ECON 184

Statistical analysis

Contents

1 Type of variables

- Continuous: income per year.
- Discrete: years of education.
- Binary (or dummy): gender.
- Ordinal: low, medium, high.

2 Cross tabulations

- Crosstabs allow us to relate a few variables.

Type of Institution	Economic growth (%)
Non-democratic regimes	zz
Democratic regimes	aa

3 Regression analysis

3.1 Tables vs. regression

- Regression analysis permits the analysis of multiple variables.
- Unlike crosstab we can identify the partial effect.
- What is the impact of x on y, after controlling for z?
- Example: x=gender, y=wages and z=experience.

3.2 Ordinary Least Squares (OLS)

- We usually assume that the relation between y and x is linear:

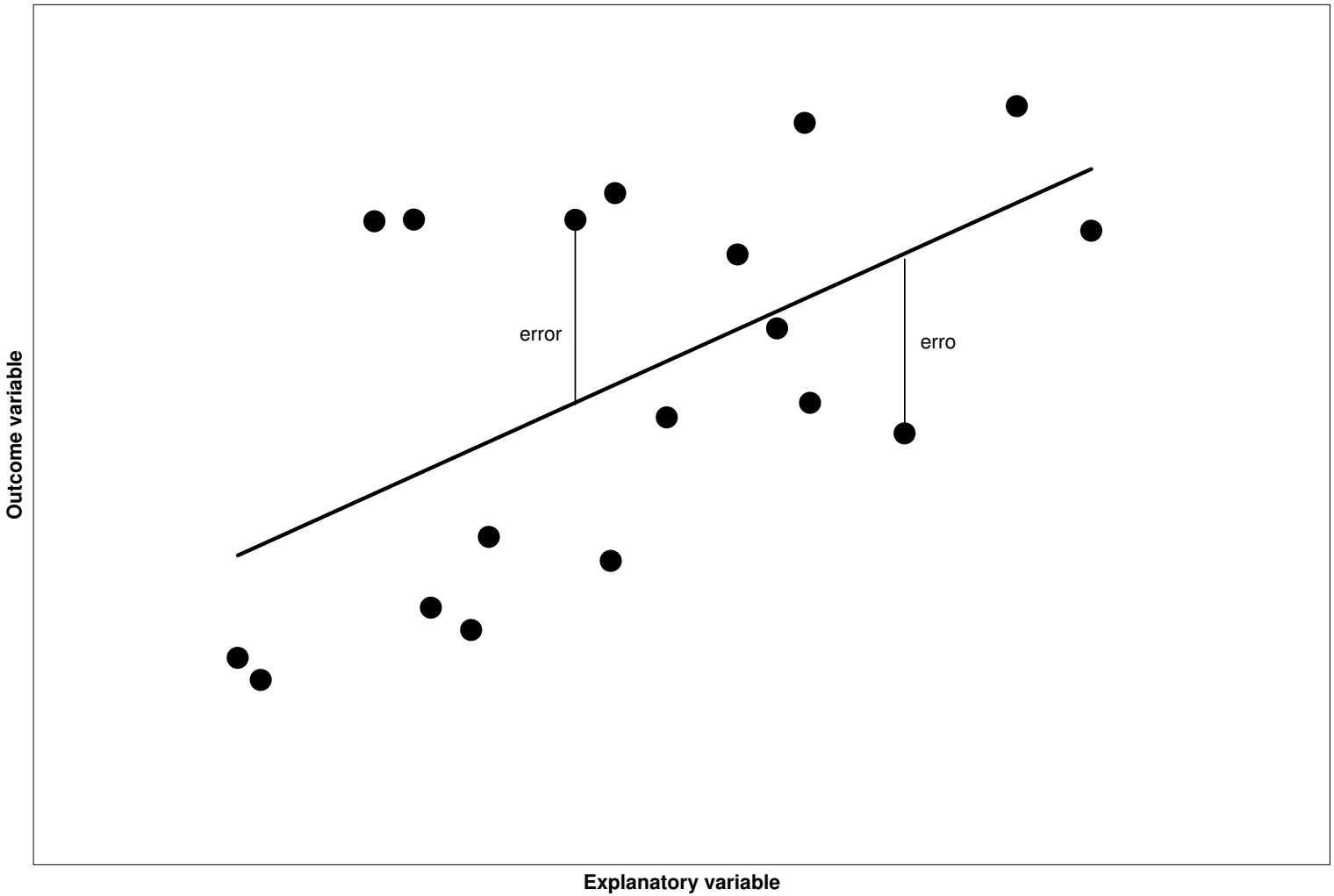
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \tag{1}$$

- Non-linear relation: $y_i = \beta_0 + \beta_1 x_{1i}^{\theta} + e_i$.
- β_0 is the **constant** or **intercept**.
- The **slopes coefficients**: β_1 and β_2.
- β_1 represents the change in y associated to changes in x_1.
- Equation (1) is a simplication of reality where e is the **error term**.

