Predicting Child Labour Risks by Norms in India

By Jihye Kim, Wendy Olsen, and Arkadiusz Wiśniowski

published in ***Work, Employment and Society* 2022**

This Appendix is at Online URL https://github.com/WendyOlsen/normslabourindia

And sj-pdf-1-wes-10.1177_09500170221091886 – Supplemental material for Predicting Child-Labour Risks by Norms in India

Corresponding author Jihye.kim@manchester.ac.uk

Appendix

Table A.1. Observed Numbers of Out-of-School Children by Working Hours (Zt)

Daily working hours[1)]	Count of out-of-school children			Sample no. of children		
	Ages 5–11	Ages 12–14	Ages 15–17	Ages 5–11	Ages 12–14	Ages 15–17
0	3,094	908	1,564	32,057	14,127	9,344
1	18	51	88	705	816	674
2	25	82	146	571	865	855
3	4	64	119	234	411	524
4	18	49	119	163	330	430
5	6	33	97	38	192	271
6	4	52	153	51	147	289
7	3	26	83	7	53	119
8	12	196	699	29	283	883
9	0	28	89	8	48	128
10	2	32	131	6	43	184
11	1	10	67	5	23	99
12	0	17	118	3	21	162
13	0	7	39	1	12	55
14	0	11	60	1	11	72
15	NA	1	17	NA	2	26
16	2	23	87	3	29	106
Total	3,189	1,590	3,676	33,882	17,413	14,221

Table A.2. Estimating Thresholds of Child Labour Hours

The time threshold model uses a Poisson distribution that is designed to estimate child labour hours that bring a significant change in children's education status. The model uses counting data $\mathbf{Z_t}$ which is the weighted count of out-of-school children at each hour of work; ***t*** denotes the number of daily working hours. The key parameter $\boldsymbol{\lambda_t}$ represents the risk rate of children being out-of-school when they work for ***t*** hours per day. The time index parameter $\boldsymbol{\theta}$ indicates daily working hours after which educational harms increase sharply. Thus, the risk parameter $\boldsymbol{\lambda_t}$ relies on two given conditions: one is $\boldsymbol{\lambda_1}$, which is the risk parameter when children's daily working hours number less than $\boldsymbol{\theta}$, and the other is $\boldsymbol{\lambda_2}$, the risk parameter when working hours are equal or more than $\boldsymbol{\theta}$. If the two parameters, $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$, are not identified, $\boldsymbol{\lambda_t}$ will have diverged numbers. Otherwise, it has two values, which are $\boldsymbol{\lambda_1}$, if $\mathbf{t} < \boldsymbol{\theta}$, or $\boldsymbol{\lambda_2}$, if $\mathbf{t} \geq \boldsymbol{\theta}$. Two conditional parameters, $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$, are positive rates; therefore, a uniform distribution of (0, 1) is assumed. The parameter $\boldsymbol{\theta}$ is supposed to have an integer uniform prior between 0 and 16 hours (maximum daily working hours given by the data). It was assumed that a continuous distribution for the time index $\boldsymbol{\theta}$ as working hours is, in principle, a continuous variable, operationalised in the data by integers. A continuous prior permits a more precise estimation of the time index than when using, for example, an integer uniform prior.

$$Z_t \sim \text{Poisson}(\lambda_t * n_t),\ t = 0, \ldots, 16 \quad (1)$$

$$\lambda_t = \begin{cases} \lambda_1, & \text{if } t \in [0, \Theta] \\ \lambda_2, & \text{if } t \in [\Theta, 16], \end{cases} \quad (2)$$

$$\lambda_1 \sim \text{uniform}(0, 1), \quad (3)$$

$$\lambda_2 \sim \text{uniform}(0, 1), \quad (4)$$

$$\Theta \sim \text{uniform}\ (0,16) \quad (5)$$

The same model is repeated for three age groups: ages 5–11, ages 12–14 and ages 15–17. After running the model, estimates of $\boldsymbol{\theta}$ are multiplied by seven under the assumption that children work for seven days a week and then they are considered as weekly time thresholds of child labour in the norm models.

Table A.3. Latent Variables

1) The Norm for Women's Work

Confirmatory factor analysis with an ordinal logit function was implemented to identify gender norms explaining people's attitudes about women's work. In implementing this, four items were chosen from the World Value Survey 2012. V45 took value 1 if people agree, 2 if people neither agree nor disagree and 3 if people disagree. V50, V52, and V54 were ordinal variables (1-4) indicating whether people agree/disagree or strongly agree/disagree. Missing values were excluded. The list of variables are below:

(V45) When jobs are scarce, men should have more right to a job than women. (1=agree, 2=neither agree nor disagree, 3=disagree)
(V50) When a mother works for pay, the children suffer. (1=strongly agree, 2=agree, 3=disagree 4=strongly disagree)
(V52) A university education is more important for a boy than for a girl. (1=strongly agree, 2=agree, 3=disagree 4=strongly disagree)
(V54) Being a housewife is just as fulfilling as working for pay (1=strongly agree, 2=agree, 3=disagree 4=strongly disagree)

All variables had significant positive coefficients for the latent variable. The latent variable is associated with higher support for the higher education of girls (V52, coef=1.33); mother's work for pay (V50, coef=0.7); women's working for pay (V54, coef=0.52) and women's right to a job (V45 was constrained as 1). Cronbach alpha test showed the items are weakly correlated (0.44), but the reason for the index is not to have a great fit but to reduce the complexity of regression. The covariance of the gender norm index is significant, indicating that it represents well whether people have a supportive norm on women's employment.

