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.

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 and Benítez-Silva 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 and Benítez-Silva (2007), Akhmedjonov (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 and Staneva (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 (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 (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 2010).

Belskaya, Sabirianova Peter, and Posso (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 considers post-Lasso IV estimation procedure, which was elaborated by Belloni and Chernozhukov (2011) and Belloni et al. (2012) a decade ago. The method performs an optimal IV selection, using the Lasso statistical learning algorithm.

The Lasso (Least Absolute Shrinkage and Selection Operator) is a regression analysis method, which was popularized by Tibshirani (1996). As it stems from the name, the method does both variable selection and shrinkage (unlike Ridge, which only shrinks). The Lasso solves a regression problem with L1 penalization of finding (Diebold 2019):

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

or equivalently:

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

s.t.

$$\sum_{i=1}^{K} |\beta_i| \leq c \tag{3}$$

where y and x are dependent and independent variables for $i = 1, 2, ..., N$, respectively; β is a regression coefficient of interest; λ is a tuning parameter, K is the number of degrees of freedom (d.f.) in the model fitting, t is the amount of independent variables.

The equations portray that the Lasso loss function is composed of the least squares estimator (or can be extended to many other estimators) and a penalty term in a form of the absolute coefficient value. The main distinctive property of this technique is the ability to conduct 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.

The d.f. results of Lasso are convenient: the effective number of parameters equals exactly to the number of the picked up (non-zero) variables. The best-fitting lambda can be chosen by reliance on various information criteria (e.g., AIC and SIC) or cross-validation. The additional advantage of the Lasso algorithm is its relative computational tractability due to the convex minimization problem (Diebold 2019).

The post-Lasso was devised for the instrument selection by excluding IVs with insignificant effects on the endogenous variable. Particularly, the post-Lasso algorithm consists of 2 steps: (1) IV selection in the first-stage regression using Lasso, (2) 2SLS estimation with the picked up instruments. The method allows to enhance the precision of inferences due to the advantage of implementing multiple instruments and avoiding the finite sample bias expansion. As long as the instruments of economic importance are filtered out, the erroneous exclusion of valid instrument with minor effects does not amend the 2SLS estimation (Belloni, Chernozhukov, and Hansen 2014). This is achieved due to the assumption that the conditional mean of a given endogenous attribute can be well approached by a small range of key instruments.

The Two-Stage Least Squares specification of interest in the present paper can be written by the following equations.

First stage:

$$x_{1i} = z_i' \pi_1 + x_{2i}' \pi_2 + v_i \tag{4}$$

Second stage:

$$y_i = x_{1i} \beta_1 + x_{2i}' \beta_2 + \varepsilon_i \tag{5}$$

where y is a logarithm of wages for $i = 1, 2, ..., N$; x_{1i} reflects years of education (an endogenous regressor); x_{2i} is a vector of exogenous variables: labor market experience and its squared term; z_i is a vector of instrumental variables, aggregated at a regional level in the Russian Federation: the number of higher education institutions, high school graduate students per school, standard deviation from the national average EGE scores (Russian school-leaving examination), net migration rate, women to men ratio, marriage rate, employment in female-dominated industries, and literacy level in 1897 in the Russian empire; β_1 is the causal effect of x_1 an y; ε_i and v_i are normally distributed error terms.

When a specified 2SLS model contains exogenous regressors, the Lasso selection technique is applied over the transformed first-stage equation where each side is multiplied by a projection matrix of control variables (i.e., they are partialed out). Then the Lasso estimator is given by:

$$\hat{\pi}_1^{\text{Lasso}} = \arg\min_{\pi_1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(x_{1i}^* - z_i^* \pi_1 \right)^2 + \frac{\lambda}{N} \|\pi_1\|_1 \right\} \tag{6}$$

where x_1^* and z^* are the projected endogenous regressor (education years) and IVs, respectively; π_1 is a regression coefficient of interest; λ stands for the regularization parameter.

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.

5 | FINDINGS 1: TRADITIONAL INSTRUMENTAL VARIABLE TECHNIQUE

Figure 5.1 displays correlations between the instrumental variables and years of education by cohorts and figure 5.2 is a correlation matrix between all the region-level instruments of focus: the number of higher education institutions *(HSGPER)*, high school graduate students per school *(high_n)*, standard deviation from the national average EGE scores *(s1z)*, net migration rate *(migrationrate)*, women to men ratio *(women2menratio)*, marriage rate *(marriagerate)*, employment in female-dominated industries *(migrationrate)*, and literacy level in 1897 in the Russian empire *(Literacy_97)*.

