Advanced Microeconometrics

Project 2

Matias Hall
Copenhagen University
Department of Economics

Thomas Theodor Kjølbye
Copenhagen University
Department of Economics

January 2023

Character Count: 10,000

Mathias Hall is responsible for: 1, 2, 2.2, 3, 5

Thomas Theodor Kjølbye is responsible for 1, 2.1, 2.3, 4 :

1 Introduction

Catch-up growth - or convergence in economic growth - is one of the most fundamental and discussed topics in macroeconomics. For most of recent history, the west has enjoyed decades of - mostly - prosperity while e.g. Asian and African countries have lagged behind when compared in terms of wealth and standard of living. However, in recent years these very same countries have experienced increased economic growth, suggesting that the theory of convergence in economic growth may have merit.

In this paper, we test the theory of convergence so as to ascertain whether countries with a lower initial level of GDP have - and still are - catching up to countries with a higher initial level of GDP. We do so so in a high dimensional framework utilising the post-double lasso estimator. We find that we reject the null of convergence.

2 Theoretical Framework

Pursuant to the pioneering work of Barro (1991), we propose to explore the convergence of economies 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, $\mathbf{z}_i$

$$g_i = \beta y_{i0} + \mathbf{z}_i \gamma + u_i \tag{2.1}$$

where u_i is an unobservable random variable, often interpreted as an error term, and β is the object of interest, capturing the association of the initial level of GDP per capita and the DGP per capita growth when controlling

for the covariates in z_i. Catch-up growth implies $\beta < 0$.

Due to the nature of the data provided, enclosing more than 80 possible covariates, with only N = 104 observations, we enter a high-dimensional domain in which we expect the simple linear model to perform poorly. Sørensen (2022) proves that the simple linear model becomes increasingly inaccurate as the number of predictors increases. Specifically, the expected average squared prediction error of the least squares estimator and the optimal linear estimator is proportional to p, the number of covariates with proportionality factor $\frac{\sigma^2}{n}$. Formally,

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i^*)^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i\hat{\beta} - x_i\beta)^2\right] = \frac{\sigma^2 p}{n} \tag{2.2}$$

where n is the number of observation, $\hat{y}_i$ is the prediction of y_i of the least ordinary squares estimator, y_i^* is the optimal linear predictor, and $\hat{\beta}$ is the ordinary least squares estimator.

2.1 Penalized Linear: Lasso

To get around the drawbacks of the simple linear model in a high-dimensional framework, we present the penalized linear model. The penalized linear model appends a penalty term to the objective function of the linear model, thus imposing parameter parsimony which in turn deteriorates the in-sample fit.

The lasso specifically imposes an ℓ_1 penalty term consistent with the es-

timates

$$\hat{\theta}^{\text{lasso}} = \underset{\theta}{\text{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (g_i - x_i\theta)^2 + \lambda \left(|\beta| + \sum_{j=1}^{p} |\gamma_j| \right) \right\} \tag{2.3}$$

where $x_i = (y_{i0}, \mathbf{z}_i)$, $\theta = (\beta, \gamma)'$, and $\lambda \geq 0$ is the complexity or tuning parameter[1]. In contrast to the simple linear, the lasso regression is not scale-invariant, and as such, the parameters are not inversely proportional to changes in the units (scales) of the variables. Consequently, we standardize data to mean 0 and standard deviation 1.

2.2 Post-Double Lasso

With the objective of inference in mind, neither lasso nor the single post lasso are suitable candidates for the choice of estimator. The simple lasso constructs estimates for which the asymptotic distribution is unknown, complicating the construction of e.g. confidence interval and leaving us unable to conduct any valid inference. The single post lasso leverages the least squares estimator after an initial simple lasso selection. The approach relies on infeasible perfect model selection, and as such, is sensitive to mistakes if the lasso discards relevant control variables which in turn creates a bias in the estimates. Consequently, we use the post-double lasso procedure.

