KØBENHAVNS UNIVERSITET
DET SAMFUNDSVIDENSKABELIGE FAKULTET

AME, E23

Assignment 2

High-dimensional Linear Models and Convergence in Economic Growth

November 2023

Division of responsibilities:

xx: Anna Kathrine Deding Nielsen

xx: Barbara Bille Tagmose

xx: Martin Per Erik Kihlstedt

Contents

1 Introduction

In this paper we examine the economic theory of convergence. The theory states, that countries starting out relatively poor will have a faster growth rate in GDP. As in Barro (1991) we examine this through the equation stated in (1).

$$g_i = \beta y_{i0} + \boldsymbol{z_i}\boldsymbol{\gamma} + u_i \tag{1}$$

Where the average annual growth rate of GDP per capita in country i (in 1970-2020) is denoted g_i, the log of initial GDP per capita (in 1970), y_{i0} and a vector of control variables, $\boldsymbol{z_i}$. The unobservable error term is denoted u_i. If the theory of convergence is correct, we would expect a negative coefficient of β.

We examine the validity of the theory of convergence by estimating a Lasso of the model. First, we present the econometric theory behind this method, and then our estimation results. We find a statistically significant negative coefficient of β, which backs up the theory. Lastly, we discuss our findings. These should be seen in the light of possible bias, stemming from exclusion of relevant observations and/or controls.

2 Econometric theory

We use a dataset containing data on 214 countries, with a long list of variables. Considering the upper bound of the sample size, n, and the possible number of explanatory variables, p, we encounter a high dimensional framework. In this section we'll refer to the vector of the model (1) parameters as $\boldsymbol{\theta} = \{\beta, \boldsymbol{\gamma}\}$ and explanatory variables as $\boldsymbol{X_i} = \{y_{i0}, \boldsymbol{z_i}\}$. If we estimate the model using OLS, we obtain a measurement error as presented in (2) (Riis-Vestergaard Sørensen, 2023a).[1]

$$E[\frac{1}{n}\sum_{i=1}^{n}(X_i'\hat{\theta} - X_i'\theta)^2] = \frac{\sigma^2 p}{n} \tag{2}$$

[1] The last equality demands the assumptions of full rank, u being independent of y and z, and $u \sim \mathcal{N}(0, \sigma^2)$.

With a sizeable p, $\frac{p}{n} \nrightarrow 0$ as $n \rightarrow \infty$. Because of this, the OLS estimator makes a poor predictor in high dimensions and, in this case, thus of g_i and we must consider estimating otherwise.

2.1 Sparsity

We can work around the complications above, by assuming that the number of non-zero parameters in θ is *sparse.* Ideally, sparsity reduces the number of explanatory variables, so that only those relevant, (s), are maintained. Reducing p to s allows us to overcome the issue of $\frac{p}{n} \nrightarrow 0$ as $n \rightarrow \infty$, as long as $\frac{s}{n} \approx 0$ as $n \rightarrow \infty$. Exact sparsity is defined as in (3).

$$s = \sum_{j=0}^{p} \mathbf{1}\left\{\theta_j \neq 0\right\} \text{ is small} \tag{3}$$

Approximate sparsity, on the other hand, is that most parameters approximately equal to zero, but a few parameters far from. This is generally the assumption made in applied settings.

2.2 Lasso

If we make the assumption of sparsity we must then determine the relevant variables. This can be done using the Lasso estimator, which does not only estimate θ, but also performs variable selection. The estimator is defined as in (4) (Hastie et al., 2017).

$$\hat{\theta}(\lambda) \in \underset{b \in R^p}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (g_i - X_i'b)^2 + \lambda \sum_{j=1}^{p} |b_j| \right\} \tag{4}$$

Where $\lambda \geq 0$ is a *penalty level* to be selected. We may note that the left term corresponds to the MSE. The right term, λ multiplied with the L1-norm of b, when minimized over b, will move less relevant coefficients to zero. Thus, in (4), θ will be estimated by balancing simplicity, i.e., removing irrelevant variables based on our λ, and model fit.