Instrumental Variables	Females older	Females younger	Males older	Males younger
women2menratio	0.01	0.02	0.01	0.02
s1z	0.15 ***	0.13 ***	0.17 ***	0.15 ***
migrationrate	0.17 ***	0.14 ***	0.23 ***	0.19 ***
marriagerate	0.05 ***	0.05 ***	0.07 ***	0.03 ***
Literacy_97	0.22 ***	0.18 ***	0.3 ***	0.26 ***
HSGPER	0.22 ***	0.18 ***	0.29 ***	0.24 ***
high_n	0.22 ***	0.17 ***	0.28 ***	0.23 ***
fem_ind_prop	-0.13 ***	-0.12 ***	-0.22 ***	-0.16 ***

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05*

FIGURE 5.1 Pearson correlation between Instrumental Variables and Years of Education by Cohorts, 2018

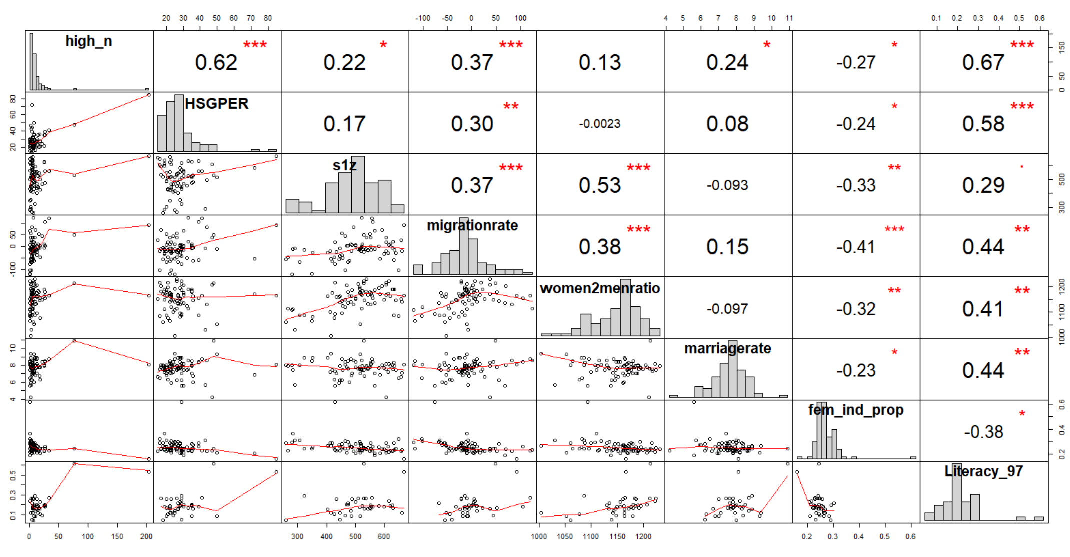


FIGURE 5.2 Pearson correlation Matrix and Distributions for Instrumental Variables, 2018

TABLE 5.1 Individual 2SLS Estimates (Standard Errors) for each IV by Cohorts and OLS Results, 2018

IVs	Female_young	N1	Female_old	N2	Male_young	N3	Male_old	N4
high_n	1.959 (0.229)	9779	0.857 (0.042)	11564	0.787 (0.047)	10636	0.64 (0.026)	9826
HSGPER	1.717 (0.174)	9779	0.887 (0.044)	11564	0.754 (0.043)	10636	0.636 (0.026)	9826
s1z	1.34 (0.195)	9779	0.602 (0.041)	11564	0.648 (0.06)	10636	0.501 (0.034)	9826
migrationrate	1.429 (0.163)	9779	0.796 (0.05)	11564	0.625 (0.045)	10636	0.534 (0.028)	9826
women2menratio	-0.439 (0.609)	9779	-0.99 (0.731)	11564	-0.105 (0.154)	10636	-2.998 (5.602)	9826
marriagerate	2.312 (0.536)	9779	2.18 (0.477)	11564	2.638 (0.705)	10636	1.988 (0.378)	9826
fem_ind_prop	1.677 (0.218)	9779	0.985 (0.079)	11564	0.829 (0.066)	10636	0.663 (0.034)	9826
Literacy_97	1.868 (0.232)	6834	0.912 (0.053)	7860	0.726 (0.044)	7247	0.641 (0.03)	6701
OLS	0.092 (0.003)	9775	0.107 (0.003)	11560	0.103 (0.003)	10632	0.114 (0.003)	9822