The post-double lasso procedure enhances the model

$$g_i = \beta y_{i0} + \mathbf{z}_i\gamma + u_i \tag{2.4}$$

[1]For $\lambda = 0$, the lasso estimator is identical to the least square estimator, though as $\lambda \to \infty$ the coefficients converge to 0.

with a first-step regression

$$y_{i0} = \mathbf{z}_i\psi + v_i \tag{2.5}$$

we then leverage the analogy principle to compute the object of interest, $\hat{\beta}$ from the estimand, β

$$\beta = \frac{E[(y_{i0} - \mathbf{z}_i\psi)(g_i - \mathbf{z}_i\gamma)]}{E[(y_{i0} - \mathbf{z}_i\psi)y_{i0}]} \tag{2.6}$$

As such, the post-double lasso entails three steps. First, we run the lasso specified in equation (2.5). Second, we run the lasso specified in (2.4). In closing, we estimate $\hat{\beta}^{PDL}$ invoking the analogy principle on equation (2.6).

Under the sparsity conditions and with the appropriate choice of penalty, we can get consistent estimates, and the post-double lasso satisfies

$$\sqrt{N}(\hat{\beta}^{PDL} - \beta) \xrightarrow{d} N(0, \sigma^2), \qquad \sigma^2 = \frac{E(v^2w^2)}{(E(w^2))^2} \tag{2.7}$$

2.3 Penalty Selection

There are three conventional approaches to determining the 'correct' level of penalty, $\hat{\lambda}$: cross-validation (CV), BRT-rule, and BCCH-rule. Though the most commonly used method of three is likely CV, the procedure is most apt at tackling problems in which prediction is the objective of interest. However, as we are concerned with inference in the context of economic convergence, we discard CV. Instead, we focus exclusively on BRT and BCCH.

The BRT penalty term is computed as

$$\hat{\lambda}^{\text{BRT}} := \frac{2c\sigma}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2p}\right)\sqrt{\max_{1\leq j\leq p}\frac{1}{n}\sum_{i=1}^{n}x_{ij}^2} \tag{2.8}$$

where Φ is the standard normal CDF, $c > 1$, $\alpha \in (0,1)$, and subscript j denotes the jth variable. The validity of the BRT penalty relies on i) the assumption of conditional homoscedasticity, i.e. $\text{Var}(u_i|\mathbf{x}) = \sigma^2$, and 2) that the variance σ^2 of the error term is known. In closing, given that we standardize our data, the BRT penalty term in (2.6) simplifies to

$$\hat{\lambda}^{\text{BRT}} := \frac{2c\sigma}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2p}\right) \tag{2.9}$$

The BCCH penalty term entails three steps. First, we choose α and c analogously to the BRT procedure, Second, we run a 'pilot' lasso $\hat{\beta}^{\text{pilot}} := \hat{\beta}(\hat{\lambda}^{\text{pilot}})$ with

$$\hat{\lambda}^{\text{pilot}} := \frac{2c}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2p}\right)\max_{1\leq j\leq p}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2x_{ij}^2} \tag{2.10}$$

Third, from the pilot, we compute the residuals, which in turn are leveraged to determine the level of penalty.

$$\hat{u}_i := y_i - x_i'\hat{\beta}^{\text{pilot}} \tag{2.11}$$

$$\hat{\lambda}^{\text{BCCH}} := \frac{2c}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2p}\right)\max_{1\leq j\leq p}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2\cdot x_{ij}^2} \tag{2.12}$$

Unlike the BRT penalty rule, the advantage of BCCH is that it relies neither on the assumption of conditional homoscedasticity nor on the preliminary knowledge of the variance of the error term.

3 Data

The cross-sectional data encloses country-level data for 214 countries, though the response variable (GDP growth per capita) is only observed for 104 countries. When selecting control variables, we make an effort to balance the trade-off there exists in mitigating any endogeneity issues by adding covariates, though it simultaneously reduces the number of observations available due to the poor quality of the data.

We considered imputing the data to increase the number of observations such that we could include more control variables without losing observations. However, we soon realized the large number of missing observations would have too great a weight in the analysis. We have chosen controls that are typical describers of growth plus investment rate and population growth as they are common predictors of growth known from amongst other the Solow model. The variables are reported below in table 1

Table 1: Caption

Institutions	`currentinst, dem, demreg, lh_bl, ls_bl, marketref`
Religion	`pcatholic, pmuslim, pprotest`
Geography	`abslat, africa, americas, area_ar, asia, distc` `landlock, oceania, oilres, precip, temp`
Extra	`investment_rate, pop_growth`

4 Empirical analysis

Using the methods mentioned in the previous section we estimate the model. Since the Lasso estimator is sensitive to rescaling the explanatory variables have been normalised. We estimate the model using both BRT and BCCH penalty terms. The results can be found in table 2. Both methods find that $\beta < 0$ which is in favor of the hypothesis of economic convergence. However, both estimates are highly insignificant, which is why we have to reject the hypothesis.

