Project 2: High-dimensional Linear Models and Convergence in Economic Growth

Due: Sunday November 6th at 22:00

1 Economic Growth

A central question in growth theory is whether economies exhibit some form of convergence. That is, whether countries that start out poor tend to grow faster and vice versa. Barro [1991] proposed to explore this by regressing the average annual growth rate of GDP per capita in country i, denoted g_i, on the log of initial GDP per capita, y_{i0}, and a vector of control variables, $\mathbf{z}_i$:

$$g_i = \beta y_{i0} + \mathbf{z}_i \boldsymbol{\gamma} + u_i, \tag{1}$$

where u_i is an econometric error term ($u$nobservable). Figure 1a shows a scatterplot of the data with g_i on the y-axis and y_{i0} on the first axis, which indicates a seemingly clear negative unconditional correlation. That is, it seems that the raw associations in the data are consistent with a negative value of β, which is referred to as 'β-convergence' or 'catch-up growth.'

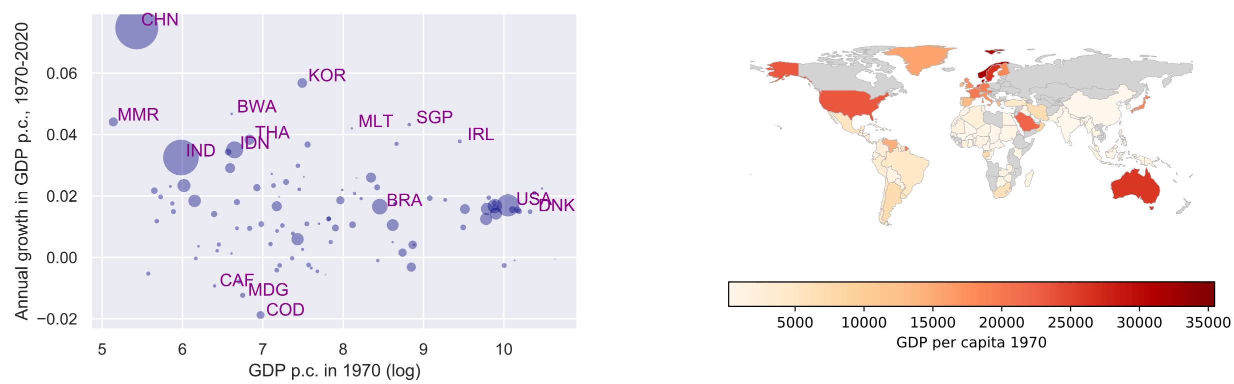
However, Figure 1b shows initial GDP on a map, from which one may be inspired to start thinking of potential omitted variables, that might explain simultaneously initial GDP and subsequent growth. For instance, although many African and Asian countries started from approximately the same levels, the subsequent growth trajectories have been remarkably different.

A key challenge in estimating β is therefore what to include in $\mathbf{z}_i$ in order to avoid omitted variable bias. Over time, researchers have emphasized a long list of possible controls, including geographic factors (like temperature, disease, agricultural suitability, and natural resources), but also historical determinants like colonialization, genetic diversity, culture, and more recently economic institutions. For example, a country may have been poor in 1970 due to a hazardous climate, which then also explains why that country remains poor and does not catch up during 1970–2020.

With as many basic candidate regressors as there are countries (or more), the empirical researcher faces a high-dimensional estimation problem when attempting to estimate β: Ideally, we want to avoid restricting the set of possible control variables ex ante, and instead let the data inform which are most important for economic growth to warrant inclusion. However, we also recognize the selection problem: with so many controls, we are forced to choose some and leave out others. (Even high-dimensional methods break down eventually.)

Figure 1: Growth and Convergence

(a) Growth and Initial GDP (b) Initial GDP Across the Globe



Note: In panel a, each dot is a country, and the size of the dot is proportional to the initial population. Names are provided for a few selected countries. In panel b, countries are colored by initial GDP per capita (1970). Grey colors indicate missing values.

