Study of the Influence of Indicator Factors on a Complex Index of the Region's Innovation Potential Using Correlation and Regression Analysis

A S Sazonova¹, A A Kuzenko¹, L B Filippova¹, R A Filippov¹

¹Bryansk State Technical University, 241035, Central Federal district, Bryansk region, Bryansk city, Bouleva 50 let Oktyabrya, 7

E-mail: asazonova@list.ru

Abstract. A special feature of Russia's innovative development is the high degree of regional development disparity. This is due to a number of factors: specific nature of each region, historically established leading industries and geographical location. The innovative potential of a country is a combination of innovative potential of regions, and the importance of considering and studying the issues of evaluating and analyzing this index is more relevant today like never before.

The innovation potential of a region is a whole complex of resources and organizational capabilities of the region aimed at providing its innovative development, ensuring an innovation continuum. In this review, the innovation potential is represented as an integrated quantity that includes a system of indicator factors.

Within the framework of this studies, the analysis of the influence of individual indicator factors on the complex index of the region's innovation potential is of particular interest. Since the relationship between indicators is incomplete and often probabilistic, there are good reasons for using correlation and regression analysis as a research method.

1. Introduction

Correlation and regression analysis is a statistical analytical method that makes the calculation of the assumed relationships between a dependent variable and one or more independent variables possible. The purpose of this analysis is to establish a specific analytical dependence of one or more performance indicators on one or more indicator factors [1].

In statistics, correlation and regression are usually considered as a combined process of statistical research showing that the change in an apparent variable depends on a large number of factors, but only a few of them have a significant impact on the studied index. The value of the region's innovation potential was chosen as an apparent variable, and the values of indicator factors were chosen as explanatory variables.

Initially, it is necessary to estimate the value of the innovation potential and obtain a figure of this value. As mentioned above, innovation potential is a complex indicator that includes a set of individual factors.

The following factors were selected as agent-indicators:

- number of staff engaged in research and development, number of people (x1);
- internal expenditures on research and development, billion rubles (x2);
number of off-the-shelf advanced manufacturing technologies, units (x3);
number of existing advanced manufacturing technologies, units (x4);
volume of innovative goods, works, services in bulk of shipped goods, performed tasks, percent (x5);
innovative activity of organizations (the share of organizations that carried out technological, organizational, and marketing innovations in the total number of organizations surveyed), percent (x6);
coefficient of inventive activity (the number of domestic patent applications for inventions filed in Russia, per 10 thousand people of the population) (x7).

The selection of these factors was made on the basis of expert views on the complex indicator of innovation potential, as well as based on the possibility of collecting statistical data.

2. Theoretical part
A mathematical model of the region's innovation potential can be presented as follows (1):

\[ IP^r = \langle x_1, x_2, x_3, x_4, x_5, x_6, x_7 \rangle \]  

where \( IP^r \) – a composite index, \( x_1\text{–}x_7 \) – indicator factors.

The calculation of the innovation potential indicator was performed for the Bryansk region, the study period is 2010–2018 [2,3,4,5,11].

To perform calculations for all the listed factors, statistical information was collected for this study period.

Due to the fact that the variables included in the final model differ both in their physical meaning and in their absolute values, it is advisable to perform a procedure for normalizing the values of the indicator factors using the formula (2):

\[ n_{ir} = \frac{x_{ir} - x_{imin}}{x_{imax} - x_{imin}} \]  

where \( n_{ir} \) – normalized value of \( i\)-th indicator factor concerning \( r\)-region in the considered extensive evidence, \( 0 \leq n_{ir} \leq 1 \), \( x_{ir} \) – absolute or relative (actual) value of the \( i\)-th factor for each region in the data set under consideration. \( x_{imin} \) and \( x_{imax} \) – the minimum and maximum values of the \( i\)-th factor in the data set under consideration. The integrated indicator of the region's innovation potential was calculated using the area diagram method [6]. The essence of the method is that the complex value is defined as the area of a flat shape, formed by rays, where length is determined by the values of indicator factors (Fig.1).
Graphical representation of the area diagram is necessary exclusively for the clarity of the method, all calculations were performed with the benefit of the analytical method using the formula (3):

$$S_t = \sum_{i=1}^{n} k_{i,t} \cdot k_{i,t+1} \cdot \sin(360/i) / 2$$  \hspace{1cm} (3)

where $S_t$ is the area of the constructed diagram for the $t$-th year, $k_{i,t}$ is the normalized value of the $i$-th indicator factor for the $t$-th year in the data set under consideration, and $i$ is the number of indicator factors selected for modeling. Thus, the values of the innovative potential of the Bryansk region in the period under review were calculated. The final data is presented in table 1.

