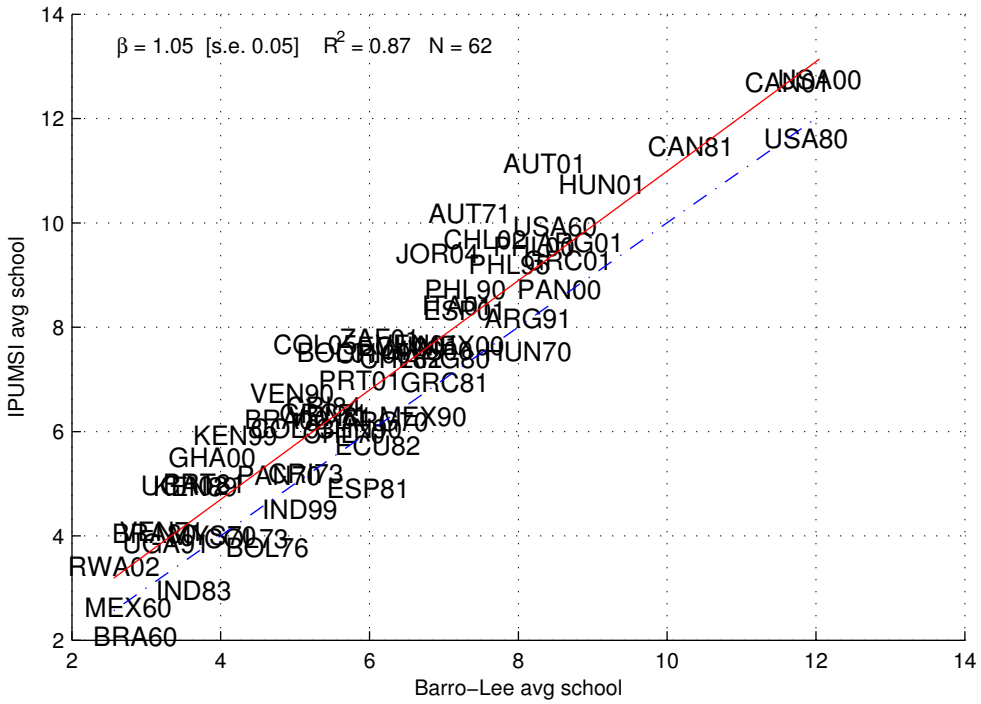

Figure 16: Average years of schooling

3.2 Comparison with Barro and Lee (2001)

Barro and Lee (2001, hereafter BL) calculate educational attainment for 142 countries based on census and school enrollment data. Figure 16 compares average years of schooling in the population aged 15+.[2] Generally, average years of schooling in IPUMS are higher by about one year compared with BL. The correlation between the two samples is high (an unweighted OLS regression results in $R^2 = 0.87$). No large outliers are visible. The largest deviations appear for AUT, which has unusable schooling data in IPUMS.

Figure 17 compares the fraction of the population aged 15+ with each of four schooling levels: no schooling, some or completed primary, secondary, and tertiary. Samples with questionable data (see above) are included. The correlations between the datasets are overall high, but some large deviations are apparent. Tables 1 through 5 show detailed data for each sample.

For a number of samples, the datasets differ for unknown reasons (CAN2001, CHL1970, CHL2002, COL (all years), CRI1973, ECU1982, GHA2000, JOR2004, KEN1999, MYS2000, PAN1970, PRT2001, RWA2002, UGA2002, ZAF2001). It is not clear which data source is more reliable.

[2]WDI reports the fraction of the labor force with primary, secondary, and tertiary education. However, the data contain large, obvious errors, such as 10 percentage point jumps in the fraction with a given schooling level from one year to the next.

