UNIVERSITY OF COPENHAGEN
FACULTY OF SOCIAL SCIENCES

Advanced Microeconometrics

Final Exam

Authors:

Caroline Bergholdt Hansen: 1.1, 1.3, 1.3.1.1, 1.3.2, 1.5, 2.2, 2.4, 2.6, 2.8, 2.10

Julie Cathrine Krabek Sørensen: 1.2, 1.3.1, 1.3.1.2, 1.4, 2.1, 2.3, 2.5, 2.7, 2.9, 2.11

13-15 January 2024

Part I: Approx. 9.897 characters

Part II: Approx. 10.585 characters

1 High-dimensional Linear Models and Convergence in Economic Growth

1.1 Introduction

A key question in growth theory concerns economic convergence, the idea of whether nations that initiate from a less wealthy position tend to undergo faster growth, than those starting with higher prosperity. This paper investigates the β-convergence hypothesis empirically by using the Post-Double-Lasso approach on a high-dimensional data set. While the accuracy of the findings of the analysis may be subject to considerable concerns, they appear to support the convergence hypothesis.

1.2 Empirical Model and Data

Barro (1991) suggested examining the convergence hypothesis statistically by regressing a nation's annual GDP[1] growth, g_i, on its initial GDP, y_{i0}, and a set of control variables, $\mathbf{z}_i$ in the following form:

$$g_i = \beta y_{i0} + \mathbf{z}_i\gamma + u_{it}, \quad \mathbb{E}[u_i|y_{i0}, \mathbf{z}_i] = 0 \tag{1}$$

where u_{it} is the unobservable error term for country $i = 1, ..., n$. The parameter β is thus the partial effect of initial GDP on GDP growth, such that $\beta < 0$ signifies aforementioned economic convergence. We aim to test the null hypothesis that β-convergence does not occur, i.e. $H_0 : \beta \geq 0$, with alternative hypothesis $H_A : \beta < 0$.

For analysing this model we include a data set which is a collection of variables from 102 different countries across the globe. The data stems from the World Bank, as well as data collected by other researchers[2]. Following the work of prominent scholars within the field, we choose to include variables with geographic (Diamond and Ordunio, 1999) and institutional (Acemoglu, Johnson, et al., 2005) character in our investigation, as they previously have proven to be effective in questions of economic growth.

1.3 Theory

The model of interest is high-dimensional (HD) in that it involves a small number of countries (n) compared to the number of candidate regressors (p). This feature of the problem renders least squares (LS) estimation to be less than optimal, as is evident by

[1] In this paper, GDP is always measured in per capita terms but shortened to just GDP.

[2] See Table 1 in the appendix for a full list of chosen variables and sources.

the formula for the variance of its prediction error:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}_i^{LS}-g_i^*\right)^2\right]=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i'\hat{\Psi}^{LS}-\mathbf{X}_i'\Psi\right)^2\right]=\frac{\sigma^2 p}{n} \tag{2}$$

where $g_i^* := \mathbb{E}[g_i|\mathbf{X}_i] = \mathbf{X}_i'\beta$ and $\mathbf{X}_i$ and Ψ are combined regressors $(y_{i0}, \mathbf{z}_i)$ with respective coefficients β and γ. Hence, the LS prediction error does not tend to zero as $\frac{p}{n} \nrightarrow 0$. Lasso, on the other hand, remains useful in HD settings as long as the underlying number of relevant regressors, s, is small relative to the number of observations, n. Lasso can be shown to outperform LS when $\frac{p}{s} \to \infty$.

1.3.1 The Lasso Estimator

Lasso encourages sparsity by introducing a penalty term to the LS minimization problem, thus penalizing the inclusion of additional regressors:

$$\hat{\beta}(\lambda) \in \arg\min_{b\in\mathbb{R}^p}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(g_i-\mathbf{X}_i'b\right)^2+\lambda\sum_{j=1}^{p}|b_j|\right\} \tag{3}$$

where $\mathbf{X}_i := [y_{i0}, \mathbf{z}_i]'$. The first term is the usual objective function of the LS, while the second term introduces aforementioned penalty. This penalty depends on a penalty term, λ, and the sum of the absolute value of the coefficients. The Lasso thus performs variable selection in the sense that a variable j is selected iff $\hat{\beta}_j(\lambda) \neq 0$. An implicit assumption of Lasso is that of sparsity, meaning that, from all the candidate regressors, p, there is an underlying set of relevant regressors, $s = \sum_{j=1}^{p} \mathbf{1}(\beta_j \neq 0)$, which is smaller.