- OLS is a method for obtaining estimates of βs using data on y and x_1, x_2, etc...
- One way to think about OLS is to find βs that minimize the sum of squared deviations

$$\sum_{i}^{N}(y_i - \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})^2$$

Data, fitted line and residual errors



Examples

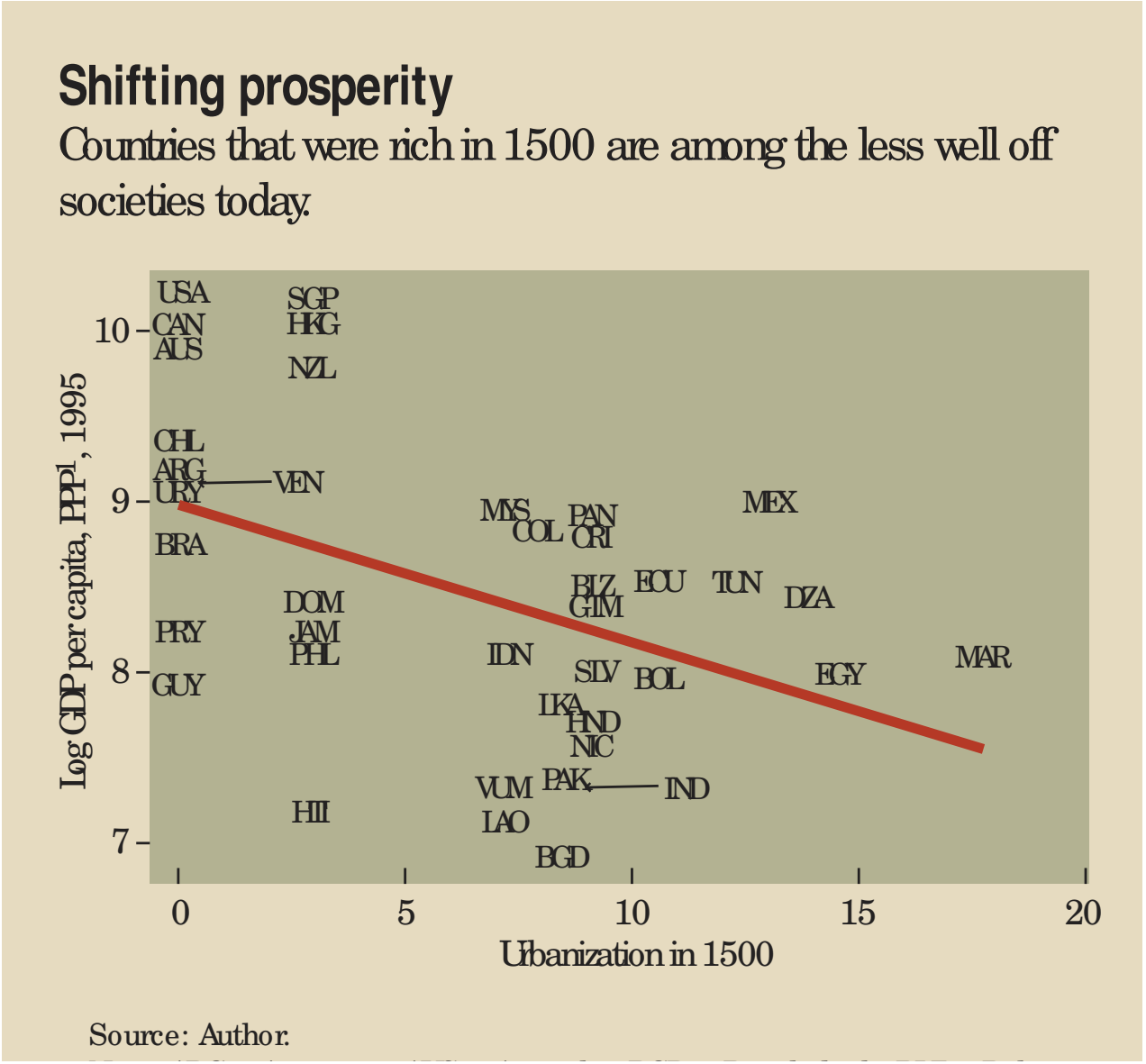
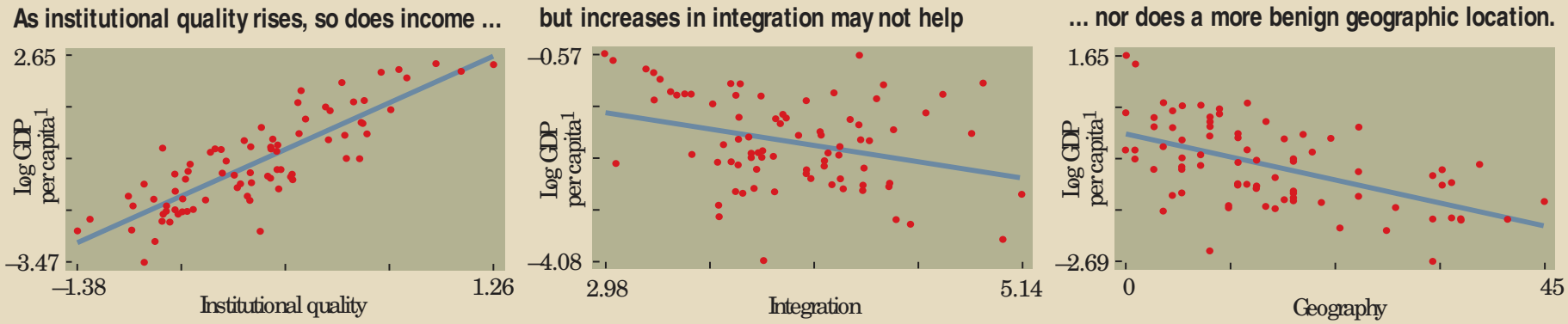