In the model with BRT penalty terms, LASSO includes the following variables: (`ls_bl, pprotest, pcatholic, demreg, asia, investment_rate`) in the first step (`currentinst, asia`), while with BCCH penalty terms LASSO only includes (`pprotest`) in the first step and excludes all controls in the second step. This is due to the high BCCH penalty value. We find it unlikely that only religion plays a part in growth, which is why the model with BCCH penalty probably suffers from some omitted variable bias. With BRT both institutional and geographical factors are included, making it seem more realistic and can be given a more straightforward causal interpretation, even though the lack of factors still is evidence of ommited variable bias. However, as stated BRT relies on the assumption of homoskedastic error terms. This implies that e.g. a characteristic such as temperature varies the same no matter the area, which is highly unrealistic. So even though BRT might be better in terms of ommited variabel bias, it does probably suffer from homoskedastic error terms, making it less efficient than BCCH.

A problem with our analysis is that many of the regressors in our model

are likely highly correlated[2]. This can cause Lasso to peform poorly. A solution the Elastic Net estimator, which is a combination of LASSO and Ridge Regression, since it provides more stability.

Table 2: Estimations of β with different penalty selection

	(1)	(2)
	PDL	PDL
β	-0.2608	-0.1354
se	1.7698	1.2494
n	71	71
p	21	21
Penalty term	BRT	BCCH
λ^{yz}	0.5395	0.8464
λ^{gx}	0.5618	1.1539

λ^{yz} is the error term in the first step and λ^{gx} in the second step of the PDL

5 Conclusion

In this project we have looked into the hypothesis of "catch-up growth", estimating a model testing it using the PDL estimator. Despite, visual evidence that there is a relation between growth and initial GDP, and using both BRT and BCCH penalty we find that initial GDP has an insignificant effect on growth and conclude that the data does not show evidence for 'catch-up growth".

[2]Take for example religion and region

Advanced Microeconometrics

Binary Response Model

Matias Hall
Copenhagen University
Department of Economics

Thomas Theodor Kjølbye
Copenhagen University
Department of Economics

November 2022

Mathias Hall is responsible for: Q1, Q3, Q5, Q7, Q9
Thomas Theodor Kjølbye is responsible for Q2, Q4, Q6, Q8, Q10:

1 Introduction: Binary Response Model

An outcome is generated according to the binary response model

$$y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$y_i = \mathbb{1}\{y_i^* > 0\}$$

where, in accordance with the exam, y_i^* is a latent variable, x_i is a random variable, and ε_i is a Cauchy distributed random variable.

2 Theoretical section

Q1

The 'success' probability $p(y_i = 1|\ x_i)$ is derived by first realising that $y_i = 1 \Leftrightarrow y_i^* > 0$ which we substitute in for the observable variable, y_i. Second, we substitute in the definition of the latent variable, y_i^*. Finally, we leverage the property of symmetry $F(a) = 1 - F(-a)$ and the definition of the cumulative distribution function (CDF) $F(a) := p(A \leq a)$ to complete the derivation

$$\begin{aligned}
p(y_i = 1|\ x_i) &= p(y_i^* > 0|\ x_i) \\
&= p(\epsilon_i > -\beta_0 - \beta_1 x_i|\ x_i) \\
&= 1 - p(\epsilon_i \leq -\beta_0 - \beta_1 x_i|\ x_i) \\
&\overset{\text{def}}{=} 1 - F(-\beta_0 - \beta_1 x_i; \mu) \\
&\overset{\text{sym}}{=} F(\beta_0 + \beta_1 x_i; \mu) \\
&= \frac{1}{2} + \frac{1}{\pi} \arctan(\beta_0 + \beta_1 x_i - \mu) \qquad (2.1)
\end{aligned}$$

Q2

We leverage the log-likelihood contribution derived in Q3 to prove by contradiction that the model mentioned above in section 1 is not identifiable. The log-likelihood

contribution $\ell_i(\boldsymbol{\beta}, \mu)$ of the ith observation is given by

$$\ell_i(\boldsymbol{\beta}, \mu) = y_i \log\left(\frac{1}{2} + \frac{1}{\pi}\arctan(\beta_0 + \beta_1 x_i - \mu)\right) + (1 - y_i)\log\left(\frac{1}{2} - \frac{1}{\pi}\arctan(\beta_0 + \beta_1 x_i - \mu)\right) \tag{2.2}$$

where $\boldsymbol{\beta} := (\beta_0, \beta_1)'$. Pursuing to Q3, we have not imposed the restriction $\mu = 0$. Now, recall from the lecture "M-Estimation I: Introduction and Asymptotic properties" that likelihood-based estimators belong to - and are special cases of - the broader M-estimator framework that are uniquely identified when