1.3.1.1 The Post-Double-Lasso Estimator (PDL)

Post-Double-Lasso estimation involves first estimating:

$$y_0 = \mathbf{z}'\psi + v, \qquad E[v|\mathbf{z}] = 0 \tag{4}$$

Using Lasso to obtain $\widehat{\psi}$ and then estimating (1) by Lasso to obtain $\widehat{\beta}$ and $\widehat{\gamma}$. (4) is known as the first stage and serves the purpose of extracting the variation in the treatment variable that is not explained by the control variables. Together, the assumptions $E[u|\mathbf{X}] = 0$ and $E[v|\mathbf{z}] = 0$ imply that:

$$E[(y_0 - \mathbf{z}'\psi)(g - \beta y_0 - \mathbf{z}'\gamma] = 0 \tag{5}$$

The added structure imposed by PDL implies the moment condition above, which in turn allows us to isolate the parameter of interest. By the analogy principle, it can be

computed with the following formula for a given sample:

$$\check{\beta} = \frac{\sum_{i=1}^{n}(D_i - \mathbf{Z}_i'\widehat{\psi}_0)(Y_i - \mathbf{Z}_i'\widehat{\gamma}_0)}{\sum_{i=1}^{n}(D_i - \mathbf{Z}_i'\widehat{\psi}_0)D_i} \tag{6}$$

1.3.1.2 Penalty-Term Selection

Selecting the appropriate penalty term λ is key to the reliability of Lasso regression. Our focus will be directed towards utilizing the Bickel-Ritov-Tsybakov rule (BRT) and Belloni-Chen-Chernozhukov-Hansen rule (BCCH)[3].

The **BRT rule** relies on the assumption of conditional homoskedasticity, which implies that the residual term ε is independent of the predictor matrix $\mathbf{X}$ and further that the variance of ε is known. The penalty term is calculated:

$$\hat{\lambda}^{BRT} := \frac{2c\sigma}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2p}\right)\sqrt{\max_{1\leq j\leq p}\frac{1}{n}\sum_{i=1}^{n}X_{ij}^2} \tag{7}$$

The econometrician must make choices for the parameters α, significance level, and c, scaling factor, which must be larger than 1. We set it to 1.1 in line with previous literature.

The **BCCH rule** is calculated similar to the BRT, but does not rely on the assumption of prior knowledge of the variance.

$$\hat{\lambda}^{BCCH} := \frac{2c}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2p}\right)\max_{1\leq j\leq p}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon_i}^2 X_{ij}^2} \tag{8}$$

The ε is obtain from the auxillary regression $\hat{\varepsilon_i} = g_i - X_i'\hat{\beta}^{pilot}$, where $\hat{\beta}^{pilot}$ is an initial estimate derived with a pilot penalty term, which is calculated the same way as $\hat{\lambda}^{BCCH}$, but where $\hat{\varepsilon_i}$ is replaced with $g_i - \bar{g}$. It is noteworthy that the assumption of a known variance does not hold in most applications and so the BCCH is often preferred. The BCCH does however often yield a higher penalty term, and is thus more restrictive.

1.3.2 Inference

Inference based on Lasso estimation can be cumbersome due to the estimator not being analytically expressible, making it difficult to construct confidence intervals and hypothesis testing. We therefore resort to a different version of Lasso, namely PDL. While the asymptotic distribution of Lasso is generally unknown, it can be shown that, under certain

[3]Cross-validation is also a typical method, but where it is a valuable tool for out-of-sample prediction, it is not necessarily well suited for inference purposes. Therefore it is left out of the analysis.

sparsity conditions, PDL converges to a standard normal distribution as $n \to \infty$ and that $\sigma^2 = \frac{E[\varepsilon^2 v^2]}{(E[v^2])^2}$. By the analogy principle, we can estimate the variance as $\check{\sigma}^2 = \frac{n^{-1}\sum_i \widehat{\varepsilon}_i^2 \widehat{v}_i^2}{(n^{-1}\sum_i \widehat{v}_i^2)^2}$. For our purposes, this is the key advantage of PDL compared to Lasso, as it allows us to compute asymptotically valid confidence intervals:

$$\widehat{CI}(1-\alpha) = \left[\check{\beta} \pm q_{1-\alpha/2}\frac{\check{\sigma}}{\sqrt{n}}\right] \tag{9}$$

where $\alpha \in (0,1)$ is the significance level and $q_\alpha := \Phi^{-1}(\alpha)$ is the standard normal quantile function. We use a significance level of $\alpha = 0.05$ all throughout the paper.

1.4 Analysis

We estimate eight different models. Table 2 summarises the results. Models (a)-(b) are estimated by OLS with (a) including the treatment variable as the only regressor, while (b) also includes the selected control variables specified in Section 2. Models (c)-(e) are estimated by PDL using the BRT rule for penalty-term selection. Model (c) includes only the original selected control variables, whereas (d) also includes technical variables in the shape of interaction terms, and (e) further adds squares. Finally, models (f)-(h) are pairwise identical to (c)-(e) except the BCCH rule is applied.