Shifting prosperity
Countries that were rich in 1500 are among the less well off societies today.
Log GDP per capita, PPP1, 1995
10
9
8
7
0
5
10
15
20
Urbanization in 1500
USA
CAN
AUS
SGP
HKG
NZL
CHL
ARG
URY
VEN
BRA
PRY
GUY
DOM
JAM
PHL
HTI
MYS
COL
PAN
CRI
MEX
BLZ
ECU
GTM
TUN
DZA
IDN
SLV
BOL
EGY
MAR
LKA
HND
NIC
VUM
PAK
IND
LAO
BGD
Source: Author.

Chart 2

Institutional quality scores high

Institutional quality can boost income significantly, while global integration and geography, on their own, do not.

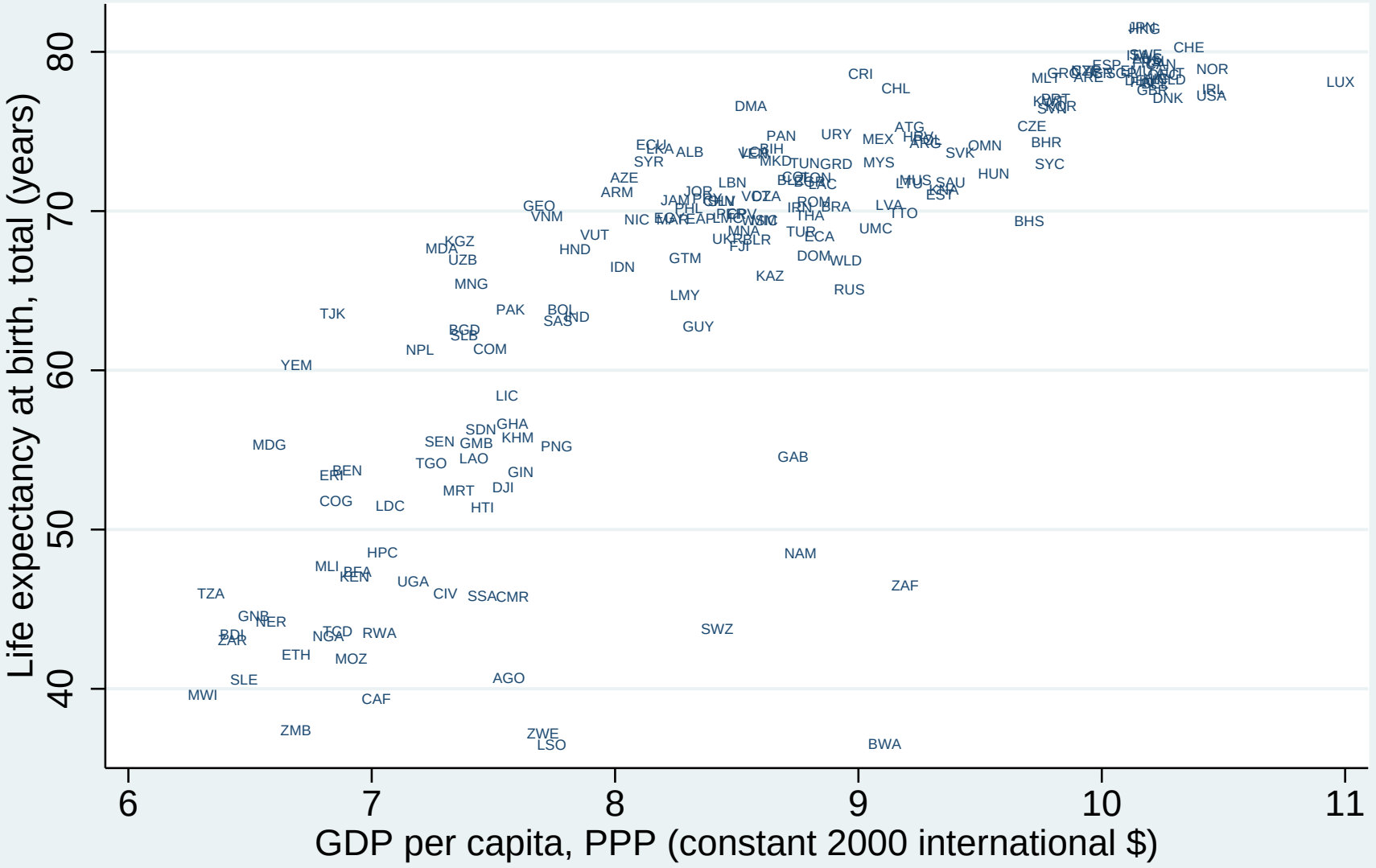


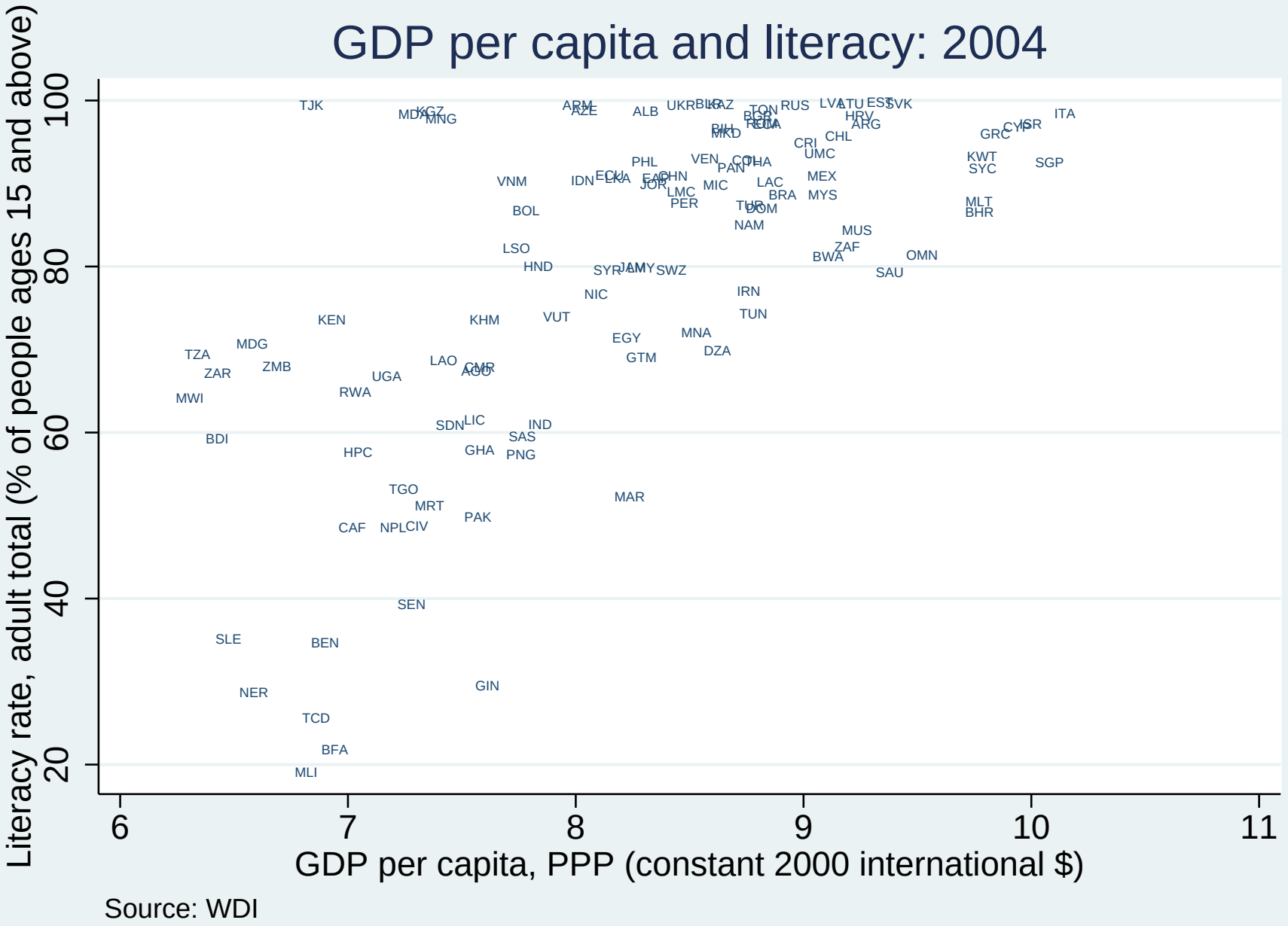
Source: Authors