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \;\; \mathbb{E}[\mathrm{q}(\mathbf{w}; \boldsymbol{\theta})] \tag{2.3}$$

where $\mathrm{q}(\mathbf{w}; \boldsymbol{\theta})$ is the criterion function, which in turn is a function of observables, $\mathbf{w} := (y_i, x_i)$, and parameters, $\boldsymbol{\theta} := (\beta_0, \beta_1, \mu)'$. The criterion function consistent with the log-likelihood contribution in (2.2) is given by

$$\mathrm{q}(\mathbf{w}_i; \boldsymbol{\theta}) = -y_i \log\left(\frac{1}{2} + \frac{1}{\pi}\arctan(\beta_0 + \beta_1 x_i - \mu)\right) - (1 - y_i)\log\left(\frac{1}{2} - \frac{1}{\pi}\arctan(\beta_0 + \beta_1 x_i - \mu)\right) \tag{2.4}$$

Now, consider the parameter vector, $\boldsymbol{\theta} = (k + \beta_0, \beta_1, k + \mu)$[1]. If we insert these values into the criterion function in (2.4) the ks cancel out, leaving the criterion function untouched. Consequently, without imposing additional structure on either β_0 or μ there is no uniquely defined solution to (2.3) - In fact, there is an infinite number of solutions, implying that we cannot identify the parameters[2].

[1]The criterion function is then

$$\mathrm{q}(\mathbf{w}_i; \boldsymbol{\theta}) = -y_i \log\left(\frac{1}{2} + \frac{1}{\pi}\arctan(\beta_0 + \beta_1 x_i - \mu)\right) - (1 - y_i)\log\left(\frac{1}{2} - \frac{1}{\pi}\arctan(k + \beta_0 + \beta_1 x_i - (k + \mu))\right)$$

[2]Another approach to tackling Q2 could be that of a simulation. In particular, we could generate data and display y_i and y_i^* for different sets of parameters. We would then observe that while the latent variable would have changed, the observable would not, thus showing non-identification as there is no unique set of parameters which generates the obseverable data.

Q3

We recognize y as a Bernoulli distributed random variable with PDF[3]

$$f(y) = p(y=1)^{\mathbb{1}_{\{y=1\}}} p(y=0)^{\mathbb{1}_{\{y=0\}}} \tag{2.5}$$

To derive the log-likelihood contribution $\ell_i(\boldsymbol{\beta}, \mu)$, we consider specific realizations, denoted by subscript i. In addition, we take the natural logarithm and condition on x i.e. $logf(y_i|x_i)$ is

$$\mathbb{1}_{\{y_i=1\}} \log[p(y_i = 1|\ x_i)] + \mathbb{1}_{\{y_i=0\}} \log[p(y_i = 0|\ x_i)] \tag{2.6}$$

Using $p(y_i = 0|\ x_i) = 1 - p(y_i = 1|\ x_i)$ and reworking the indicator functions we get

$$y_i \log[p(y_i = 1|\ x_i)] + (1 - y_i) \log[1 - p(y_i = 1|\ x_i)] \tag{2.7}$$

In closing, leveraging our result from Q1

$$p(y_i = 1|\ x_i) = F(\beta_0 + \beta_1 x_i; \mu) \tag{2.8}$$

where $F(\cdot)$ is the CDF of the Cauchy distribution, we obtain the log-likelihood contributions

$$\begin{aligned}
\ell_i(\boldsymbol{\beta}, \mu) &= y_i \log[F(\beta_0 + \beta_1 x_i; \mu)] + (1 - y_i) \log[1 - F(\beta_0 + \beta_1 x_i; \mu)] \\
&= y_i \log\left(\frac{1}{2} + \frac{1}{\pi}\arctan(\beta_0 + \beta_1 x_i - \mu)\right) \\
&\qquad + (1 - y_i) \log\left(\frac{1}{2} - \frac{1}{\pi}\arctan(\beta_0 + \beta_1 x_i - \mu)\right)
\end{aligned} \tag{2.9}$$

From which it is straightforward to set $\mu = 0$ to obtain the log-likelihood contribution $\ell_i(\boldsymbol{\beta})$ reported in the exam.