In our way of specifying the model, it is implicit that β is the object of interest, while the presence of the control variables is to avoid omitted variable bias rather than an interest in the coefficient estimates for these variables.

Our eight regressions yield the following results. While the simple OLS regression (a) does not deliver a statistically significant estimate for the coefficient on the treatment variable, the resulting estimates from the OLS model including control variables (b) seem to confirm the hypothesis of β-convergence as the coefficient estimate on y_{i0} is negative and statistically significant. Similarly, the models estimated by PDL also yield negative and significant estimates for the coefficient on log initial GDP. The penalty from the BCCH however removes all control variables, while with the BRT at most 2 variables remain. This means that for the BCCH models (f)-(g) the estimates are the same for all three models.

1.5 Discussion and Conclusion

Even though most of the results confirm the economic divergence hypothesis, there are several reasons we might not have complete faith in their statistical validity. First of all,

it is of concern that the models estimated by PDL lead to a very small set of regressors with non-zero coefficient estimates. In model (c), 'Africa' and 'Asia' are the only non-zero control variables out of 26, while the only non-zero control variable in model (d) is the interaction term between the absolute latitude and 'Asia', and in model (e) it is only the interaction term between geodesic centroid longitude and 'Oceania'. While the former case might be explained by Africa and Asia experiencing the highest GDP growth, it is not immediately clear why the two interaction terms in models (d) and (e) are the only control variables with a non-zero coefficient.

However, that so few variables have non-zero estimates could be an indicator that the penalties selected by the BRT and BCCH are too restrictive, making it difficult to extract information and conclude upon the results, i.e., none of the models estimated by PDL using the BCCH result in any of the control variables having a non-zero coefficient estimate. This could indicate that the assumptions of the BCCH which lead to a more restrictive penalty term, make it unfit for our setting. This is of great concern as we do not believe that the assumption of a known variance necessary for the BRT holds. Ultimately, although it is an advantage of PDL that it performs variable selection, it is not very useful if it barely selects *any* variables. It could, on the other hand, also be the case that none of the included control variables are relevant, and that this is why none of them are 'selected'. This seems a bit unlikely, given the results from the previous literature.

Another concern is that of the suitability of OLS estimation for HD data. As already mentioned, OLS typically produces imprecise estimates in HD settings due to the variance of its prediction error depending on $\frac{p}{n}$. While adding the 26 control variables to model (b) does lead to substantially larger standard errors, we are nonetheless still able to establish statistical significance of the coefficient estimate of main interest. It is not certain that this would still be the case if we included even more control variables, in which case the problem would be 'truly' HD.

In conclusion, all the obtained estimates for β (except for model (a), whose validity is highly questionable), confirm the hypothesis of economic convergence, but the fact that OLS is typically not well-suited for this type of problem and PDL in this case leads to only very few variables being selected indicate that the results might not be completely reliable.

2 Heckmania

2.1

As the selection criterion is that $\delta_o x_i + v_1 \in (-\infty, +\infty)$ it means that we have no truncation of the sample i.e. $s_i = 1$ for all i, and thus y_i is always observed. When we have no selection we will not have any selection bias i.e. $\gamma_o = 0$. Further we have that u_i and x_i are mean independent conditional on v_i, which means that

$$\mathbb{E}[u_i | x_i = x, v_i = v] = \mathbb{E}[u_i | v_i = v] = \gamma_o v = 0 \tag{10}$$

This implies that the OLS.1 assumptions hold which means that

$$\mathbb{E}(\mathbf{x}'u) = \mathbf{0} \tag{11}$$

If we premultiply the outcome equation by $\mathbf{x}'$ and take expectations we get the following:

$$\mathbb{E}(\mathbf{x}'y) = \mathbb{E}(\mathbf{x}'\mathbf{x})\beta_o + \mathbb{E}(\mathbf{x}'u) \tag{12}$$

Equations (11) and (12) then implies $\mathbb{E}(\mathbf{x}'y) = \mathbb{E}(\mathbf{x}'\mathbf{x})\beta_o$.

Since $\mathbf{x}$ is a vector of scalar random variables with full support, there will be variability in its values, and the elements will not be constant. As a result, the outer product $\mathbf{x}'\mathbf{x}$ will be a matrix with non-constant elements, and $\mathbb{E}(\mathbf{x}'\mathbf{x})$ will have full rank. OLS.2 then holds, which means that $\mathbb{E}(\mathbf{x}'\mathbf{x})$ is invertible and thus β_o can be identified as:

$$\beta_o = [\mathbb{E}(\mathbf{x}'\mathbf{x})]^{-1}\mathbb{E}(\mathbf{x}'y) \tag{13}$$

2.2

We use the analogy principle to turn the population problem into its sample counterpart. This means that instead of the expectations of the population we sub in the sample means. By this, our estimator for β_o becomes:

$$\hat{\beta} = \left(\frac{1}{N}\sum_{i=1}^{N} x_i' x_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} x_i' y_i\right) \tag{14}$$