Table 1. The values of the innovation potential of the Bryansk region for 2010-2018.

| Years | The importance of the innovative potential of the Bryansk region |
|-------|---------------------------------------------------------------|
| 2010  | 0.042959746                                                   |
| 2011  | 0.270575563                                                   |
| 2012  | 0.358414629                                                   |
| 2013  | 0.061178042                                                   |
| 2014  | 0.375530036                                                   |
| 2015  | 0.576496818                                                   |
| 2016  | 1.227948656                                                   |
| 2017  | 1.184218176                                                   |
| 2018  | 0.818450327                                                   |

3. Practical significance and the results of implementation

Before creating a regression model it is inevitable to test the linearity of the original model for determining which independent variables should be used. To analyze the linear relationship between the dependent variable and independent ones, scatters (dispersion) were plotted. While studying the
obtained relationships, it was found that all the dependencies are linear in nature, so we can use the line function equation of the form (4):

\[ y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \] \hspace{1cm} (4)

where \( y_i \) – experimental value of an independent variable at point \( i \), \( \beta_0 \) – y-intercept (segment on the coordinate axis, constant value), \( \beta_n \) – coefficient of regression or deviation of the independent variable \( N \) at point \( i \), \( x_n \) – value of the variable \( N \) at point \( i \), \( \varepsilon \) – error in the regression equation.

At the next stage of research, it is necessary to construct a matrix of correlation coefficients to identify the tightness of relationships between the studied independent variables. The matrix of correlation coefficients is presented in table 2.

| \( x_1 \) | \( x_2 \) | \( x_3 \) | \( x_4 \) | \( x_5 \) | \( x_6 \) | \( x_7 \) | \( Y \) |
|---|---|---|---|---|---|---|---|
| 1 | -0.10 | 0.22 | -0.15 | 0.23 | 0.09 | 0.51 | 0.41 |
| 0.91 | 0.98 | 0.32 | 0.14 | -0.97 | 0.60 | 0.74 | 0.74 |
| 1 | 1 | 0.87 | 0.32 | -0.89 | 0.76 | 0.88 | 0.88 |
| 0.24 | -0.94 | 0.57 | 0.56 | -0.23 | 0.56 | 0.71 | 0.71 |
| 1 | -0.23 | -0.65 | 0.61 | -0.77 | 0.94 | 0.61 | 0.61 |

By reference to this table it can be seen that such indicator factor as "number of employees engaged in research and development" can be excluded from the model, since it has a weak impact on the indicator of the region's innovation potential. The constructed regression equation has the following form (5):

\[ Y = 0.00035 x_2 + 0.04385 x_3 - 0.00036 x_4 + 0.00001 x_5 - 0.00583 x_6 + 0.72148 x_7 - 0.68452 \] \hspace{1cm} (5)

where \( Y \) – the value of innovative potential of Bryansk region, \( x_2 \) – internal research and development costs, \( x_3 \) – number of advanced developed manufacturing technologies, \( x_4 \) – number of used advanced manufacturing technologies, \( x_5 \) – volume of innovative goods, works, and services in the bulk of shipped goods and performed tasks, \( x_6 \) – innovative activity of organizations, \( x_7 \) – coefficient of inventive activity. The regression indicators are calculated in table 3.

| Name of the indicator | Value of indicator |
|-----------------------|--------------------|
| Multiple R            | 0.9913152          |
| R-square              | 0.9827058          |
| Normalized R-square   | 0.9308231          |
| Standard error        | 0.1167289          |
| Observations          | 9                  |

The determination coefficient denoted as R-square, measures how well the regression equation models actual data points. The resulting model has R-squared value of 0.982, which is a good indicator of the model's accuracy.

The standard error of the discrepancy measures the accuracy that makes regression model prediction of values with new data possible. In the resulting model, it has a value of 0.116, which also
indicates the high accuracy of the constructed model. For determining the significance as a component of the analysis, the following statistical parameters are used:

- F-statistics and associated p-value;
- T-statistics and associated p-value;
- confidence belts;

F-statistics - a global statistical indicator returned by the F-criterion that shows the model's prediction capabilities by calculating regression coefficients in the model that differ significantly from 0. F-criterion analyzes the combined effect of independent variables, rather than evaluating each one individually. The significance of the F-criterion in the resulting model is 0.05, which indicates that the model actually has relationships between variables (that is, the revealed regularity is not random). As for t-statistic indicator it can be seen that this local statistical indicator returned by the t-criterion, shows the prediction capabilities for each independent variable separately. Each t-statistic value has an associated p-value that indicates the significance of the independent variable. The results are shown in table 4.