2 Data

The dataset, `growth.csv`, contains data on 214 countries for a long list of variables (see Appendix A). For the outcome variable, g_i, you should use the variable `gdp_growth`, and the key variable of interest, y_{i0}, is `lgdp_initial`. As the appendix table shows, `gdp_growth` is only available for 102 countries, so this will provide an upper bound on the number of observations.

You must choose which variables to include in addition to these two. For instance, it is common to include the investment rate, `investment_rate`, as a control, but what additional controls you permit is up to you. Aim to strike a balance between permitting as many as possible and avoiding variables that restrict your sample size too heavily.

3 Assignment

Test whether growth is consistent with the theory of β-convergence. In doing so, you must treat $\mathbf{z}_i$ as high dimensional.

4 Hints

(1) You may start by discarding some of the available variables and never considering those in your project. For example, you could write your entire paper only considering the

geographical variables (plus g_i and y_{i0}, of course). You should then proceed by treating that list of variables in a high-dimensional paradigm.

(2) When using an estimation procedure, carefully discuss the assumptions required to derive the estimator and establish properties thereof, and assess whether these assumptions are likely to be satisfied in the current empirical setting. If not, what are the consequences for the estimator in question (and your results)? Strive to provide a real-world example of behavior that might invalidate a given assumption, carefully linking the behavior or mechanisms to the mathematical symbols in the model.

(3) Be precise about the statistical tests you use for testing various hypotheses. Explain which null hypothesis you are testing, the alternative you are testing against, how the test statistic is constructed, the decision rule you employ, and the conclusion you reach.

5 Formal Requirements

- You must hand in a report that presents the econometric model, presents your estimation results and results of formal statistical tests (including interpretation and statements on economic and statistical significance), and discusses the potential weaknesses of the model, data and approach. If presenting many estimates of the same parameters (e.g. estimators based on different assumptions, or varying the controls or sub-sample used), it may be helpful to present the estimates together in one table to facilitate comparison.

- The report must be written in English and uploaded to Peergrade via Absalon as a single PDF file.

- The report must be at most five normal pages long, plus at most two pages of output in the form of properly formatted and labelled tables or graphs.

  - Five normal pages correspond to 12,000 characters including spaces and math.
  - Use fontsize = 12p, line spacing = 1.5, and 2.5 cm page margins (as in this document).
  - Unfortunately, there is no standard way of counting mathematical characters. Copy-paste the text including math and formulas from your PDF document into: https://charcounter.com/ or https://charactercounttool.com. Report the number of characters on the front page. We will allow a 20 percent buffer, as different methods of counting the characters give different results. If your report

clearly exceeds the limit, it will not be assessed. (If your report is $\geqslant 20$ percent math, then you are doing it wrong.)

- You must submit your Python code alongside your report. Make sure that your code runs with only minor and obvious modifications (e.g. changing paths).
- You are allowed (and strongly encouraged) to work in groups of up to three people. List all group members on the front page of your report in alphabetical order of surnames.
- The assessment criteria are given on the course website in Absalon.

References

Daron Acemoglu, Simon Johnson, and James A Robinson. The colonial origins of comparative development: An empirical investigation. *American economic review*, 91(5):1369–1401, 2001.

Daron Acemoglu, Simon Johnson, and James A Robinson. The colonial origins of comparative development: An empirical investigation: Reply. *American Economic Review*, 102(6):3077–3110, 2012.

Daron Acemoglu, Suresh Naidu, Pascual Restrepo, and James A Robinson. Democracy does cause growth. *Journal of political economy*, 127(1):47–100, 2019.

David Y Albouy. The colonial origins of comparative development: an empirical investigation: comment. *American economic review*, 102(6):3059–76, 2012.

Quamrul Ashraf and Oded Galor. The'out of africa'hypothesis, human genetic diversity, and comparative economic development. *American Economic Review*, 103(1):1–46, 2013.