[3]Strictly speaking, y_i is a discrete random variable which is why the PDF is the probability mass function. In addition, the binary response model specifies a model for the conditional mean of y given x. As a result, we condition on x.

3 Cross Section Data Analysis

Q4

As stated in Q2 we can estimate model defined in the intro in the M-estimation framework, with the log-likelihood contributions as the criterion function by solving the sample problem:

$$\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{N} \sum_{i=1}^{N} q(\boldsymbol{w}_i, \boldsymbol{\theta}) \tag{3.1}$$

With N being the number of observations and $\boldsymbol{w}_i = (y_i,\, x_i)$. We solve the problem for the M-estimator, $\hat{\theta}$, with a numerical optimiser. We use the BFGS method, which is a gradient based optimiser. Since the criterion function is smooth and differentiable a gradient based optimiser will find the optimum efficiently. Under the assumption of (Wooldridge 2010) Theorem 12.1 and assuming that the M-estimand $\boldsymbol{\theta}_0$ is identified, the M-estimator solves equation 3.1 and is consistent. Linking this to the condtional maximum likelihood estimator we use, it is by construction continuous and strictly concave in $\boldsymbol{\theta}$, why as long as the distribution for ε_i is correctly specified, $\hat{\theta}$ is consistent. We assume that ε_i is correctly specified, as we have no way to argue for or against it. Provided, that Theorem 12.3 also holds the M-estimator is asymptotically normal. We use this property to compute asymptotic variance on the estimates. Further, under the conditional information matrix equality(CIME) we get $\boldsymbol{A}_0 = \boldsymbol{B}_0$ and the asymptotic variance simplify to $Avar(\hat{\theta}) = \boldsymbol{A}_0^{-1}/N$.[4] We use the following specification of $\boldsymbol{A}_0$ as $\boldsymbol{A}_0 == \big(-\sum_{i=1}^{N} \boldsymbol{H}_i(\hat{\theta})\big)^{-1}$ as it provides a good trade-off, between low effort computation and efficiency.

The results are reported below in table 1. Both estimates are significant. We find that an increase to x_i increases the latent variable, thereby, increasing the probability of success. We will compute by how much in the next question. We also find that the mean probability of success given x_i, $\frac{1}{N}\sum_{i=1}^{N}[F(\beta_0 + \beta_1 x_i)] = 86.72$ pct.

	Estimate	Standard error	t-value
$\hat{\beta}_0$	0.9355	0.1410	6.6339
$\hat{\beta}_1$	2.6806	0.5284	5.0726

Table 1: Model parameters for cross-sectional data

[4] $\boldsymbol{A}_0$ and $\boldsymbol{B}_0$ are defined as in the maximum likelihood slides by Jesper Sørensen

Q5-Q6

More often than not, in any discrete or binary response model, the object of interest is the partial effect of x_i on the response probability. In the conventional simple linear model, such effects are naively computed as $\frac{\partial \mathbb{E}(y|\ x)}{\partial x_j} = \beta_j$, and thus independent of x. Generally - assuming the effects are not governed by a linear process - the partial effects ought to be evaluated at some $x = x^0$. While the binary response models accommodate this shortcoming, the models complicate the interpretation of the β - in particular, the magnitude of the β becomes difficult to interpret. For partial effects β decides the sign but not the full magnitude of the effect. Instead, we compute the partial effects of the discrete covariates as

$$\text{PE}_{x=1} = p(y = 1|\ x = 1) - p(y = 1|\ x = 0)$$
$$\text{PE}_{x=2} = p(y = 1|\ x = 2) - p(y = 1|\ x = 1)$$

where $p(\cdot)$ is simply given by the Cauchy CDF, owing to the result in Q1[5]. Rewriting we get

$$\text{PE}_{x=1} = F(\boldsymbol{z}\hat{\boldsymbol{\beta}}|\ x = 1) - F(\boldsymbol{z}\hat{\boldsymbol{\beta}}|\ x = 0) \tag{3.2}$$
$$\text{PE}_{x=2} = F(\boldsymbol{z}\hat{\boldsymbol{\beta}}|\ x = 2) - F(\boldsymbol{z}\hat{\boldsymbol{\beta}}|\ x = 1) \tag{3.3}$$

Where $\hat{\boldsymbol{\beta}}$ is the estimated coefficients reported in Q4 and $\boldsymbol{z} = (1, x)$. There is a continuous equivalent, though omitted here as we consider only discrete variables. We assess that, since the variable only take discrete values, that an effect around a point is meaningless, why we instead look at the full effect of jumping from one value to the next.