**Table 4. t-statistics and p-values for independent variables in the resulting regression model.**

| Name of the indicator | t-statistics | P-values  |
|-----------------------|--------------|-----------|
| Y-crossing            | -0.43972     | 0.703091  |
| Variable X 2          | 0.234574     | 0.836367  |
| Variable X 3          | 1.165876     | 0.363893  |
| Variable X 4          | -0.30129     | 0.791634  |
| Variable X 5          | 1.140320     | 0.372306  |
| Variable X 6          | -0.03409     | 0.975903  |
| Variable X 7          | 2.189627     | 0.159975  |

Confidence interval is a zone where the true regression line lies with a 95% probability for repeated samples from this population. Confidence intervals for independent variables are shown in table 5.

**Table 5. Confidence intervals.**

| Name of the indicator | Coefficients | Low 95% | Upper 95% |
|-----------------------|--------------|---------|-----------|
| Y-crossing            | -0.684520496 | -7.38252 | 6.013477  |
| Variable X 2          | 0.000353673  | -0.00613 | 0.006841  |
| Variable X 3          | 0.043857893  | -0.118  | 0.205715  |
| Variable X 4          | -0.000365582 | -0.00559 | 0.004855  |
| Variable X 5          | 0.000012557  | -0.00035 | 0.000060  |
| Variable X 6          | -0.005836974 | -0.7426 | 0.730928  |
| Variable X 7          | 0.721483154  | -0.69624 | 2.139209  |

For some variables the confidence interval is quite large, but in this particular case (in this particular aggregate), if there is a small sample of data, this assumption was accepted. In consequence of this analysis the resulting model is recognized as suitable for research and has a fairly high accuracy. The obtained results are of practical value, as the given model allows calculating the predicted values of innovation potential depending on the values of the indicator factors. Prediction problem is one of the most important tasks of managing innovation and innovative development, since it enables adjusting the policy in this industry making an effective distribution of resources in the region possible.
4. References

[1] Glinsky V V, Ionin V G 2002 Statistical analysis: textbook (Moscow: INFRA-M publ.)

[2] The Federal statistics service Official statistics Available at: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/science_and_innovations

[3] Science and innovation in Russian regions Available at: http://regions.extech.ru

[4] The Regions of Russia Main characteristics of the subjects of the Russian Federation Available at: http://www.gks.ru

[5] Rospatent Russian Federal Service for Intellectual Property Available at: https://rupto.ru/

[6] Kozhukhar V M 2010 Monitoring and forecasting of regional needs for specialists of higher scientific qualification: Bryansk state technical University (Bryansk)

[7] Filippova 2017 In: Abraham A, Kovalev S, Tarassov V, Snasel V, Vasileva M, Sukhanov A (eds) Text : electronic Proceedings of the Second International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’17) IITI 2017 Advances in Intelligent Systems and Computing vol 679 (Springer, Cham) pp 95-103 DOI:10.1007/978-3-319-68321-8_10

[8] Leonov Yu A 2019 Selection of rational schemes automation based on working synthesis instruments for technological processes (Yelm, WA, USA: Science Book Publishing House LLC) 192 p ISBN: 978-5-9765-4023-1 Text : unmediated

[9] Averchenkov A V 2018 Development of a mathematical model of an information system for inventory and monitoring of software and hardware based on fuzzy logic methods Text: direct Quality. Innovations. Education 7 (Moscow: European center for quality) pp 105-112 ISSN: 1999-513X

[10] Kuzmenko A A 2018 Development of the WEB resource structure based on the needs of the end user Text: direct In the collection: New information technologies in scientific research Materials of the XXIII all-Russian scientific and technical conference of students, young scientists and specialists Vol 2 (Ryazan: Ryazan state radio engineering University) pp 183-185 ISBN: 978-5-7722-0274-6

[11] Absimetov V E, Saltanova E V, Solovev D B 2020 Possibility Justification of Application of the Mecanochemical Method during Preparation of the Metal Structures’ Surfaces under Repair and Reconstruction Materials Science Forum 992 585-589. [Online]. Available: https://doi.org/10.4028/www.scientific.net/MSF.992.585