6 | FINDINGS 2: POST LASSO IV RESULTS

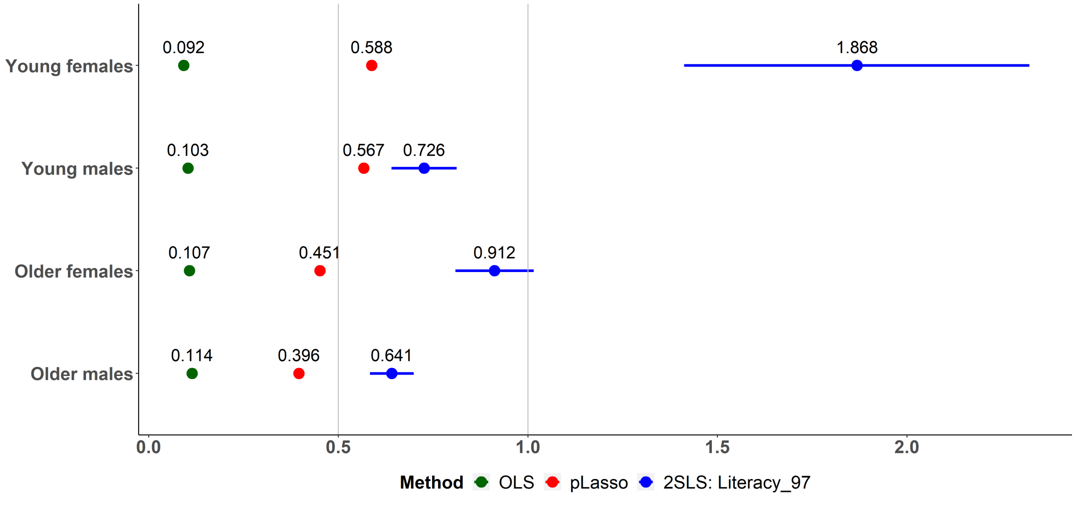
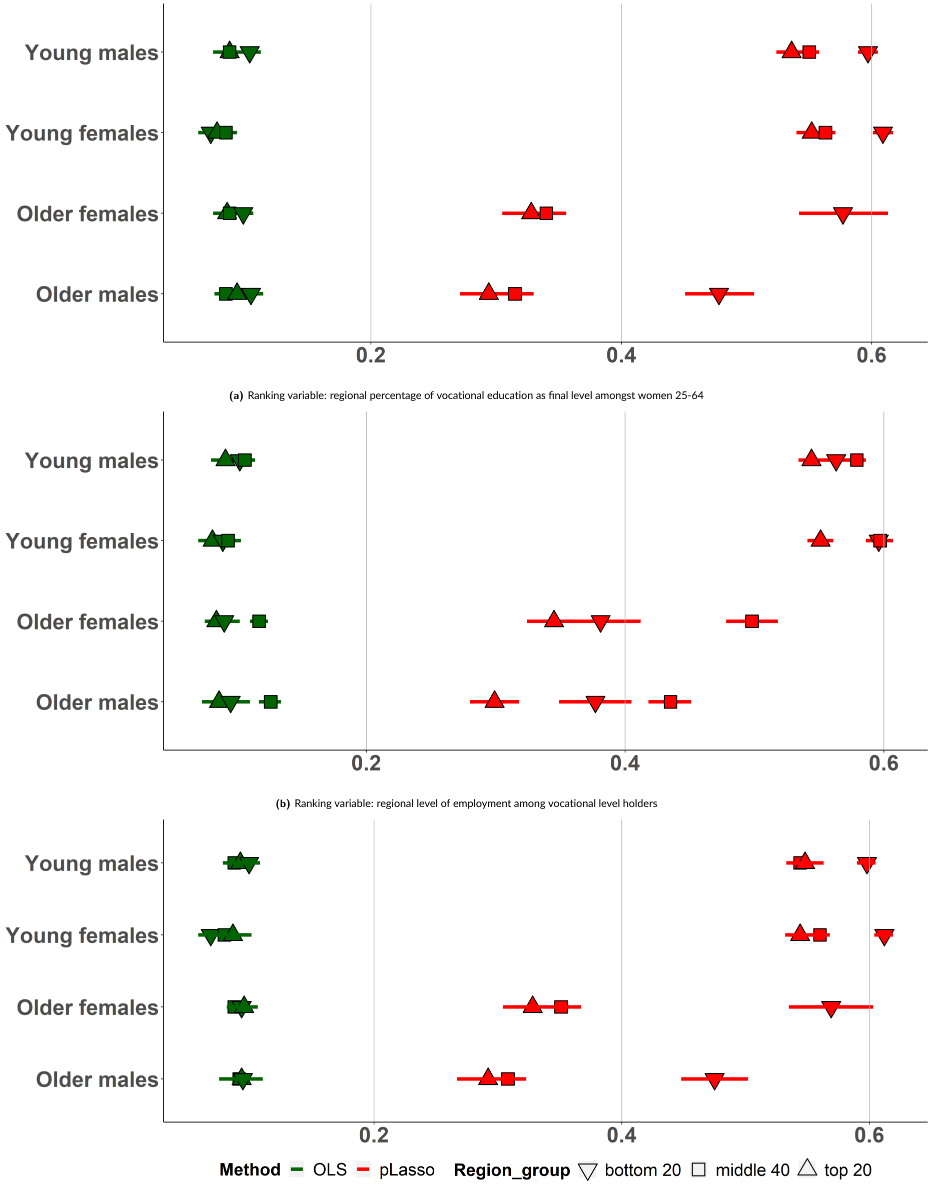