For consistency, Lasso relies first on the condition that $\mathbf{X'}u/n$, must be "small". This condition is fulfilled by choosing the appropriate hyperparameter λ, which will be discussed in section 2.2.2. Second, $\mathbf{X'X}/n$ should be well behaved, i.e., invertible. However, the assumption of sparsity ensures that condition is met as long as the submatrix of relevant regressors $\mathbf{X_j'X_j}/n$ is invertible. So even if $\mathbf{X'X}/n$ is singular, the condition may still be met depending on the choice of X_j's.

2.2.1 Post-Double Lasso

It is usually a challenge to do inference based on Lasso estimates. This is so since we are lacking many of the traditional features for constructing confidence intervals. The problem can be dealt with b by using the Post-Double Lasso (PDL). Estimation with PDL is done in a few steps (Riis-Vestergaard Sørensen, 2023b). We seek to estimate how initial GDP per capita, $\boldsymbol{y_{i0}}$, affects the growth rate, $\boldsymbol{g_i}$, controlling for a variety of characteristics, $\boldsymbol{z_i}$. First we run $y_{i0} = z_i'\hat{\psi} + v$ to estimate $\hat{\psi}$ (with $E[v|Z] = 0$) using Lasso. We then run $g_i = \hat{\beta} y_{i0} + z_i'\hat{\gamma} + u_i$, (with $E[u|y_0, z] = 0$). We then estimate $\check{\beta}$ as in (5).

$$\check{\beta} := \frac{\sum_{i=1}^{n}(y_{i0} - z_i'\hat{\psi})(g_i - z_i'\hat{\gamma})}{\sum_{i=1}^{n}(y_{i0} - z_i'\hat{\psi})y_{i0}} \tag{5}$$

We note that this approach mimics that of an IV estimation, hence accounting for some correlation in $\boldsymbol{X_i}$. (5) in effect corresponds to estimating the impact of y_{i0} on g_i (i.e., $\check{\beta}$). Thus, assuming sparsity and regularity, PDL will satisfy $\frac{\sqrt{n}(\hat{\beta}-\beta_0)}{\sigma_0} \xrightarrow{d} \mathcal{N}(0,1)$ as $n \to \infty$ even in a high dimensional scenario. We estimate the variance by $\check{\sigma}^2 := \frac{n^{-1} \cdot \sum_i \hat{u_i}_i^2 v_i^2}{(n^{-1} \sum_i v_i^2)^2}$ and are able to do inference and construct confidence intervals.

2.2.2 Choosing the hyperparameter

When estimating using the Lasso, we need to chose the optimal level of penalization, and hence the hyperparameter, λ. There are several ways to determine this. We resort to using the (BCCH)-rule defined by Belloni, Chen, Chernozhukov, and Hansen (2012). In order to estimate λ^{BCCH} we choose the significance level $\alpha = 0.05$ and set $c = 1.1$. First, we estimate a pilot hyperparameter, λ^{pilot} as in (6).

$$\lambda^{\text{pilot}} = \frac{2c}{\sqrt{n}} \Phi^{-1}(1 - \frac{\alpha}{2p}) \max_{1 \leq j \leq p} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (g_i - \bar{g})^2 X_{ij}^2} \tag{6}$$

Using λ^{pilot}, we run a pilot Lasso, and obtain the residuals $\hat{u}_i = Y_i - X_i'\hat{\beta}^{\text{pilot}}$. These are now used in the estimation of the final hyperparameter, λ^{BCCH}, as in (7).

$$\lambda^{\text{BCCH}} = \frac{2c}{\sqrt{n}}\Phi^{-1}(1 - \frac{\alpha}{2p}) \max_{1\leq j\leq p} \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{u_i}^2 X_{ij}^2} \tag{7}$$

Compared to methods like sample splitting and cross-validation, which are good for out-of-sample prediction, this rule is better when we want to deal with variable selection and model parameters. Additionally, with this rule, we do not need to know the variance of u_i or assume homoskedastic of u_i, as is the case with the BRT rule. The last is favorable, since it is hard to imagine the variance of i.e. initial GDP being the constant across countries.