2) Asset Index Score

Similarly, confirmatory factor analysis was used to create the asset index category based on the IHDS 2011/12. The selected variables were toilet facilities, floor materials, the ownership of a vehicle, phone, and TV. Cronbach alpha test showed that the variables are moderately correlated (0.66). The ownership of a phone was constrained as 1. TV (coef=1.56) and floor (coef=1.55) have a strong positive coefficient value. Toilet facilities (coef =1.05), roof (coef=0.63) and vehicle (coef=0.32) also explained the latent variable. The variance of the latent variable is significant, indicating the appropriateness of the use of the variable as a proxy of the asset index.

Table A.4. Correlation Matrix

	Child labour	Female	Urban	Lowest assets	Land category	Dalit	Adivasi	Norm 1	Norm 2	Norm 3
Child labour	1.00									
Female	-0.11	1.00								
Urban	-0.37	0.00	1.00							
Lowest assets	0.54	0.00	-0.49	1.00						
Land category	0.00	0.00	-0.49	0.02	1.00					
Dalit	0.17	0.00	-0.10	0.12	-0.38	1.00				
Adivasi	0.20	0.00	-0.31	0.48	0.14	-0.14	1.00			
Norm 1 (seclusion)	0.22	0.00	-0.10	0.43	0.13	-0.10	0.09	1.00		
Norm 2 (benevolence)	-0.04	0.00	-0.30	-0.11	0.07	0.09	0.01	-0.08	1.00	
Norm 3 (women's work)	-0.25	0.00	0.19	-0.37	-0.07	0.14	-0.26	-0.52	-0.32	1.00

Table A.5. Model-checking

A Bayesian posterior predictive p-value of the likelihood ratio statistics showed that the threshold models demonstrated no significant discrepancy between predictions and observations (Lunn et al., 2012; Zhang, 2014). As the threshold model did not include any explanatory variables, a Bayesian R-squared test[i] was not applied. For children aged 5 to 11 years, the data were sparse and contained some missing values as well as zero counts that are likely an effect of low exposures. This did not allow the model parameters to be identified (the MCMC algorithm did not achieve convergence). Therefore, the results for this age group are not presented in the following sections and this investigation is left for future research.

1) Test Statistics of Threshold Models

Models	Bayesian p-value	MSE	DIC	Deviances	pD
Ages 5–11	0.45	375.62	173.6	138.95	34.7
Ages 12–14	0.48	190.49	212.7	210.70	2.0
Ages 15–17	0.65	429.00	221.1	219.10	2.0

Notes: a Bayesian p-value of the likelihood ratio statistics; MSE=Mean squared errors; DIC=deviance+pD; pD = var(deviance)/2

In the main models, an out-of-sample test was implemented: for this, one occupational group (each occupational group included 17 cases, i.e. 17 states) was excluded from the total sample. Each simulation predicted the number of child labourers from the left-out samples. The test results showed that around 40 percent of simulations predicted the number of child labourers with a Bayesian posterior predicted p-value between 0.05–0.95, indicating that the models did not generate extreme predictions. Furthermore, a Bayesian R-squared test confirmed that all three models showed high enough Bayesian R-squared values: more than 50 percent of the variation was explained by each model. Model 3 provided a high value for explained variance at around 70–80 percent. In terms of DIC, Model 3, which included state-level norms and interactions, was the best performing model as it had the lowest value of DIC. There was no sign of overdispersion in the three models. Thus, a Poisson distribution sufficiently describes the variability of the data.

2) Out-of-sample Test Using Norm Models

	Bayesian R-squared test			Bayesian P-value test			DIC		
Models	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Simulation 1	0.69	0.75	0.76	1.00	0.99	0.97	2176.70	2117.60	2099.30
Simulation 2	0.59	0.62	0.66	0.81	0.80	0.75	2273.60	2208.30	2188.20
Simulation 3	0.81	0.85	0.83	0.72	0.62	0.78	2160.10	2111.70	2093.40
Simulation 4	0.51	0.56	0.56	0.98	0.98	0.96	2153.10	2092.50	2056.20
Simulation 5	0.30	0.33	0.29	1.00	1.00	1.00	2252.00	2188.10	2162.30
Simulation 6	0.74	0.80	0.83	1.00	0.99	0.98	2204.20	2150.00	2131.40
Simulation 7	0.91	0.88	0.89	0.05	0.01	0.02	2231.20	2151.10	2129.30
Simulation 8	0.51	0.55	0.60	0.67	0.65	0.58	2250.30	2188.80	2171.60
Simulation 9	0.73	0.69	0.68	0.01	0.01	0.01	2249.00	2177.40	2157.30
Simulation 10	0.76	0.81	0.83	0.94	0.83	0.69	2220.30	2145.00	2121.40

Notes: a Bayesian p-value of the likelihood ratio statistics (Lunn et al., 2013)

Table A.6. Rstan Code for the Model

The code is available at URL https://github.com/WendyOlsen/normslabourindia

[i] A Bayesian R-squared is 'the variance of the predicted values divided by the variance of predicted values plus the expected variance of the errors' (Gelman et al., 2019, p.307).