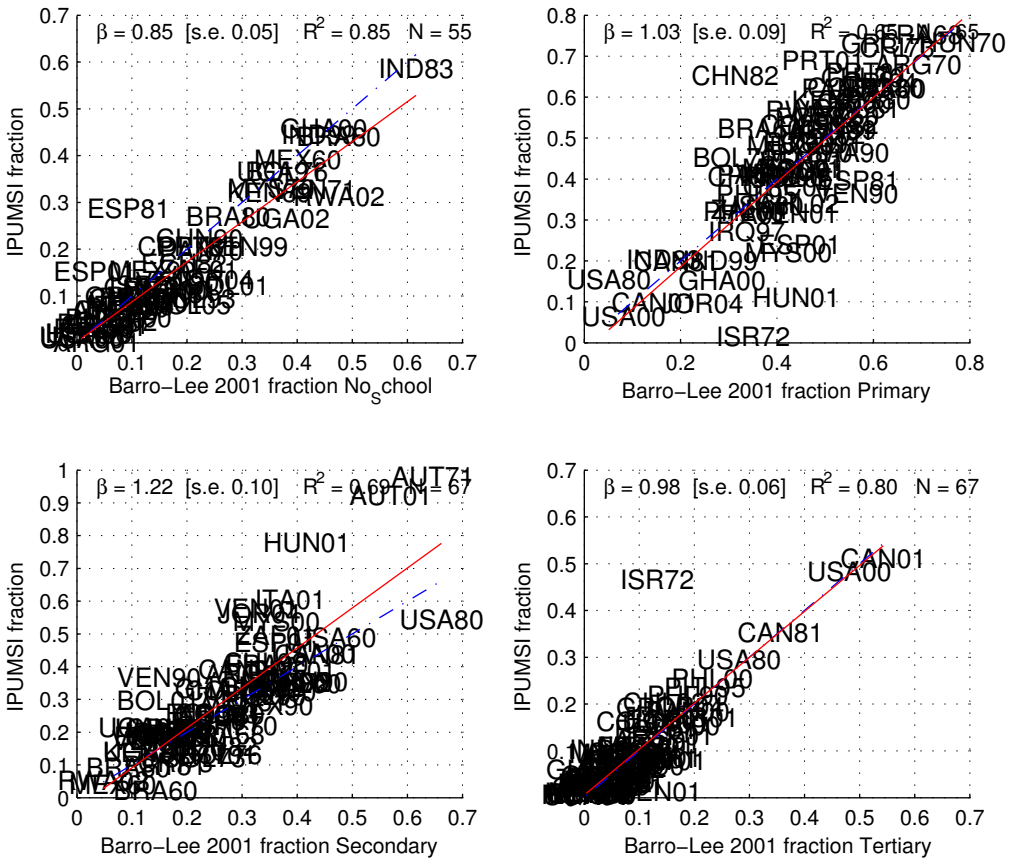

Figure 17: Educational attainment

Table 1: Educational attainment: Comparison with Lee and Barro (2001)

Sample	Missing	No school	Primary	Secondary	Tertiary
USA 1960	0.0	2.0	33.6	48.7	15.7
		2.0	36.8	46.7	14.5
		-0.0	-3.2	2.0	1.2
USA 1980	0.0	0.8	15.3	54.3	29.5
		0.9	5.1	66.2	28.1
		-0.1	10.2	-11.9	1.4
USA 2000	0.0	1.3	6.4	43.9	48.3
		0.8	8.2	42.9	48.1
		0.5	-1.8	1.0	0.2
ARG 1970	2.3	7.1	69.7	19.1	4.1
		7.0	69.3	19.3	4.4
		0.1	0.4	-0.2	-0.3
ARG 1980	5.2	0.0	65.8	26.8	7.4
		5.7	61.4	26.2	6.7
		-5.7	4.4	0.6	0.7
ARG 1991	1.1	3.5	52.5	31.8	12.2
		4.8	51.2	30.7	13.2
		-1.3	1.3	1.1	-1.0
ARG 2001	3.5	0.0	43.3	39.0	17.6
		3.6	45.2	31.1	20.1
		-3.6	-1.9	7.9	-2.5
AUT 1971	0.0	0.0	0.0	97.8	2.2
		3.4	29.8	64.6	2.2
		-3.4	-29.8	33.2	-0.0
AUT 2001	0.0	0.0	0.0	92.1	7.9
		2.8	25.5	57.0	14.7
		-2.8	-25.5	35.1	-6.8
BOL 1976	2.1	37.4	46.2	13.5	3.0
		38.1	31.1	26.1	4.7
		-0.7	15.1	-12.6	-1.7
BOL 2001	1.2	12.3	43.1	30.0	14.7
		27.4	44.8	14.8	13.0
		-15.1	-1.7	15.2	1.7
BRA 1960	0.6	44.3	52.7	2.1	0.8
		47.5	36.4	14.3	1.8
		-3.2	16.3	-12.2	-1.0
BRA 1980	0.1	27.0	59.5	9.3	4.3
		27.5	59.0	9.3	4.3
		-0.5	0.5	-0.0	-0.0
BRA 2000	0.0	11.2	62.1	18.8	7.9
		16.0	62.2	14.4	7.5
		-4.8	-0.1	4.4	0.4