3 Analysis

In this section we present our process of selecting relevant variables for our study. After this, we present the estimation of the coefficient of the growth rate of GDP per capita, β using the empirical framework presented in section 2. We are testing the null-hyphotesis of $\mathcal{H}_0 : \beta \geq 0$ against the alternative, $\mathcal{H}_A : \beta < 0$. A rejection of $\mathcal{H}_0$ supports the theory of convergence.

3.1 Variable selection

In selecting relevant variables for our estimation of model (1), we face a trade-off. On the one hand, we wish to include as many relevant explanatory variables as possible. The variables contain less data for some countries, so some observations would have to be excluded if more variables are used. On the other hand, we wish to have as many countries included in the analysis as possible, for the aim of inference.

Other papers have found, that i.e. institutions, democracy and genetic diversity play a role in determining economic performance (Acemoglu et al., 2001; Acemoglu et al., 2019; Ashraf & Galor, 2013). We take this into account by including variables that measure the strength of these factors. The 48 variables for 75 countries included in our analysis are presented in Table 2. The variables in $\boldsymbol{z}_i$ are cubed and interacted, in order to take correct specification into account. As a result of this, we end up with 18,871 control variables in total. The 75 countries are presented in Table 3.

3.2 Results

In Table 1 our estimated coefficient on initial GDP is presented. The coefficient, β is -0.16, implying that an increase in initial GDP (in 1970) at 1 pct. results in a -0.16 pct. point decrease in the annual growth in GDP per capita. As seen, this is significant at a 95 pct. level. This result backs up the theory of convergence.

4 Discussion and concluding remarks

Our result should be viewed in the light of the variables and observations that we included. As mentioned in section 3.1, we faced a trade-off regarding a large feature matrix and a large sample-size. Because of this, we ended up only including 75 countries in our estimations. If the observations that we end up excluding have certain systematic attributes, our conclusion in favour of the theory of convergence could be biased. Also, we must consider the possibility of us having excluded some relevant variables in the selection process. This can result in omitted variable bias, where some excluded variables actually have power in predicting the growth rate in GDP, g_i. However this should be less severe than estimating with Post Single Lasso (PSL), due to the IV nature of the PDL. However the PDL only allows us to estimate a single parameter, whereas PSL allows us to interpret the roles of the controls as well.

Our focus on democratic and institutional factors could be misleading since it seems reasonable to assume that the interplay of a range factors create differing conditions for growth across countries. Correspondingly, this poses a problem to our consistency assumption for Lasso, since it might simply not be the case that $\mathbf{X}$'u/n is small.

However it is also conceivable that our analysis has taken too many factors into account. It is worth keeping in mind that no countries have identical conditions for economic growth, making it difficult to isolate the effect of initial GDP in a reliable manner. A potential way to mitigate these issues would be to e.g., perform multiple analyses, comparing our results to groupings of fewer countries with as similar conditions for growth as possible. Such an analysis could at least provide complementary insight on the reliability of our results.

Tables and Graphs

Table 1: Estimation result

β	-0.166
	$(-0.196, -0.135)$
Polynomial degrees	3
Number of observations	75
Number of control variables	$47\ (18, 871)$

Note: $\lambda = 1.53$ is calculated according to the BCCH-rule.
Number of controls incl. third-order polynomials and interactions in parenthesis.

