WORLD BANK EDUCATION GLOBAL PRACTICE

Russian Federation: Analytical Services and Advisory Activity: P170978

Returns to Education in the Russian Federation: Application of a machine learning instrumental variable technique for policy development in priority regions

Ekaterina Melianova* |**Suhas Parandekar*** |**Harry Patrinos*** |**Artëm Volgin***

*Education Global Practice, Europe and Central Asia

Correspondence
Email: sparandekar@worldbank.org

Data and Code
Thanks are due to the Higher School of Economics, Moscow for making the Russian Longitudinal Monitoring Study (RLMS) Household data readily available for reseachers around the world. The code used for this paper is made freely available for all researchers at `https://bitbucket.org/zagamog/edreru/src/master/`

This is a generic template designed for use by multiple journals, which includes several options for customization. Please consult the author guidelines for the journal to which you are submitting in order to confirm that your manuscript will comply with the journal's requirements. Please replace this text with your abstract.

KEYWORDS
keyword 1, *keyword 2*, keyword 3, keyword 4, keyword 5, keyword 6, keyword 7

1 | MOTIVATION FOR THIS PAPER

1.1 | Stylized fact 1: Returns trend and looking at cohorts

1.2 | Stylized fact 2: Priority regions and policy options for education

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.

Acknowledgments: *Country Director:* Renaud Seligman; *Regional Director:* Fadia Saadah; *Practice Manager:* Harry Patrinos; *Program Leader:* Dorota Nowak; *Peer Reviewers*: Cristian Aedo; Ruslan Yemtsov; Husein Abdul-Hamid; *Team members:* Polina Zavalina; Zhanna Terlyga. Thanks to seminar participants at the World Bank Moscow office on Jan. 29, 2020 for useful feedback. Any errors are a responsibility of the authors.

Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

First level:

$$LogWage_{ij} = b_{0j} + b_{1j} \cdot Educ + b_{2j} \cdot Exp + b_{3j} \cdot Exp^2 + b_{4j} \cdot Gender + \epsilon_{ij} \quad (1)$$

Second Level:

$$b_{0j} = \gamma_{00} + \gamma_{0n} \cdot Z + u_{00}; \qquad b_{1j} = \gamma_{10} + \gamma_{1n} \cdot Z + u_{10}; \qquad b_{ij} = \gamma_{i0} \quad for \quad i \neq 0 \quad (2)$$

where an individual i is nested withing a region j, $LogWage$ is a logarithm of monthly wage, $Educ$ stands for highest attained level of education, Exp and Exp^2 reflect the years of working experience and its quadratic term respectively, $Gender$ is a dummy variable for gender, Z is an $n \times i$ matrix of regional characteristics, ϵ and u_{00}, u_{10} are the first- and second-level errors respectively.

We refer to **card_070._1999** and belloni$_0$52.$_2$011.

2 | LITERATURE REVIEW: USE OF INSTRUMENTAL VARIABLES IN RETURNS TO EDUCATION STUDIES

General Literature

Russia Specific

The usage of instrumental variables as a way of correcting for the endogeneity of educational attainments in the estimation of returns to schooling in Russia is relatively rare. Nevertheless, there are several papers that leveraged IV technique and made a number of important conclusions.

One research ascertained that during the transition returns to education in Russia were not improving and remained among the most deficient in the world (**cheidvasser_educated_2007**). The researchers instrumented years of schooling, employing a policy experiment in Russia from the 1950s to 1960s, and corrected for a selectivity bias by adding an equation for the labor market participation. It was highlighted that the excess of well-educated workers seemed to be the main underpinning factor of wage differentials in Russia after Soviet Union dissolution. Additionally, the study showed that heterogeneity in rates of returns to education in Russia also hails from gender differences similar to the global patterns: women receive greater returns to higher education than men.