FIGURE 6.1 Returns to Education Estimates and 95% CIs for Post-Lasso, 2SLS, and OLS by Cohorts, 2018

TABLE 6.1 Means and Standard Deviations for Education Years and Wages by Ranked Groups of Regions, 2018

Variable	Group	Top 20	Middle 40	Bottom 20	Whole Sample
edu_yrs	Female older	13.38 (2.62)	13.91 (2.73)	13.91 (2.7)	13.76 (2.71)
	Female young	14.09 (2.75)	14.55 (2.66)	14.8 (2.61)	14.49 (2.68)
	Male older	12.51 (2.34)	13.21 (2.66)	13.26 (2.68)	13.03 (2.6)
	Male young	13.04 (2.58)	13.83 (2.74)	14.02 (2.75)	13.67 (2.73)
wage	Female older	19379.02 (11665.53)	27784.51 (18811.94)	26329.79 (16935.15)	25132.36 (17104.73)
	Female young	19731.68 (10639.46)	30433.91 (20443.96)	27890.92 (16672.82)	27170.08 (18149.23)
	Male older	26392.05 (18023.65)	36872.17 (26937.81)	35302.24 (24426.05)	33650.33 (24684.13)
	Male young	31539.2 (18744.09)	40202.54 (24260.45)	37103.45 (24277.62)	37266.07 (23242.51)
Priority regions		Pskovskaya Obl., Resp. Karelia, Resp. Mariy El	Altayskiy Kray, Kurganskaya Obl., Chuvashskaya Resp., Resp.a Altay	Resp. Adygeya, Resp. Kalmykiya, Resp. Tyva	Resp. Adygeya, Pskovskaya Obl., Altayskiy Kray, Kurganskaya Obl., Resp. Kalmykiya, Chuvashskaya Resp., Res. Altay, Resp. Karelia, Resp. Tyva, Resp. Mariy El

Note: the ranking was based on the regional level of employment among vocational level holders.



(a) Ranking variable: regional percentage of vocational education as final level amongst women 25-64

(b) Ranking variable: regional level of employment among vocational level holders

(c) Ranking variable: regional employment level of vocational education holders
multiplied with reciprocal of employment level of higher education holders

FIGURE 6.2 Returns to Education Estimates and 95% CIs for Post-Lasso and OLS by Cohorts and Ranked Groups of Regions, 2018