Table 2: Educational attainment: Comparison with Lee and Barro (2001)

Sample	Missing	No school	Primary	Secondary	Tertiary
CAN 1981	1.4	0.0	20.0	44.3	35.7
		1.6	19.1	43.6	35.6
		-1.6	0.9	0.7	0.1
CAN 2001	0.0	0.0	9.8	39.1	51.1
		1.4	14.5	29.8	54.3
		-1.4	-4.7	9.3	-3.2
CHL 1970	0.0	9.5	64.8	21.9	3.8
		9.6	57.6	28.9	3.8
		-0.1	7.2	-7.0	-0.0
CHL 1982	0.0	6.9	53.3	33.5	6.3
		6.9	52.5	33.5	7.1
		0.0	0.8	-0.0	-0.8
CHL 2002	0.0	4.5	33.9	41.2	20.4
		7.1	44.3	34.1	14.5
		-2.6	-10.4	7.1	5.9
CHN 1982	0.0	0.0	65.2	33.8	0.9
		34.0	31.3	33.7	0.9
		-34.0	33.9	0.1	0.0
CHN 1990	0.0	22.2	40.4	35.3	2.1
		22.2	34.6	41.3	1.9
		-0.0	5.8	-6.0	0.2
COL 1973	2.3	21.1	63.9	12.5	2.6
		18.6	55.0	23.0	3.3
		2.5	8.9	-10.5	-0.7
COL 1993	2.0	9.6	54.3	26.2	9.8
		21.2	45.2	25.4	8.3
		-11.6	9.1	0.8	1.5
COL 2005	2.6	8.4	41.4	33.6	16.6
		20.3	42.8	27.2	9.8
		-11.9	-1.4	6.4	6.8
CRI 1973	0.0	11.8	72.0	10.7	5.4
		11.7	64.9	17.9	5.4
		0.1	7.1	-7.2	0.0
CRI 1984	0.0	8.0	63.4	19.4	9.2
		10.8	61.3	17.4	10.6
		-2.8	2.1	2.0	-1.4
CRI 2000	0.0	5.4	58.3	20.9	15.4
		10.4	56.0	15.7	17.8
		-5.0	2.3	5.2	-2.4
ECU 1982	6.1	18.0	58.6	16.5	6.9
		19.4	48.4	25.0	7.2
		-1.4	10.2	-8.5	-0.3
ECU 2001	0.5	8.5	49.1	25.7	16.7
		15.1	46.0	23.8	15.1
		-6.6	3.1	1.9	1.6

Table 3: Educational attainment: Comparison with Lee and Barro (2001)