As $y_i = x\beta_o + u_i$ we substitute that into equation (14) and get:

$$\hat{\beta} = \beta_o + \left(\frac{1}{N}\sum_{i=1}^{N} x_i' x_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} x_i' u_i\right) \tag{15}$$

We know by the law of large numbers (LLN) that the sample mean converges in probability to the population mean as $N \to \infty$. This implies that:

$$\frac{1}{N}\sum_{i=1}^{N} x_i' u_i \xrightarrow{p} \mathbb{E}(\mathbf{x}'u), \qquad \frac{1}{N}\sum_{i=1}^{N} x_i' x_i \xrightarrow{p} \mathbb{E}(\mathbf{x}'\mathbf{x}) \tag{16}$$

By OLS.2 we know that $\mathbb{E}(\mathbf{x}'\mathbf{x})$ is invertible, which means that by Slutskys theorem:

$$p\text{-lim}\left(\frac{1}{N}\sum_{i=1}^{N} x_i' x_i\right)^{-1} = [\mathbb{E}(\mathbf{x}'\mathbf{x})]^{-1} \tag{17}$$

Due to this, it is possible to infer that the last part of equation (15) will tend to 0 as $N \to \infty$:

$$\begin{aligned}
\hat{\beta} &= \beta_o + \left(\frac{1}{N}\sum_{i=1}^{N} x_i' x_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} x_i' u_i\right) \xrightarrow{p} \beta_o + [\mathbb{E}(\mathbf{x}'\mathbf{x})]^{-1}\mathbb{E}(\mathbf{x}'\mathbf{u}) \\
&= \beta_o + [\mathbb{E}(\mathbf{x}'\mathbf{x})]^{-1}\mathbf{0} \\
&= \beta_o + \mathbf{0} \\
&= \beta_o
\end{aligned} \tag{18}$$

This means that $\hat{\beta}$ converges in probability to the true parameter β_o and thus that our estimator $\hat{\beta}$ is consistent.

2.3

The asymptotic distribution of the OLS estimator is derived by writing:

$$\sqrt{N}(\hat{\beta} - \beta_o) = \left(\frac{1}{N}\sum_{i=1}^{N} x_i' x_i\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i' u_i\right) \tag{19}$$

For the first part of the right-hand side (RHS) we know that (17) holds due to OLS.2. Given that $\mathbf{x}$ is random, the central limit theorem and OLS.1 hold, then the second element on the RHS is:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i' u_i \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbb{E}(u^2\mathbf{x}'\mathbf{x})) \tag{20}$$

Hence, equation (19) becomes the following (by using the product rule):

$$\sqrt{N}(\hat{\beta}_o - \beta_o) = \left(\frac{1}{N}\sum_{i=1}^{N} x_i' x_i\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i' u_i\right) \xrightarrow{d} [\mathbb{E}(\mathbf{x}'\mathbf{x})]^{-1}\mathrm{N}(\mathbf{0}, \mathbb{E}(u^2\mathbf{x}'\mathbf{x})) \tag{21}$$

Given the normal property, we can conclude that:

$$\sqrt{N}(\hat{\beta}_o - \beta_o) = \mathrm{N}(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})$$

as $N \to \infty$. Where $\mathbf{A} := \mathbb{E}(\mathbf{x}'\mathbf{x})$ and $\mathbf{B} := \mathbb{E}(u^2\mathbf{x}'\mathbf{x})$. Thus, under OLS.1–2, OLS is ($\sqrt{N}$-)asymptotically normal. For $\hat{\beta}$ to be asymptotically efficient OLS.3 needs to hold as well i.e. $\mathbb{E}(u^2\mathbf{x}'\mathbf{x}) = \sigma^2\mathbb{E}(\mathbf{x}'\mathbf{x})$, where $\sigma^2 = \mathbb{E}(u^2) = var(u)$. This means that $\hat{\beta}$ is asymptotically efficient when the variation in u is uncorrelated with the regressor. This is the case in our model when $\gamma = 0$ then $var(y|x, s = 1) = var(y|x) = var(u)$. This means we have homoscedasticity in the error term and $\hat{\beta}$ is asymptotically efficient.

2.4

We now suppose that $a = 0$, which means that we have a selected sample. For the following we suppose that the probability of selection is greater than zero, i.e $P(s_i = 1) > 0$.