Table 2 reports the estimated partial effects. We find that going from $x = 0$ to $x = 1$ increases the probability of success with 17.5 pct-points while going from $x = 1$ to $x = 2$, the probability of success increases only 3.6 pct-points.

[5]The partial effect from $x = 0$ to $x = 2$ is identical to the sum of the two partial effects presented in table 1.

	Estimate	Standard error	t-value
$PE_{x=1}$	0.175	0.026	6.680
$PE_{x=2}$	0.036	0.004	9.414

Table 2: Partial Effects for the cross-sectional data

In closing, the standard errors of the partial effects are computed by the delta method

$$\text{Avar}(h(\hat{\beta})) \approx [\nabla h(\hat{\beta})]\text{Avar}(\hat{\beta})[\nabla h(\hat{\beta})]' \tag{3.4}$$

where $h(\hat{\beta})$ is the partial effect of the discrete random variable defined in equations (3.2) and (3.3).

Q7

To test the hypothesis "x_i has no impact on the probability of success.", we can either test for $\beta_1 = 0$ or $PE_{x=1} = 0 \vee PE_{x=2} = 0$. We argue that the second option is the most relevant in regards to the hypothesis, as β_1 does not have a direct interpretation in terms of the probability of success. Hence, we have two null with fellow alternative hypothesis.

$$H_0 : PE_{x=1} = 0, \quad H_A : PE_{x=1} \neq 0 \tag{3.5}$$

$$H_0 : PE_{x=2} = 0, \quad H_A : PE_{x=2} \neq 0 \tag{3.6}$$

We test using a Wald test. We compute the Wald test statistic as

$$W = \frac{(\hat{\theta} - \theta_0)^2}{var(\hat{\theta})}$$

Where θ_0 is the hypothesised value. We have that $W \overset{a}{\sim} \chi^2_k$, where k is the number of restrictions. We get a Wald-statistic of 50.43 and 94.78 respectively. Both are well outside the critical value and we reject both nulls. Hence, we conclude that x_i has an impact on the probability of success.

4 Panel Data Analysis

Q8

We start by remembering that the likelihood contribution from the ith observation is defined as $f(y_i|x_i)$ where f is a probability density function(PDF). We begin the derivation there and then use the definition of marginal conditional densities and conditional densities. We can then use the fact that c_i is IID. Defining the normal PDF of a random variable, a as $\phi(a)$ we can substitute out $f(x)$ as well.

$$\begin{aligned} f(y_i|x_i) &= \int_{-\infty}^{\infty} f(y_i,c_i|x_i)\, dc \\ &= \int_{-\infty}^{\infty} f(y_i|x_i,c_i) f(c|x_i)\, dc \\ &= \int_{-\infty}^{\infty} f(y_i|x_i,c_i)\phi(c)\, dc \end{aligned}$$

Next we take the log and using dynamic completeness[6]. We also insert the PDF for y, the Bernoulli distribution as mentioned in Q3.

$$\begin{aligned} \log f(y_i|x_i) &= \log \int_{-\infty}^{\infty} \prod_{t=1}^{T} f(y_{it}|x_{it},c_i)\phi(c)\, dc \\ &= \log \int_{-\infty}^{\infty} \prod_{t=1}^{T} p(y_{it}=1|x_{it},c_i)^{\mathbb{1}\{y_{it}=1\}} p(y_{it}=0|x_{it},c_i)^{\mathbb{1}\{y_{it}=0\}} \phi(c)\, dc \end{aligned}$$

Deriving $p(y_{it}=1|x_{it},c_i)$ is very similar to the derivations in Q1 and yields

$$\begin{aligned} p(y_{it}=1|x_{it},c_i) &= F(\beta_0+\beta_1 x_{it}+c_i) \\ &= \frac{1}{2}+\frac{1}{\pi}\arctan(\beta_0+\beta_1 x_{it}+c_i) \end{aligned} \tag{4.1}$$

We then rework the indicator functions, and insert equation (4.1). Finally we define $c_i = \sigma_c c$. We then get