Sample	Missing	No school	Primary	Secondary	Tertiary
ESP 1981	0.2	28.8	39.8	23.4	8.0
		9.4	56.8	25.8	8.0
		19.4	-17.0	-2.4	-0.0
ESP 2001	0.5	15.2	24.0	47.4	13.4
		3.3	44.7	36.0	16.0
		11.9	-20.7	11.4	-2.6
FRA 1968	2.6	0.0	77.2	19.6	3.1
		0.8	69.9	26.6	2.7
		-0.8	7.3	-7.0	0.4
FRA 1990	2.1	0.0	47.5	41.3	11.2
		1.0	54.7	34.5	9.8
		-1.0	-7.2	6.8	1.4
GHA 2000	0.0	45.7	15.1	33.3	5.9
		44.8	28.6	25.7	1.0
		0.9	-13.5	7.6	4.9
GRC 1971	4.1	0.0	76.2	18.9	5.0
		16.6	62.5	17.1	3.8
		-16.6	13.7	1.8	1.2
GRC 1981	0.0	9.8	57.3	24.8	8.2
		9.5	56.4	25.5	8.7
		0.3	0.9	-0.7	-0.5
GRC 2001	0.0	4.0	38.0	38.7	19.3
		5.1	42.1	39.0	13.8
		-1.1	-4.1	-0.3	5.5
HUN 1970	0.0	2.0	73.3	21.0	3.6
		2.1	78.4	14.8	4.9
		-0.1	-5.1	6.2	-1.3
HUN 2001	0.0	0.0	11.1	77.9	11.0
		2.4	43.9	41.7	12.1
		-2.4	-32.8	36.2	-1.1
IND 1983	0.1	58.7	20.3	18.6	2.4
		61.6	16.5	19.0	2.8
		-2.9	3.8	-0.4	-0.4
IND 1999	0.1	44.6	20.0	25.1	10.2
		43.9	28.2	23.8	4.1
		0.7	-8.2	1.3	6.1
IRQ 1997	43.7	0.0	48.1	35.4	16.5
		44.3	33.7	15.7	6.4
		-44.3	14.4	19.7	10.1
ISR 1972	4.8	12.9	1.5	36.7	48.9
		12.5	35.0	39.4	13.2
		0.4	-33.5	-2.7	35.7

Table 4: Educational attainment: Comparison with Lee and Barro (2001)

Sample	Missing	No school	Primary	Secondary	Tertiary
ITA 2001	0.0	0.0	31.4	60.3	8.3
		12.4	34.8	38.7	14.2
		-12.4	-3.4	21.6	-5.9
JOR 2004	1.5	13.8	9.6	56.9	19.7
		24.6	24.4	33.1	17.9
		-10.8	-14.8	23.8	1.8
KEN 1989	1.5	32.5	52.8	14.0	0.8
		34.9	51.9	12.4	0.7
		-2.4	0.9	1.6	0.1
KEN 1999	0.0	20.5	59.1	19.3	1.1
		30.7	51.8	16.4	1.1
		-10.2	7.3	2.9	0.0
KHM 1998	32.6	1.4	43.9	54.1	0.6
MEX 1960	0.4	39.2	55.8	4.0	1.0
		40.1	52.2	6.5	1.3
		-0.9	3.6	-2.5	-0.3
MEX 1990	0.0	15.4	48.1	28.2	8.3
		13.7	42.8	35.0	8.5
		1.7	5.3	-6.8	-0.2
MEX 2000	3.5	8.2	44.1	36.4	11.3
		9.7	41.8	37.9	10.6
		-1.5	2.3	-1.5	0.7
MYS 1970	0.0	32.9	46.4	20.0	0.8
		35.3	46.7	16.6	1.5
		-2.4	-0.3	3.4	-0.7
MYS 2000	3.3	11.8	22.8	55.7	9.7
		16.2	42.4	36.2	5.2
		-4.4	-19.6	19.5	4.5
PAN 1970	0.1	19.7	62.1	14.5	3.8
		22.5	53.8	19.4	4.3
		-2.8	8.3	-4.9	-0.5
PAN 2000	0.7	7.6	41.3	32.6	18.6
		9.1	36.2	35.8	18.8
		-1.5	5.1	-3.2	-0.2
PHL 1990	1.6	5.2	41.2	31.6	22.1
		5.2	41.4	34.2	18.8
		-0.0	-0.2	-2.6	3.3
PHL 1995	0.8	4.1	36.5	35.7	23.8
		3.7	35.8	38.4	21.9
		0.4	0.7	-2.7	1.9
PHL 2000	4.8	3.1	33.7	36.5	26.7
		3.1	33.1	40.6	23.2
		0.0	0.6	-4.1	3.5

Table 5: Educational attainment: Comparison with Lee and Barro (2001)