Valentina A Assenova and Matthew Regele. Revisiting the effect of colonial institutions on comparative economic development. *PloS one*, 12(5):e0177100, 2017.

Robert J Barro. Economic growth in a cross section of countries. *The quarterly journal of economics*, 106(2):407–443, 1991.

A Variable Labels

The Table below shows the descriptions and sources for all variables in the dataset in `growth.csv`. There are 208 rows – one per country – so the column "Obs." indicates how many of the 208 countries have non-missing values for the given variable. The column "Source" indicates where the data comes from, with the abbreviations:

- WB: World Bank,
- AG: Ashraf and Galor [2013],
- ANRR: Acemoglu et al. [2019],
- AR: Assenova and Regele [2017]. Their data is based on Acemoglu et al. [2001].[1]

Variable	Description	Source	Obs.
abslat	Absolute latitude	AG	205
africa	Africa dummy	AG	208
americas	Americas dummy	AG	208
area	Total land area	AG	208
area_ar	Arable land area	AG	196
asia	Asia dummy	AG	208
capital_growth_pct_gdp_initial	Gross capital formation in 1970 (% of GDP)	WB	97
capital_growth_pct_gdp_now	Gross capital formation in 2020 (% of GDP)	WB	123
cenlong	Geodesic centroid longitude	AG	208
code	World Bank country code	AG	214
cons00a	constraint on executive in 1900	AR	91
currentinst	IFC ease of doing business index/rank, 2012 (** our measure)	AR	155
dem	Democracy measure by ANRR	ANRR	155
demBMR	Democracy measure by BMR	ANRR	154
demCGV	Democracy measure by CGV	ANRR	150
democ00a	democracy in 1900	AR	87
democ1	democracy in first year of independence	AR	87
demreg	Average democracy in the region*initial regime (leaving own country out)	ANRR	183
distc	mean distance to coast	ANRR	159
distcr	mean distance to coast or river	ANRR	159
distr	mean distance to river	ANRR	159

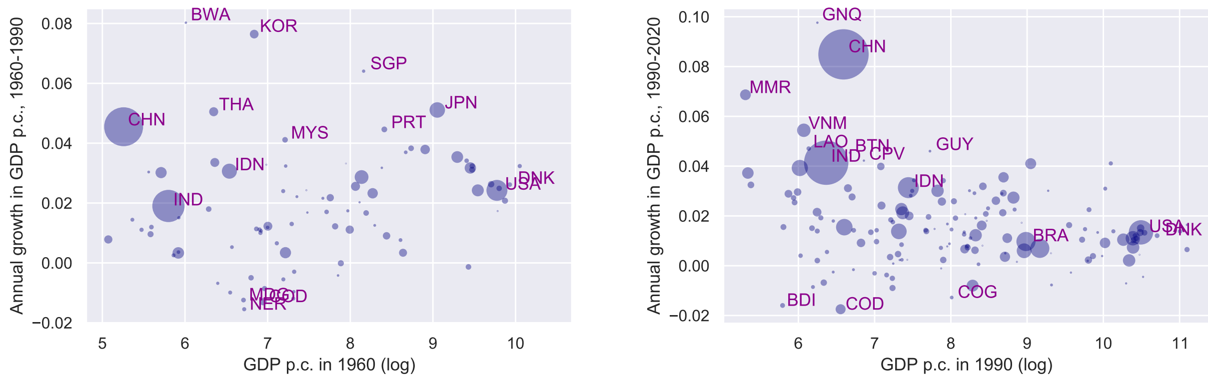
[1] Acemoglu et al. [2001] argue that settler mortality can be an instrument for settler mortality, as proxied by disease prevalence and temperature. This strategy relies on the assumption that temperature and disease prevalence has no direct effect on economic output; i.e. that all effects run through their effects on colonialization and subsequently on economic institutions. This was later debated by Albouy [2012] (AJR reply in Acemoglu et al., 2012), and then by AR.