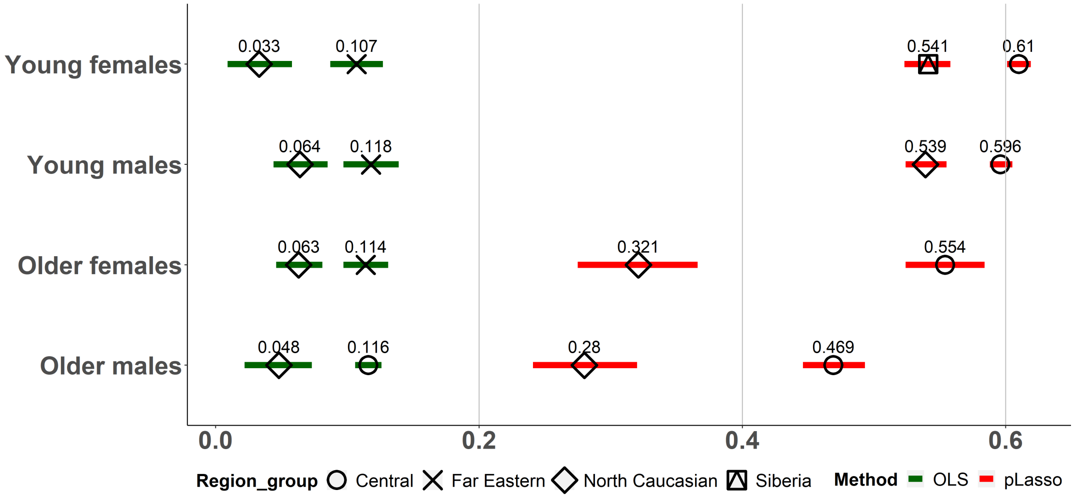


FIGURE 6.3 Returns to Education Estimates and 95% CIs for Post-Lasso and OLS by Cohorts and Federal Districts (Maximum and Minimum for each Method), 2018

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.

References

Akhmedjonov, Alisher. 2011. “Do higher levels of education raise earnings in post-reform Russia?” Publisher: Taylor & Francis, *Eastern European Economics* 49 (4): 47–60.

Arabsheibani, G. Reza, and Anita Staneva. 2012. “Returns to education in Russia: Where there is risky sexual behaviour there is also an instrument”. Publisher: IZA Discussion Paper.

Belloni, Alexandre, and Victor Chernozhukov. 2011. “High dimensional sparse econometric models: An introduction”. In *Inverse Problems and High-Dimensional Estimation*, 121–156. Springer.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. “High-dimensional methods and inference on structural and treatment effects”. *Journal of Economic Perspectives* 28 (2): 29–50.

Belloni, Alexandre, et al. 2012. “Sparse models and methods for optimal instruments with an application to eminent domain”. Publisher: Wiley Online Library, *Econometrica* 80 (6): 2369–2429.

Belskaya, Olga, Klara Sabirianova Peter, and Christian Posso. 2014. “College expansion and the marginal returns to education: Evidence from Russia”.

Cheidvasser, Sofia, and Hugo Benítez-Silva. 2007. “The educated Russian’s curse: Returns to education in the Russian Federation during the 1990s”. *Labour* 21 (1): 1–41.

Diebold, F. X. 2019. “Econometric Data Science: A Predictive Modeling Approach”. *Department of Economics, University of Pennsylvania*. `http://www.ssc.upenn.edu/~fdiebold/Textbooks.html`.

Kyui, Natalia. 2016. “Expansion of higher education, employment and wages: Evidence from the Russian Transition”. *Labour Economics* 39:68–87.

—. 2010. “Returns to Education and Education-Occupation Mismatch within a Transition Economy. Empirical Analysis for the Russian Federation”.

Tibshirani, Robert. 1996. “Regression shrinkage and selection via the lasso”. Publisher: Wiley Online Library, *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–288.

Appendix

References

Akhmedjonov, Alisher. 2011. "Do higher levels of education raise earnings in post-reform Russia?" Publisher: Taylor & Francis, *Eastern European Economics* 49 (4): 47–60.

Arabsheibani, G. Reza, and Anita Staneva. 2012. "Returns to education in Russia: Where there is risky sexual behaviour there is also an instrument". Publisher: IZA Discussion Paper.

Belloni, Alexandre, and Victor Chernozhukov. 2011. "High dimensional sparse econometric models: An introduction". In *Inverse Problems and High-Dimensional Estimation*, 121–156. Springer.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-dimensional methods and inference on structural and treatment effects". *Journal of Economic Perspectives* 28 (2): 29–50.

Belloni, Alexandre, et al. 2012. "Sparse models and methods for optimal instruments with an application to eminent domain". Publisher: Wiley Online Library, *Econometrica* 80 (6): 2369–2429.

Belskaya, Olga, Klara Sabirianova Peter, and Christian Posso. 2014. "College expansion and the marginal returns to education: Evidence from Russia".

Cheidvasser, Sofia, and Hugo Benítez-Silva. 2007. "The educated Russian's curse: Returns to education in the Russian Federation during the 1990s". *Labour* 21 (1): 1–41.

Diebold, F. X. 2019. "Econometric Data Science: A Predictive Modeling Approach". *Department of Economics, University of Pennsylvania*. `http://www.ssc.upenn.edu/~fdiebold/Textbooks.html`.

Kyui, Natalia. 2016. "Expansion of higher education, employment and wages: Evidence from the Russian Transition". *Labour Economics* 39:68–87.

— . 2010. "Returns to Education and Education-Occupation Mismatch within a Transition Economy. Empirical Analysis for the Russian Federation".

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso". Publisher: Wiley Online Library, *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–288.