Note: The graphs capture the causal impact of each of the determinants on income, after controlling for the impact of the others. The indicators of integration and geography used are the ratio of trade to GDP and distance from the equator, respectively. For further details, see Rodrik, Subramanian, and Trebbi (2002).

[1]Expressed in terms of purchasing power parity, 1995.

GDP per capita and life expectancy: 2002




GDP per capita and literacy: 2004
Literacy rate, adult total (% of people ages 15 and above)
20
40
60
80
100
6
7
8
9
10
11
GDP per capita, PPP (constant 2000 international $)
TJK
KGZ
MDA
MNG
ARM
AZE
ALB
UKR
BLR
KAZ
TON
RUS
LVA
LTU
EST
SVK
HRV
ARG
ITA
CYP
ISR
GRC
MKD
CHL
CRI
UMC
KWT
SYC
SGP
PHL
VEN
COL
PAN
THA
VNM
IDN
ECU
LKA
EAP
CHN
JOR
MIC
LAC
MEX
LMC
BRA
MYS
PER
TUR
DOM
MLT
BHR
BOL
NAM
MUS
LSO
ZAF
BWA
OMN
HND
SYR
JAM
SWZ
SAU
NIC
IRN
KEN
KHM
VUT
TUN
EGY
MNA
MDG
DZA
TZA
GTM
ZMB
LAO
CMR
ZAR
AGO
UGA
RWA
MWI
SDN
LIC
IND
BDI
SAS
HPC
GHA
PNG
TGO
MAR
MRT
PAK
CAF
NPL
CIV
SEN
SLE
BEN
GIN
NER
TCD
BFA
MLI

Source: WDI

3.3 How to read a regressions table

- A regression table is where the results of regression analysis are reported.
- It provides estimates from slope coefficients.
- Below is an example from a model that tries to explain economic growth ($y = GR6085, GR7085$) from a sample of countries between 1960 and 1985.