[6]Since y_{it} is serially independent if y_{it}^* is, and since y_{it}^* is serially independent conditioning on x_{it} and c_i and u_{it} is IID also conditioning on x_{it} and c_i, dynamic completeness holds

$$\ell_i(\boldsymbol{\theta}) = \log\left(\int_{-\infty}^{\infty}\prod_{t=1}^{T}\left[\frac{1}{2}+\frac{1}{\pi}\arctan(\beta_0+\beta_1 x_{it}+\sigma_c c)\right]^{y_{it}}\right.$$
$$\left.\cdot\left[\frac{1}{2}-\frac{1}{\pi}\arctan(\beta_0+\beta_1 x_{it}+\sigma_c c)\right]^{1-y_{it}}\phi(c)\,dc\right) \tag{4.2}$$

Where $\boldsymbol{\theta}$ is as defined in Q9. Equation (4.2) is exactly what we should show.

Q9

The model specified in the exam is a binary response model with random effects. Since the model includes time variation we can no longer solve by maximum likelihood(MLE) but instead need to utilise simulated MLE. In order to solve the model, we need to solve the integral in equation (4.2). We can do so by simulation or using quadrature[7]. We opt for quadrature due to the superior precision. Since, we are working with a low dimensional model, extra notes require little computing power, why, we include 24 nodes. More nodes improve precision, although when using quadrature for a polynomial of order $2n-1$ the number of nodes beyond n only improves precision by a small amount. Again, since the computing power required is so low, we opt for the barely improved precision anyways. We have checked the results with simulation and we get the exact same results. The same properties as described in Q4 needs to be valid for identification, consistency and efficiency, since, it is an M-estimator. We report the results below in tablE 3.All estimates are significant. Again we find a positive $\hat{\beta}_1$ so higher x leads to a higher probability of success. The mean probability of success is approximately[8]

	Estimate	Standard error	t-value
$\hat{\beta}_0$	0.6571	0.0414	15.8908
$\hat{\beta}_1$	0.5296	0.0517	10.2357
$\hat{\sigma_c}$	0.7252	0.0425	17.0538

Table 3: Model parameters for panel data

[7]Because $f(c) = \phi(c)$ is normally distributed

[8]Since we do not observe c we can only approximate it by simulating c's with the found distribution

Q10

The methods leveraged to compute the partial effects and their standard errors are covered in Q5/Q6. For the case were c is imputed to zero, what we will now call the partial effect on the average[9] it is straight forward to just set $c = 0$ in $F(\boldsymbol{z}_{panel}\hat{\boldsymbol{\theta}})$ so it becomes $F(\boldsymbol{z}\hat{\boldsymbol{\beta}})$, where $\boldsymbol{z}_{panel} = (1,x,c)$. The expected (wrt. c_i) partial effect, $\mathbb{E}_c[F(\boldsymbol{z}_{panel}\hat{\boldsymbol{\theta}})]$ we compute utilising quadrature. We use 24 nodes again for consistency. For the standard errors we do the same, that is set $c = 0$ for the first and use quadrature for the second. We report the results below in table 4.

	Estimate	Standard error	t-value
$PE_{c=0}$	0.0921	0.0085	10.7774
$PE_{\mathbb{E}_c}$	0.1706	0.0154	11.0513

Table 4: Partial Effects for the panel data

The difference between the two effects are if you are interested in the partial effect of x on the average person, or if you are interested in the partial effect of x happening or not. This is best illustrated with an example. If y is the probability of of passing the Advanced Microeconometrics exam and x is whether the student attended exercise classes. Then we can think of c as inherent ability. Setting $c = 0$ then becomes the effect of attending exercise classes for the average student, while the expected partial effect becomes the expected effect of having attended exercise classes v. not attending. In effect the later is an analogy to the average partial effect. If we continue our analogy with the exam, the results then show that the effect on the average student is smaller compared to the average partial effect. This means that the effect on the students with less ability is very large. Hence, in this case the biggest payoff is to get the students with less ability to attend more, while getting average students to attend is less important.

[9]Since the average of x_i is not 0 this is not quite true, but since there can be no effect if $x_i = 1$ we can think of the case this way

References

Wooldridge, Jefrey M. 2010. *Econometric analysis of cross section and panel data.* The MIT Press. ISBN: 9780262232586, accessed January 14, 2023. http://www.jstor.org/stable/j.ctt5hhcfr.