Our previously suggested estimator now has to be rewritten to take the selection process into account:

$$\hat{\beta} = \left(\frac{1}{N}\sum_{i=1}^{N} s_i x_i' x_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} s_i x_i' y_i\right) \tag{22}$$

By the same reasoning as presented in 2.2, equation (16) instead take the form:

$$\frac{1}{N}\sum_{i=1}^{N} s_i x_i' u_i \xrightarrow{p} \mathbb{E}(s_i x_i' u_i), \qquad \frac{1}{N}\sum_{i=1}^{N} s_i x_i' x_i \xrightarrow{p} \mathbb{E}(s_i x_i' x_i) \tag{23}$$

The last part of equation (23) can be rewritten as:

$$\mathbb{E}(s_i x_i' x_i) = P(s_i = 1)\mathbb{E}(x_i' x_i | s_i = 1) \tag{24}$$

We assumed $P(s_i = 1) > 0$ it is also feasible given the provided sample. $P(s_i = 1)$ is just a scalar that does not influence the rank condition as long as it is greater than zero. However, the OLS.2 assumption only holds if $\mathbb{E}(x_i' x_i)$ has full rank conditional on selection.

Even under the rank condition for β to be identified we still need OLS.1 to hold i.e. $\mathbb{E}(u|\mathbf{x}) = 0$. To find whether this holds we take the conditional expectation of y_i given $s_i = 1$:

$$\mathbb{E}[y_i | x_i = x, s_i = 1, v_i = v] = \mathbb{E}[\beta_o x_i + u_i | x_i = x, v_i = v] \tag{25}$$

as we know $\mathbb{E}(u_i | x_i = x, v_i = v)$ is not conditional on x the following holds:

$$\mathbb{E}[y_i | x_i = x, s_i = 1, v_i = v] = \beta_o x + \mathbb{E}[u_i | v_i = v] \tag{26}$$

where $\mathbb{E}[u_i | v_i = v] = \gamma_o v$. Thus $\mathbb{E}(u|\mathbf{x}) = \mathbb{E}[u_i | v_i = v] = \gamma_0 v$.

This means that $\hat{\beta}$ is only identified iff $\gamma_o = 0$, which means that our previously suggested estimator is only identified if there is no selection bias.

2.5

As $a = 0$ and $b = \infty$ we know that $s_i = 1$ iff $\delta_o x_i + v_i > 0$:

$$P(s_i = 1|x_i = x) = P(\delta_o x_i + v_i > 0)$$

We use the complement rule to rewrite the probability in terms of the relationship between the regressors and the error term:

$$= 1 - P(v_i \leq -\delta_o x)$$

We divide both sides of the inequality with the variance of v_i:

$$= 1 - P\left(\frac{v_i}{\sigma(x, \alpha_o, \eta_o)} \leq \frac{-\delta_o x}{\sigma(x, \alpha_o, \eta_o)}\right)$$

This is done as we know that the error term divided by its variance will be normally distributed i.e. $\frac{v_i}{\sigma(x,\alpha_o,\eta_o)} \sim N(0, 1)$, which means we can rewrite the term as:

$$= 1 - \Phi\left(\frac{-\delta_o x}{\sigma(x, \alpha_o, \eta_o)}\right)$$

As we know the functional form of σ we insert it into the equation:

$$= 1 - \Phi\left(\frac{-\delta_o x}{e^{(\alpha_o + \eta_o x)}}\right)$$

By the symmetry of the normal distribution's CDF, the expression is:

$$= \Phi\left(\frac{\delta_o x}{e^{(\alpha_o + \eta_o x)}}\right)$$

Thus we have shown that:

$$P(s_i = 1|x_i = x) = \Phi\left(\frac{\delta_o x}{e^{(\alpha_o + \eta_o x)}}\right) \tag{27}$$

2.6

Based on the selection equation we define a latent outcome variable $s_i^* = \delta_o x_i + v_i$, where $s_i = \mathbf{1}(s^* > 0)$. As s_i^* is unobserved this implies the following:

$$s_i = \mathbf{1}(s^* > 0) = \mathbf{1}(\delta_o x_i + v_i > 0) = \mathbf{1}((c\delta_o)x_i + (cv_i) > 0) \quad \text{for any } c > 0 \tag{28}$$

This means because cv_i shares the properties of v_i it is not possible to distinguish $c\delta_o$ from δ_o, and thus δ_o is not identifiable.

We can conclude from (27) that δ_o and $\sigma(x, \alpha, \eta)$ are only identifiable in a ratio between the two. To account for this identification issue, we must fix the scale i.e. σ, in order to identify δ_o. As an example we can fix $\sigma = 1$. We know the functional form of $\sigma(\alpha_o, \eta_o, x)$, which means that we can calculate it given values of α_o, σ_o and x. With the given form of σ it is not possible to identify a unique value for both α and η as multiply solutions is possible. e.g. $\alpha = 0.5, \eta = -0.5x = 1$ or $\alpha = 0, \eta = -1, x = -1$, or $\alpha = 1, \eta = -0.5, x = 2$. This means that even if we fix σ, to identify δ_o it will not be possible to uniquely identify α_o, η_o. Thus it is not possible to identify the triplet $(\delta_o, \alpha_o, \eta_o)$.