- The explanatory variables (x) are listed here:

Table 2. Definitions of Variables in Tables 1

GR6085 (GR7085): Growth rate of real per capita GDP from 1960 to 1985 (1970 to 1985).
GDP60 (GDP70, GDP85): 1960 (1970, 1985) value of real per capita GDP (1980 base year).
g^c/y: Average from 1970 to 1985 of the ratio of real government consumption (exclusive of defense and education) to real GDP.
SEC50 (SEC60): 1950 (1960) secondary-school enrollment rate.
PRIM50 (PRIM60): 1950 (1960) primary-school enrollment rate.
REV: Number of revolutions and coups per year (1960–1985 or subsample).
ASSASS: Number of assassinations per million population per year (1960–1985 or subsample).
PPI60DEV: Magnitude of the deviation of 1960 PPP value for the investment deflator (U.S. = 1.0) from the sample mean.
AFRICA: Dummy variable for sub-Saharan Africa.
LAT. AMER.: Dummy variable for Latin America.

Source: Barro (1991), p. 439.

Table 1. Regressions for per Capita Growth

	(1)	(2)	(3)	(4)
Dep. var.	GR6085	GR7085	GR6085	GR6085
No. obs.	98	98	98	98
Const.	0.0302	0.0287	0.0288	0.0345
	(0.0066)	(0.0080)	(0.0065)	(0.0067)
GDP60	–0.0075	–0.0089	–0.0073	–0.0068
	(0.0012)	(0.0016)	(0.0011)	(0.0009)
SEC60	0.0305	0.0331	0.0254	0.0133
	(0.0079)	(0.0137)	(0.0110)	(0.0070)
PRIM60	0.0250	0.0276	0.0324	0.0263
	(0.0056)	(0.0070)	(0.0077)	(0.0060)
SEC50	—	—	0.0183	—
			(0.0121)	
PRIM50	—	—	–0.0085	—
			(0.0064)	
g^c/y	–0.119	–0.142	–0.121	–0.094
	(0.028)	(0.034)	(0.027)	(0.026)
REV	–0.0195	–0.0236	–0.0189	–0.0167
	(0.0063)	(0.0071)	(0.0060)	(0.0062)
ASSASS	–0.0333	–0.0485	–0.0298	–0.0201
	(0.0155)	(0.0185)	(0.0130)	(0.0131)
PPI60DEV	–0.0143	–0.0171	–0.0141	–0.0140
	(0.0053)	(0.0078)	(0.0052)	(0.0046)
AFRICA	—	—	—	–0.0114
				(0.0039)
LAT.AMER.	—	—	—	–0.0129
				(0.0030)
R^2	0.56	0.49	0.56	0.62
$\hat{\sigma}$	0.0128	0.0168	0.0129	0.0119

Source: Barro (1991), pp. 410–413.

- Column (1): An extra dollar of GDP in 1960 (x =GDP60) is associated with a 0.75 percentage points decrease in annual growth.
- This is suppose to reflect a convergence property.
- Richer countries are expected to grow a lower rates than poorer countries.
- But, it we can also get an estimated β if we include the number of colors in the official flag of a country. Would that estimated slope mean anything?
- That is why we need **standard errors** (or **t-stats**).

- With SE we can answer this question: is the relation between x and y statistically significant?
- In the previous table the SE are in parenthesis, below the estimated coefficients.
- We can construct a confidence interval (at 95%) for the estimated parameters as follows

$$b_j - 1.96 * s_j \leq \beta_j \leq b_j + 1.96 * s_j \tag{2}$$

- b_j and s_j are the estimated parameter and its corresponding SE.
- A parameter is statistically different from zero when the intervals does not include zero.
- An alternative is to construct a t-stat: $\frac{b_j}{s_j}$.
- If the absolute value of t-stat is above 1.96, we reject the null hypothesis that the slope is zero.

3.4 OLS vs. causality

- Correlation (even partial correlation) does not imply causation.
- Examples: education and wages, institutions and economic growth.
- We can do regression analysis with these variables.
- But assigning one as y and the other as x does not guarantee causation.
- How would you test for causation?

- Selected methods in economics
  1. Random assignments: select randomly treatment and control groups for x. Any differences in y between groups is coming from the treatment.
  2. Instrumental variables: find an "exogenous" source of variation for x (variable z) that is not related to y. To evaluate impact of institutions (x) on economic growth(y), use initial settlements as the source of exogenous variation (z).
  3. Natural experiments: Find some changes in policy. What is the impact of increases in minimum wage (MW, x variable) on employment (y)? Maybe some states change their MW and others not. Compare people living on each side of the state border.