Sample	Missing	No school	Primary	Secondary	Tertiary
PRT 1981	0.0	20.5	65.9	9.9	3.7
		21.9	58.7	16.0	3.6
		-1.4	7.2	-6.1	0.1
PRT 2001	0.0	9.5	68.8	13.1	8.5
		11.5	49.6	25.1	13.8
		-2.0	19.2	-12.0	-5.3
ROM 1992	0.4	4.7	19.9	67.6	7.8
ROM 2002	0.1	4.4	14.9	69.7	11.0
RWA 2002	6.2	33.2	60.3	5.7	0.8
		47.6	47.1	4.8	0.5
		-14.4	13.2	0.9	0.3
UGA 1991	0.4	36.7	48.9	14.1	0.4
		36.9	48.8	12.6	0.4
		-0.2	0.1	1.5	-0.0
UGA 2002	0.0	26.5	50.9	21.5	1.0
		38.0	49.3	11.8	0.8
		-11.5	1.6	9.7	0.2
VEN 1971	6.2	34.8	44.0	18.7	2.5
		43.1	40.0	13.9	2.9
		-8.3	4.0	4.8	-0.4
VEN 1981	9.9	11.7	49.1	32.5	6.7
		17.9	45.1	30.5	6.4
		-6.2	4.0	2.0	0.3
VEN 1990	7.1	14.0	39.2	39.5	7.3
		18.2	56.4	14.9	10.5
		-4.2	-17.2	24.6	-3.2
VEN 2001	0.8	8.8	31.7	58.2	1.3
		9.8	44.0	32.5	13.7
		-1.0	-12.3	25.7	-12.4
VNM 1989	0.3	13.6	64.2	20.3	1.9
VNM 1999	1.1	8.6	62.8	25.9	2.7
ZAF 2001	0.0	15.6	32.5	49.4	2.5
		22.1	34.5	36.1	7.3
		-6.5	-2.0	13.3	-4.8

Likely errors in Barro and Lee (2001):

- BOL2001: BL's data have the unusual feature that the fraction of secondary and higher education is roughly the same (14%). In IPUMS, the pattern is more typical compared with countries at similar levels of development. BOL1976 also differs from BL (fraction primary / secondary).
- BRA1960: BL show falling education (primary and secondary) between 1960 and 1980. IPUMS shows rising education.
- ITA2001: BL's data show 47% with primary or no schooling. This appears high for a country at Italy's level of development. IPUMS data fail to distinguish no school and some primary.
- USA1980: Comparing U.S. data over time suggests that BL assign about 10% of the population to primary instead of secondary education.
- VEN1990: BL show a large rise in the fraction with primary schooling between 1981 and 1990, followed by a large fall between 1990 and 2000. Secondary schooling shows the opposite trend. This may reflect a coding error, which the IPUMS data do not show.

Problems in IPUMS data:

- AUT: see above.
- CHN1982 and GRC1971 do not distinguish no school / primary schooling.
- HUN2001: implausible differences relative to BL and to HUN1970.
- VEN2001: the fraction with tertiary schooling is implausibly low in IPUMS. See notes above.

Correctable coding issues:

- ARG: EDUCAR is inconsistently recoded into EDATTAND. The solution is to recreate EDATTAND from EDUCAR.
- ESP2001: The fractions with primary and secondary schooling differ from BL. A large part is due to the IPUMS definition of primary / secondary schooling for Spain. The original codes call grade 8 primary, while IPUMS codes it as lower secondary. The large fraction of persons without schooling in IPUMS seems questionable, but is a feature of the original data (EDUCES).

4 Other

The fraction of persons residing in an urban area is highly correlated with WDI data ($R^2 = 0.97$), as is the fraction of persons aged 15-64 ($R^2 = 0.99$).

References

Barro, R. J., and J.-W. Lee (2001): "International Data on Educational Attainment: Updates and Implications," *Oxford Economic Papers*, 53(3), 541–563.

International Labour Office (2009): "LABORSTA," Machine-readable database.

Lee, J.-W., and R. J. Barro (2001): "Schooling Quality in a Cross-Section of Countries," *Economica*, 68(272), 465–488.

Minnesota Population Center (2009): "Integrated Public Use Microdata Series - International: Version 5.0," Machine-readable database.

World Bank (2009): "World Development Indicators Online," Machine-readable database.