Table 2: Selected variables

Variable name	Label	Source
ls_bl	Percentage of population with at most secondar...	ANRR
lh_bl	Percentage of population with tertiary educati...	ANRR
demCGV	Democracy measure by CGV	ANRR
demBMR	Democracy measure by BMR	ANRR
dem	Democracy measure by ANRR	ANRR
demreg	Average democracy in the region*initial regime...	ANRR
distcr	mean distance to coast or river	ANRR
distc	mean distance to coast	ANRR
distr	mean distance to river	ANRR
tropicar	% land area in geographical tropics	ANRR
africa	dummy=1 for Africa	AR
asia	dummy=1 for Asia	AR
landlock	=1 if landlocked	AR
goldm	Natural minerals: gold	AR
iron	Natural mineral: iron	AR
silv	Natural mineral: silver	AR
zinc	Natural mineral: zinc	AR
oilres	oil reserves	AR
yellow	=1 if vector yellow fever present today	AR
.....		

pop1000	Population in 1000 CE	QG
pop1500	Population in 1500 CE	QG
pd1000	Population density in 1000 CE	QG
pd1500	Population density in 1500 CE	QG
pdiv	Predicted genetic diversity	QG
pdiv_aa	Predicted genetic diversity (ancestry adjusted)	QG
cenlong	Geodesic centroid longitude	QG
area	Total land area	QG
area_ar	Arable land area	QG
abslat	Absolute latitude	QG
suitavg	Land suitability for agriculture	QG
suitgini	Land suitability Gini	QG
elevavg	Mean elevation	QG
elevstd	Standard deviation of elevation	QG
rough	Terrain roughness	QG
temp	Temperature	QG
precip	Precipitation	QG
distcr	Mean distance to nearest waterway	QG
kgatr	Percentage of population living in tropical zones	QG
malfal	Percentage of population at risk of contractin...	QG
pprotest	Share of Protestants in the population	QG
pcatholic	Share of Roman Catholics in the population	QG
pmuslim	Share of Muslims in the population	QG
uvdamage	Ultraviolet exposure	QG
africa	Africa dummy	QG
asia	Asia dummy	QG
oceania	Oceania dummy	QG
americas	Americas dummy	QG
ln_yst	Log [Neolithic transition timing]	QG
investment_rate	Capital formation (% of GDP per year, avg. of ...	WB
lgdp_initial	GDP per capita in 1970 (log)	WB
pop_growth	Annual growth in population, 1970-2020	WB

Note: The table presents the dependent variable, g_i, the independent, y_{i0} and the 47 controls. WB: World Bank, ANRR: Acemoglu et al. (2019), AR: Assenova and Regele (2017).

Table 3: Country codes for included obs.

ARG	AUS	AUT	BDI	BEL
BEN	BOL	BRA	BWA	CAF
CHL	CHN	CIV	CMR	COG
COL	CRI	DEU	DNK	DOM
DZA	ECU	EGY	ESP	FRA
GAB	GBR	GHA	GMB	GRC
GTM	HND	HTI	IDN	IND
IRL	IRN	IRQ	ITA	KEN
KOR	LKA	LSO	MAR	MEX
MLI	MMR	MRT	MYS	NER
NIC	NLD	NOR	NPL	PAK
PAN	PER	PHL	PRT	PRY
RWA	SAU	SDN	SEN	SLE
SLV	SWE	TGO	THA	TUN
TUR	USA	ZAF	ZMB	ZWE

Literature

Acemoglu D., Johnson S., and Robinson J. (2001). The colonial origins of comparative development: An empirical investigation. American economic review, 91(5):1369–1401.

Acemoglu D., Naidu S., Restrepo P., and Robinson J. (2019). Democracy does cause growth. Journal of political economy, 127(1):47–100.

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

Assenova V. & Regele M. (2017). Revisiting the effect of colonial institutions on comparative economic development. PloS one, 12(5):e0177100.

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

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6), 2369–2429.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. CRC press.

Riis-Vestergaard Sørensen, J. (2023a). Linear model in high dimensions, 1: Introduction and implementation. University Lecture, UCPH

Riis-Vestergaard Sørensen, J. (2023b). Linear model in high dimensions, 2: Estimation and Inference. University Lecture, UCPH