2.7

The parameters associated with the selection equation are δ_o, η_o and α_o. Given that we know $\alpha_o = 0$, it is possible to find a unique value of η_o, and thus δ_o. From η can calculate σ, and thus we can identify δ from the ratio δ/σ.

To estimate the parameters we use the first step in the Heckit procedure and equation (27), such that we probit regress s_i on x_i to estimate δ_o and η_o using the following probit log-likelihood:

$$\ell_i(\delta, \eta) = y_i \ln \Phi\left(\frac{\delta x_i}{e^{\eta x_i}}\right) + (1 - y_i) \ln\left[1 - \Phi\left(\frac{\delta x_i}{e^{\eta x_i}}\right)\right] \tag{29}$$

Thus, the probit estimators are:

$$\delta, \eta = \underset{\delta,\eta}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \ell_i(\delta, \eta) \tag{30}$$

For the standard errors to be valid the assumptions associated with maximum likelihood estimation must hold. Table 3 shows the most important assumptions for $\hat{\delta}$[4]. 12.2 (1) holds if δ_o are identified such that any other parameterization yields a different density than the true one. 12.2 (2) is taken as given. 12.2 (3) holds as error term v_i is assumed to be normally distributed with a continuous variance function, it implies continuity of the error term and, thus, of $\ell_i(\delta, \eta)$ as well. Thus the MLE is consistent. 12.3 (1) holds as δ_o is identified and therefore known to be an interior solution. The functional form of $\ell_i(\delta, \eta)$ lives up to 12.3 (2). The M-estimator is then asymptotic normally distributed:

$$\sqrt{N}(\hat{\delta} - \delta_o) \xrightarrow{d} N(\mathbf{0}, \mathbf{A}_o^{-1}\mathbf{B_o}\mathbf{A}_o^{-1}), \mathbf{A}_o = E[H(w_i, \delta_o)], \mathbf{B}_o = E[z(w_i, \delta_o)z(w_i, \delta_o)'] \tag{31}$$

[4]Further technical assumptions necessary are available in Wooldridge (2010)

where δ_o are the true parameters and $w = (\mathbf{s}, \mathbf{x})$. By the analogy principle, we substitute expectations with averages and true parameters with corresponding estimates, such that:

$$\hat{\mathbf{A}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{H}}_i = \mathbf{H}(\mathbf{w}, \hat{\delta}), \qquad \hat{\mathbf{B}} = \frac{1}{N_i}\sum_{i=1}^{N_i}\hat{z}_i\hat{z}_i', \quad \hat{z}_i := \mathbf{z}(s_i, x_i, \hat{\delta}) \tag{32}$$

For MLE, the Information Matrix Equality holds, which means that $\mathbf{A}_o = \mathbf{B}_o$, and the variance formula simplifies to $Avar(\hat{\delta}) = \mathbf{A}_o^{-1}/N$. Nonetheless, we compute standard errors using the sandwich variance as it is arguably more robust. We use the same reasoning for η.

From Table 4 the results are shown. We see that δ_o is 1.5041. The belonging standard error and t-ratio are 0.0909 and 16.5397, respectively, thus δ_o are significant. We see that $\eta_o = 0.4888$ with a standard error of 0.0566 and a t-ratio equal to 8.6306, hence η_o is significant as well.

2.8

We know that for a continuous random variable x_{ij} the partial effects can be calculated with the general formula:

$$PE = \frac{\partial P}{\partial x_j}(\mathbf{x}) = g(\mathbf{x}\boldsymbol{\beta_o})\beta_{o,j} \tag{33}$$

where $p(x) = G(\mathbf{x}\boldsymbol{\beta_o})$, and g is ther derivative of G. Our model only has one regressor which means that we are only calculating one type of partial effect and so the j subscript is not relevant. However, it is important to note that the partial effect is dependent on the value of x, which indicates that the partial effect will vary depending on its starting point x. We insert the derivative of our expression from equation (27):

$$PE = \frac{\partial P}{\partial x}(x) = \varphi\left(\frac{\delta_o x}{e^{\eta x}}\right)\left(\frac{\delta_o}{e^{\eta x}}\right) \tag{34}$$

where φ is the PDF of the standard normal distribution.

As seen in equation (27) we know that the selection probability only depends on $\frac{\delta_o}{\sigma}$, which makes our partial effects identifiable as long as the ratio of $\frac{\delta_o}{\sigma}$ is known.