elevavg	Mean elevation	AG	184
elevstd	Standard deviation of elevation	AG	161
europe	Europe dummy	AG	208
excolony	=1 if was colony FLOPS definiti	AR	78
gdp_growth	Annual growth in GDP per capita, 1970-2020	WB	102
gdp_initial	GDP in 1970	WB	109
gdp_now	GDP in 2020	WB	167
gdp_pc_initial	GDP per capita in 1970	WB	109
gdp_pc_now	GDP per capita in 2020	WB	167
ginv	Gross investment as a share of GDP	ANRR	104
goldm	Natural minerals: gold	AR	159
imputedmort	imputed mortality rate from logem4 measure (=exp(logem4))	AR	78
imr95	infant mortality rate (1995)	AR	60
investment_rate	Capital formation (% of GDP per year, avg. of available years 1970-2020)	WB	179
iron	Natural mineral: iron	AR	159
kgatr	Percentage of population living in tropical zones	AG	160
landlock	=1 if landlocked	AR	163
leb95	life expectancy at birth (1995)	AR	60
legor_fr	French legal origin dummy	AG	202
legor_uk	British legal origin dummy	AG	202
lgdp_initial	GDP per capita in 1970 (log)	WB	109
lh_bl	Percentage of population with tertiary education (Barro-Lee)	ANRR	143
ln_yst	Log [Neolithic transition timing]	AG	164
ln_yst_aa	Log [Neolithic transition timing (ancestry adjusted)]	AG	158
logem4	log of mortality rate (IV)	AR	87
lp_bl	Percentage of population with at most primary education (Barro-Lee)	ANRR	143
lpop_initial	Population in 1970 (log)	WB	199
ls_bl	Percentage of population with at most secondary education (Barro-Lee)	ANRR	143
lt100km	amount of territory within 100 km of the coast (McArthur and Sachs)	AR	61
malfal	Percentage of population at risk of contracting malaria	AG	164
marketref	Index of market reforms (1960)	ANRR	136
mortality	mortality measure (deaths per 1000 soldiers) from Appendix Table A2, AJR	AR	61
oceania	Oceania dummy	AG	208
oilres	oil reserves	AR	154
pcatholic	Share of Roman Catholics in the population	AG	204
pd1	Population density in 1 CE	AG	155
pd1000	Population density in 1000 CE	AG	177

pd1500	Population density in 1500 CE	AG	184
pd1500	Population density in 1500 CE	AG	184
pdiv	Predicted genetic diversity	AG	207
pdiv_aa	Predicted genetic diversity (ancestry adjusted)	AG	164
pdivhmi	Mobility index-predicted genetic diversity	AG	139
pdivhmi_aa	Mobility index-predicted genetic diversity (ancestry adjusted)	AG	132
pmuslim	Share of Muslims in the population	AG	204
polity	Polity (measure of democracy)	ANRR	122
polity2	Polity 2 (measure of democracy)	ANRR	122
pop1	Population in 1 CE	AG	155
pop1000	Population in 1000 CE	AG	177
pop1500	Population in 1500 CE	AG	184
pop_growth	Annual growth in population, 1970-2020	WB	198
population_initial	Population in 1970	WB	199
population_now	Population in 2020	WB	198
pother	Share of other religions in the population	AG	201
pprotest	Share of Protestants in the population	AG	201
precip	Precipitation	AG	184
rough	Terrain roughness	AG	184
silv	Natural mineral: silver	AR	159
suitavg	Land suitability for agriculture	AG	155
suitgini	Land suitability Gini	AG	160
temp	Temperature	AG	184
tropicar	% land area in geographical tropics	ANRR	159
uvdamage	Ultraviolet exposure	AG	207
yellow	=1 if vector yellow fever present today	AR	163
zinc	Natural mineral: zinc	AR	159

Figure 2

(a) 1960–1990 ($N = 88$) (b) 1990–2020 ($N = 151$)



Note: The two panels are identical, except that they are created based on the periods 1960–1990 and 1990–2020, respectively. The size of each dot is proportional to the population of the country in the baseline year.