Spatial and nonlinear effects of universities' research on innovations: Evidence from Russian regions

D. Tereshchenko, HSE University, St. Petersburg

1 Introduction

Recent work suggests that universities play an important role in regional innovation systems, but the nature of this role varies depending on the context. In addition, it is worth mentioning that research methods vary greatly from one paper to another. However, it is worth noting the lack of empirical studies of the impact of the research activities of universities for regional innovations. Also, the Russian context requires more detailed consideration. In addition, I consider the non-linear nature of the impact of university research on innovation.

2 Models and methods

2.1 Baseline model

Equation (1) presents a baseline research model that takes into account regional and yearly fixed effects, as well as the nonlinear relationship between university research and innovation:

$$y_{it} = p_{it}\beta_1 + p_{it}^2\beta_2 + \mathbf{C}_{it}^{\prime} \quad + \alpha_i + \nu_t + u_{it} \tag{1}$$

where i, t are region and year indices, y_{it} is a measure of innovation, p_{it} is a measure of university research, $\mathbf{C}_i$ is a vector of control variables, β_1, β_2 are parameters of interest, reflecting the impact of university research,   is a vector of coefficients on control variables, α_i, ν_t are region and year fixed effects, u_{it} is error term.

For the interpretation, I am interested primarily in analyzing the statistical significance of the marginal effect presented in the equation:

$$ME_p = \beta_1 + 2\beta_2 p$$

Note that the value and statistical significance of the marginal effect changes when the measure of university research activity (p) changes. The estimation of this marginal effect for different values of p can be obtained using estimates of coefficients β_1, β_2. One can also compute the standard errors of the marginal effect estimate for each p using the Delta method. In this way, one can estimate the statistical significance of the estimated marginal effect at each point.

2.2 Spatial effects model

The baseline model considers the average impact of university research activities and other variables on innovation within each region. However, in reality, there may be complex spatial relationships between regions that need to be taken into account in the model. One way to address this problem is adding the spatial lags of the explanatory variables to the model. Such model is represented in the equation (2):

$$y_{it} = p_{it}\beta_1 + p_{it}^2\beta_2 + \bar{p}_{it}\beta_3 + \bar{p}_{it}^2\beta_4 + \mathbf{C}_{it}^{'} \; + \overline{\mathbf{C}}_{it}^{'} \; + \alpha_i + \nu_t + u_{it} \tag{2}$$

where $\bar{p}_{it}$ is a "spatial lag", or spatially weighted average of p_{it} within regions neighboring to the region i, while $\overline{\mathbf{C}}_{it}$ is a vector of such averages for control variables.

The model presented in the equation (2) resembles so-called SLX model with a small correction. It is usually assumed that the spatial lags of all explanatory variables are included in the right-hand side of the equation. However, in my model it is not quite so, because instead of the spatial lag of the square of university research activity I include *the square of the spatial lag of university research activity*, which makes more sense.

In this setting, the coefficients β_1, β_2 reflect the direct impacts of the variables p_{it}, p_{it}^2, while β_3, β_4 refer to the indirect impacts of those variables. Then one can derive a formulas for the direct and indirect marginal effects:

$$ME_p^{direct} = \beta_1 + 2\beta_2 p$$

$$ME_p^{indirect} = \beta_3 + 2\beta_4 p$$

2.3 Dynamic model

Another way to enrich the model is to add dynamic elements to it. First of all, it applies to the dependent variable, because innovations are likely to be dynamic, i.e. innovations in a given period of time depend on their previous values. On the other hand, some of the explanatory variables may have a lagged effect on innovation. All this can be taken into account by adding time lags of the dependent and explanatory variables to the model. Meanwhile, adding a lag of the dependent variable creates endogeneity in the model, which causes traditional methods to lead to inconsistent estimates.

By combining spatial and dynamic effects, one can start by introducing a model in the form presented in the formula (3):

$$y_{it} = p_{it}\beta_1 + p_{it}^2\beta_2 + \bar{p}_{it}\beta_3 + \bar{p}_{it}^2\beta_4 + \mathbf{C}_{it}^{'} \; + \overline{\mathbf{C}}_{it}^{'} \; + y_{i,t-1}\rho + \alpha_i + \nu_t + u_{it} \tag{3}$$

where $y_{i,t-1}$ is time lag of dependent variable. Without serious complications, it is possible to add to the model even more time lags of denedent variable, i. e. $y_{i,t-2}$, as well as vectors $\mathbf{C}_{it}^{'}$, $\overline{\mathbf{C}}_{it}^{'}$ can contain time lags of control variables and their spatial lags respectively.

Four types of impacts can be calculated for statistical inference: direct long-term, indirect long-term, direct short-term and indirect short-term. Since the marginal effect I am interested in depends on two variables, p_{it} and p_{it}^2, four impact types should be estimated for each of them. As a result, estimates of long-term and short-term, direct and indirect marginal effects can be obtained for each point of interest, as well as their standard errors.

2.4 Estimation strategy

The estimation strategy in the study aims to deal with endogeneity, which can be caused either by the presence of time lags of the dependent variable in the model, as well as omitted variables or reverse causality. However, inclusion of spatial lags of explanatory variables in the model, as well as their time lags, does not lead to the endogeneity problem, so it does not significantly affect the estimation procedure.

Therefore, for illustrative purposes, I estimate the baseline model and the "SLX" model using traditional methods, i.e., OLS with within-group transformation. In doing so, I am cautious about the interpretation of the results obtained.

In contrast, the inclusion of a time lag of the dependent variable makes OLS estimators inconsistent, so I use the Arellano-Bond (difference GMM) and Blundell-Bond (systematic GMM) methods to estimate such models. These methods also help to deal with endogeneity caused by omitted variables and reverse causality. The main way to justify the robustness of the results is to align the findings obtained using a different number of instruments used in GMM estimations.

It should be noted that the definition of innovation in Russian statistics changed in 2017. I account for this by including year fixed effects in the model, and I also estimate models on a panel of Russian regions for both the full period 2011-2019 and a reduced panel for 2011-2016.

3 Data and variables

The information base of the study is balanced panel data on Russian regions for 2011-2019. The main source of data was the annual collection of Rosstat "Regions of Russia", from which data on regional innovation activity and socio-demographic characteristics of the regions were extracted. Universities' research was measured based on data on the publication activity of the OpenAlex project (Priem, Piwowar, and Orr, n.d.). All monetary variables are adjusted to take into account the cost of the regional consumer basket.

To measure the dependent variable, i.e. innovations, I use the *share of innovative goods, works, services* in their total volume as the dependent variable.

As a proxy for university research, I use OpenAlex data on the number of publications and citations. In doing so, I transform the variables and use such relative indicators as *works count per faculty member* and *citation count per work*.

As control variables, I use the share of R&D expenditures in GRP, the number of students per 10,000 population, the share of R&D organizations in the total number of organizations, the share of middle-aged population, the share of older population, and the number of women per 1,000 men.

4 Results

Preliminary results are not entirely unambiguous. In terms of direct effects, the static models partially confirm the presence of an inverse U-shaped dependence of innovation on the number of publications per author in the region. Dynamic models, in turn, suggest the presence of a U-shaped dependence of innovation on the number of publications per author in the region, i.e. a parabola with upward branches. As for indirect effects, they turned out to be insignificant in all models.

References

Priem, Jason, Heather Piwowar, and Richard Orr. n.d. "OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts." https://doi.org/10.48550/arXiv.2205.01833.