Another option is to use the estimated $\hat{\delta}$ derived in 2.7 from our maximum likelihood estimation. This means that we can estimate a plug-in version of the partial effect:

$$PE = \frac{\partial P}{\partial x}(x) = \varphi\left(\hat{\delta}x\right)\hat{\delta} \tag{35}$$

2.9

To test "x_i has zero partial effect on selection", the null hypothesis is $H_0 : PE = 0$, with alternative hypothesis $H_A : PE \neq 0$. Given the use of the parameters estimated in 2.7, we estimate the PE from equation (35). We use the Delta method to calculate robust standard errors, as the partial effects are a function of the estimated parameters. In the following we write $PE = h(\hat{\delta})$. The Delta method states that if the estimated coefficients are $\sqrt{N}$-Asymptotic normally distributed with mean 0 and variance V, $h(\hat{\delta})$ are also $\sqrt{N}$-Asymptotic normally distributed with mean 0 and variance $\nabla h(\delta_o) V \nabla h(\delta_o)'$ We use the consistent estimator of the variance of $h(\hat{\delta})$, $\hat{V} = \nabla h(\hat{\delta}) \hat{\mathbf{\Sigma}} \nabla h(\hat{\delta})'$, where $\hat{\mathbf{\Sigma}}$ is the estimated covariance matrix for $\hat{\delta}$. As previously mentioned we assume the MLE to be $\sqrt{N}$-Asymptotic normally distributed. This allows us to compute asymptotically valid standard errors to test the null hypothesis stated above. From Table 4 we see that PE is 0.39, with an associated standard error of 0.0033 and t-ratio of 119.6143, thus the result is very significant. Therefore, we can reject our null and x_i has partial effect on selection.

2.10

To derive an expression for the conditional mean function we take the conditional expectation of y_i:

$$\mathbb{E}[y_i | x_i = x, s_i = 1, v_i = v] = \mathbb{E}[\beta_o x_i + u_i | x_i = x, v_i = v] \tag{36}$$

We then take $\beta_o x$ out of the expectation, as it a constant when $x_i = x$:

$$\mathbb{E}[y_i | x_i = x, s_i = 1, v_i = v] = \beta_o x + \mathbb{E}[u_i | x_i = x, v_i = v] \tag{37}$$

As we know $E[u_i | x_i = x, v_i = v]$ does not depend on x we can rewrite:

$$\mathbb{E}[y_i | x_i = x, s_i = 1, v_i = v] = \beta_o x + \mathbb{E}[u_i | v_i = v] = \beta_o x_i + \gamma_o v \tag{38}$$

Then by the law of iterated expectation we get the following:

$$\mathbb{E}[y_i | x_i = x, s_i = 1, v_i = v] = \beta_o x + \mathbb{E}[\gamma_o v | x_i = x, s_i = 1] \tag{39}$$

Then $s_i = 1$ are substituted for its expressions and γ_o is taken out of the expectation as it is a constant:

$$= \beta_o x + \gamma_o \mathbb{E}[v_i | v_i > -\delta_o x]$$

Due to truncation that says that $\mathbb{E}[Z|Z < c] = \frac{\varphi(c)}{1-\Phi(c)}$ it is possible to chance the expectations operator:

$$= \beta_o x + \gamma_o \frac{\varphi(-\delta_o x)}{1 - \Phi(-\delta_o x)}$$

Because of the symmetry of the standard normal CDF and PDF we rewrite the expression:

$$= \beta_o x + \gamma_o \frac{\varphi(\delta_o x)}{\Phi(\delta_o x)}$$

We use the inverse Mill ratio such that $\lambda(x) = \frac{\varphi(x)}{\Phi(x)}$:

$$= \beta_o x + \gamma_o \lambda(\delta_o x)$$

And thus we have derived the conditional mean function:

$$\mathbb{E}[y_i|x_i = x, s_i = 1] = \beta_o x + \gamma_o \lambda(\delta_o x) \tag{40}$$

2.11

The model parameters associated with the outcome equation are β_o and γ_o. To estimate those the second step in the Heckit procedure is carried forward, hence OLS is used to regress y_i on x_i and λ_i using the selected sample only. The inverse Mill ratio can be found from the $\hat{\delta}$ found in the first stage in 2.7, using the formula for λ introduced in 2.10. From Table 4 it is seen that $\hat{\beta}$ is 2.0885 with belonging standard error and t-ratio of 0.0639 and 32.6720, respectively. Hence, the parameter is significant. The selection bias parameter, $\hat{\gamma}$, is -0.0300. The associated standard error is 0.0956 and the t-ratio is -0.3144, thus it is insignificant, which indicates no selection bias is present in this model.

We have two problems with inference when $\gamma \neq 0$. Firstly, we have heteroscedasticity as $var(y|x, s = 1) \neq var(y|x)$ is not constant. Further, we have that $\hat{\lambda}$ is a generated regressor. The heteroscedasticity issue can be amended by using the robust covariance matrix $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$, but it does not solve the generated regressor problem. Given that we have generated λ by ML, the *Avar* must be adjusted. One option is to use the bootstrap method for estimating the standard error.