Utilizing the same instruments as **cheidvasser_educated_2007**, **akhmedjonov_higher_2011** evaluated the magnitude of the education premiums in Russia between 2000 and 2002. Both OLS and 2SLS estimates (8% and 19.1% respectively) were indicative of significant returns and the formation of a more flexible wage structure in the Russian Federation.

arabsheibani_returns_2012 introduced the age of sexual intercourse as another instrumental variable to tackle endogeneity in schooling. In line with some past research, the scholars show that OLS undervalues returns to education compared to the IV estimates. The study adopts an instrumental variable quantile approach over the wage distribution in addition to the conditional mean estimation.

kyui_expansion_2016, using the amount available slots as an instrument, showed that returns to schooling in Russia declined for those who took advantage of higher education expansion in a post-communist Russia (1990-2005) in comparison to youths who obtained university degree in preceding periods. In an earlier study, **kyui_returns_2010** used the accessibility of tertiary education and the education levels of other household members as instruments in a wage equation (2010). The scholar demonstrated that a growth in returns to education in the Russian Federation does not indicate its closeness to the developed countries (**kyui_returns_2010**).

belskaya_college_2014 evaluated a large-scale college expansion in Russia after the breakdown of the Soviet Union (2014). Using the number of campuses in the municipality of residence at age 17 as an instrument, the research contended that as the number of university campuses grew, individuals with low returns to schooling grew as well. But for a marginal person, who switched into a treatment group as a result of new campuses opening, the total gains from attending a college are considerable and positive. Furthermore, the scholars found that students with higher returns are attracted more intensively by new campuses opened in constrained municipalities (small non-capital cities or those lacking higher education institutions before college expansion) in comparison to the unconstrained ones.

3 | LASSO AND POST-LASSO INSTRUMENTAL VARIABLES: OVERVIEW OF ESTIMATION METHOD

This study uses post-Lasso IV estimation, which was elaborated by **belloni_high_2011** and **belloni_sparse_2012** a decade ago. The method performs an optimal IV selection, using Lasso technique.

The Lasso is a regression analysis method, popularized by Tibshirani **tibshirani_regression_1996**, that does both variable selection and shrinkage (unlike Ridge, which only shrinks). The Lasso solves a regression problem with L1 penalization of finding (**diebold_econometric_2019**):

$$\hat{\beta}_{LASSO} = \text{argmin}_{\beta} \left(\sum_{i=1}^{N} \left(y_i - \sum_{i} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} |\beta_i| \right) \quad (3)$$

or equivalently:

$$\hat{\beta}_{LASSO} = \text{argmin}_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{i} \beta_i x_{it} \right)^2 \quad (4)$$

s.t.

$$\sum_{i=1}^{K} |\beta_i| \leq c \quad (5)$$

The equations portray that the Lasso loss function is composed of the least squares estimator (or can be extended to many other estimator) and a penalty term in a form of the absolute coefficient value. The main distinctive property of this technique is the ability conducting feature selection procedure by nullifying regression coefficients of low importance, which is of particular relevance in situations with large feature sets. In case of zero lambda, the method gets back to OLS, while huge lambdas reduce coefficients to zero and may lead to under-fitting.

4 | EMPIRICAL STUDIES USING POST-LASSO IV METHODOLOGY

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

5 | FINDINGS 1: TRADITIONAL INSTRUMENTAL VARIABLE TECHNIQUE

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

6 | FINDINGS 2: POST LASSO IV RESULTS

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

7 | CONCLUSIONS

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem.

Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Appendix

TABLE A1 Results of Estimating Human Capital Depreciation for the Female sample, RLMS

	1994	1998	2003	2006	2012	2018
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	9.725***	3.786***	5.464***	6.946***	8.133***	8.767***
	(0.381)	(0.322)	(0.301)	(0.247)	(0.186)	(0.242)
Educ, years (S)	0.122***	0.153***	0.158***	0.118***	0.087***	0.066***
	(0.025)	(0.022)	(0.020)	(0.016)	(0.012)	(0.015)
Educ X Exper (TS)	−0.002*	−0.002***	−0.002**	−0.0002	−0.0001	0.0004
	(0.001)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.001)
Exper (T)	0.074***	0.080***	0.055***	0.013	0.020**	0.020*
	(0.019)	(0.016)	(0.015)	(0.013)	(0.010)	(0.011)
Exper squared (T^2)	−0.001***	−0.001***	−0.001***	−0.0003**	−0.0005***	−0.001***
	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)
Observations	1,645	1,667	2,093	2,630	4,057	3,312
R^2	0.051	0.089	0.110	0.139	0.104	0.092
Adjusted R^2	0.049	0.087	0.108	0.138	0.103	0.091
Residual Std. Error	0.853	0.728	0.731	0.664	0.641	0.597
F Statistic	22.179***	40.520***	64.342***	106.385***	117.366***	83.993***

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE A2 Results of Estimating Human Capital Depreciation for the Male sample, RLMS

	1994	**1998**	**2003**	**2006**	**2012**	**2018**
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	10.357***	5.029***	7.334***	8.067***	8.771***	9.094***
	(0.433)	(0.360)	(0.282)	(0.243)	(0.157)	(0.185)
Educ, years (S)	0.136***	0.123***	0.080***	0.077***	0.077***	0.077***
	(0.028)	(0.024)	(0.019)	(0.016)	(0.010)	(0.012)
Educ X Exper (TS)	−0.002*	−0.001	0.0004	−0.0003	−0.0004	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.001)
Exper (T)	0.054**	0.032*	0.002	0.007	0.035***	0.037***
	(0.023)	(0.017)	(0.014)	(0.013)	(0.009)	(0.010)
Exper squared (T^2)	−0.001***	−0.0004**	−0.0003*	−0.0003*	−0.001***	−0.001***
	(0.0003)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)
Observations	1,392	1,433	1,763	2,170	3,360	2,800
R^2	0.057	0.070	0.078	0.074	0.153	0.110
Adjusted R^2	0.054	0.067	0.076	0.072	0.152	0.108
Residual Std. Error	0.951	0.803	0.754	0.688	0.598	0.570
F Statistic	20.989***	26.879***	37.362***	43.281***	151.868***	86.125***

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE A3 Results of Multilevel Modeling with Coverage by Vocational Education, Rosstat 2018

	Null model	Mincerian	Random Slope	Cross-Level Interaction
	(1)	(2)	(3)	(4)
Constant	10.178***	10.032***	10.056***	10.065***
	(0.034)	(0.034)	(0.036)	(0.036)
Vocational		0.283***	0.279***	0.267***
		(0.009)	(0.021)	(0.021)
Higher		0.638***	0.641***	0.622***
		(0.009)	(0.025)	(0.025)
Coverage VE X Vocational				0.050**
				(0.025)
Coverage VE X Higher				0.083***
				(0.030)
Experience		−0.026***	−0.027***	−0.027***
		(0.002)	(0.002)	(0.002)
Experience squared		−0.065***	−0.065***	−0.065***
		(0.002)	(0.002)	(0.002)
Females		−0.403***	−0.404***	−0.404***
		(0.005)	(0.005)	(0.005)
Coverage VE			−0.101***	−0.142***
			(0.039)	(0.043)
Variance of Intecept	0.09	0.08	0.09	0.09
Variance of Vocational			0.02	0.02
Variance of Higher			0.04	0.04
Residual Deviance	0.45	0.35	0.34	0.34
Observations	49,187	49,187	49,187	49,187
Log Likelihood	−59,755.060	−53,289.500	−53,094.620	−53,096.640
Akaike Inf. Crit.	119,516.100	106,595.000	106,217.200	106,225.300
Bayesian Inf. Crit.	119,542.500	106,665.400	106,340.500	106,366.100
Note:				*p<0.1; **p<0.05; ***p<0.01