However, it is possible to test for no selection bias because under the $H_0 : \gamma = 0$ then $var(y|x, s = 1) = var(y|x) = var(u)$. This means we have homoscedasticity in the error term and no influence of the generated regressor. This means that the standard errors from the OLS regression are valid. As our model does not reject $H_0 : \gamma = 0$ it indicates that the OLS standard errors are valid in our model.

References

Acemoglu, Daron, Simon Johnson, and James A Robinson (2005). "Institutions as a fundamental cause of long-run growth". In: *Handbook of economic growth* 1, pp. 385–472.

Acemoglu, Daron, Suresh Naidu, et al. (2019). "Democracy does cause growth". In: *Journal of political economy* 127.1, pp. 47–100.

Ashraf, Quamrul and Oded Galor (Feb. 2013). "The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development". In: *American Economic Review* 103.1, pp. 1–46. DOI: 10.1257/aer.103.1.1. URL: https://www.aeaweb.org/articles?id=10.1257/aer.103.1.1.

Assenova, Valentina A and Matthew Regele (2017). "Revisiting the effect of colonial institutions on comparative economic development". In: *Plos one* 12.5, e0177100.

Barro, Robert J (1991). "Economic growth in a cross section of countries". In: *The quarterly journal of economics* 106.2, pp. 407–443.

Diamond, Jared M and Doug Ordunio (1999). *Guns, germs, and steel*. Vol. 521. Books on Tape New York.

Wooldridge, Jeffrey M (2010). *Econometric analysis of cross section and panel data*. MIT press.

Appendix A1

Table 1: Included variables

Variable	**Source**
Democracy measure by ANRR	ANRR
Average democracy in the region*initial regime (leaving own country out)	ANRR
Index of market reforms (1960)	ANRR
mean distance to coast or river	ANRR
mean distance to coast	ANRR
mean distance to river	ANRR
% land area in geographical tropics	ANRR
dummy =1 if landlocked	AR
Geodesic centroid longitude	QG
Total land area	QG
Arable land area	QG
Absolute latitude	QG
Land suitability for agriculture	QG
Land suitability Gini	QG
Mean elevation	QG
Standard deviation of elevation	QG
Terrain roughness	QG
Temperature	QG
Precipitation	QG
Percentage of population living in tropical zones	QG
Africa dummy	QG
Asia dummy	QG
Oceania dummy	QG
Americas dummy	QG
Capital formation (% of GDP per year, avg. of available years 1970-2020)	WB
GDP per capita in 1970 (log)	WB
Annual growth in GDP per capita, 1970-2020	WB
Annual growth in population, 1970-2020	WB

Sources:

WB: World Bank

AR: Acemoglu, Naidu, et al. (2019)

QG: Ashraf and Galor (2013)

ANRR: Assenova and Regele (2017)

Table 2: Results

	OLS		PDL (BRT)			PDL (BCCH)		
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Initial log (gdp)	-0.087	-1.203	-0.131	-0.137	-0.137	-0.132	-0.132	-0.132
Standard error	0.106	0.183	0.015	0.015	0.015	0.016	0.016	0.016
λ^{dz}			0.549	0.672	0.675	0.916	1.110	1.115
λ^{yx}			0.568	0.693	0.696	1.180	2.582	2.594
CI low	-0.295	-1.561	-0.160	-0.167	-0.167	-0.163	-0.163	-0.163
CI high	0.121	-0.845	-0.102	-0.108	-0.107	-0.102	-0.102	-0.102
t-statistic	-0.821	-6.574	-8.733	-9.133	-9.133	-8.250	-8.250	-8.250
Observations	102	76	76	76	76	76	76	76
Controls	0	26	26	344	370	26	344	370
Controls post-penalty	0	26	2	1	1	0	0	0

Appendix A2

Table 3: Maximum Likelihood Assumptions

Theorem	Assumption	Cont.	A.N.
12.2 (1)	$\boldsymbol{\beta}_o$ uniquely minimizes $E[-\ell(\boldsymbol{\beta})]$	*	
12.2 (2)	$\boldsymbol{\beta} \subseteq \mathbb{R}^P$ compact (i.e. closed and bounded)	*	
12.2 (3)	$\ell(\boldsymbol{\beta})$ is a continuous function in $\boldsymbol{\beta}$	*	
12.3 (1)	$\boldsymbol{\beta}_o$ is an interior solution to $\mathbb{R}^P$	*	*
12.3 (2)	$\ell(\boldsymbol{\beta})$ is continuous and twice differentiable on the interior of the compact parameter space	*	*

Table 4: Maximum Likelihood Estimates and Partial Effects

	Estimate	Standard Error	t-value
δ_o	1.5041	0.0909	16.5397
η_o	0.4888	0.0566	8.6306
PE	0.3900	0.0033	119.6143
β_o	2.0885	0.0639	32.6720
γ_o	-0.0300	0.0956	-0.3144