Regression Analysis as a Tool for Assessing the Causal Relationships Between Urban Form and Socioeconomic Indicators of Urban Life

Ph.D. Eng. Arch.
MONIKA MARIA CYSEK-PAWLAK, PROF. UNIVERSITY
Lodz University of Technology
Institute of Architecture and Urban Planning
ORCID: 0000-0002-8175-6779

Ph. D. Eng. Arch.
SYLWIA KRZYSZTOFIK
Lodz University of Technology
Institute of Architecture
and Urban Planning
ORCID: 0000-0002-9053-0042

MSc. Eng. Arch.
ANDRII POLISHCHUK
Lodz University of Technology
Institute of Architecture
and Urban Planning
ORCID: 0009-0008-8378-8863

The purpose of this article is to depict the application of regression analysis for assessing the causal relationships between urban form and socioeconomic indicators of urban life.

Urban studies is a vast scientific field encompassing disciplines such as urban planning and the socioeconomic assessment of urban life. Jane Jacobs depicts the city as an "organized complexity" [1], where urban form both depends on and shapes social life simultaneously. This assumption allows us to conceptualize the relationships between the urban environment and urban life as a mathematical function, where the value of one variable depends on the value of another. While urban life is the domain of social sciences, the urban environment is shaped by urban development activities governed by local authorities and based on urban planning, which falls within the purview of architects and urban planners. Designers are the first instance to designate the values of specific urban form metrics as an outcome of their design. Hence, a reasonable question is whether those metrics are target-oriented. The experience of modernistic experiments in urban development has shown that urban form, which seemed perfect from an architectural point of view, does not necessarily appear to be viable and vibrant.

On the other hand, social sciences, such as economics, widely employ quantitative research to detect, predict, and adjust the power of causal relationships between resource management and its impact on quality of life, well-being, and public health. The major technique used for such investigations is regression analysis, which allows distinguishing deterministic data relationships from random ones. For example, the following causal relationships are observed and depicted within the domain of socioeconomic research: the dependency between average human height and relative prices [2], body mass index and income [3], poverty and crime [4].

Another aspect of the application of quantitative research is the issue of research bias. The unbiasedness of qualitative research is ensured by applying intersubjective research methods and techniques. Quantitative research, on the other hand, employs probability theory to estimate so-called confidence intervals, which allow calculating the probability of false-positive conclusions (Type I error). For example, if the confidence interval for rejecting a particular statistical hypothesis is less than 5%, then in 19 out of 20 cases, the rejection of this hypothesis is correct.

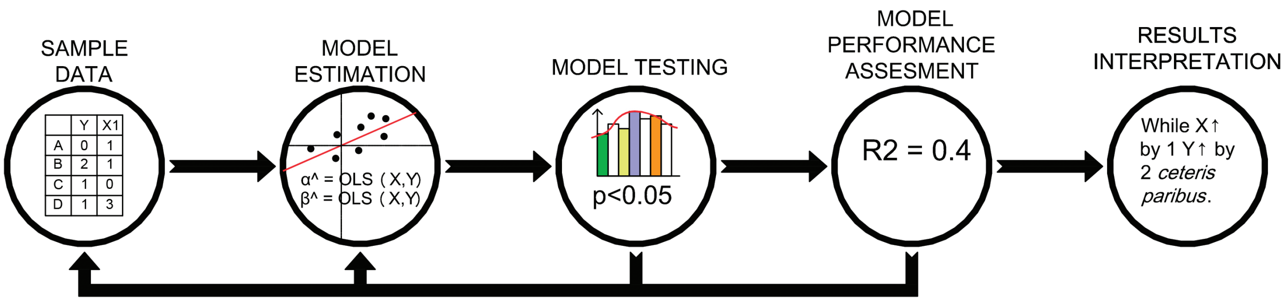

Fig. 1. Regression analysis algorithm; source: own elaboration

Statement of the article's purpose and research question

Despite the important role of quantitative methods, it is worth noting that social research could not be possible without qualitative expertise. The verification of scientific theories requires hypothesis formulation and interpretation based on previous knowledge, research intuition, and other qualitative factors. However, quantitative methods allow depicting these relationships in the form of mathematical equations, which create possibilities for interpreting the strength of the relationships and predicting future behavior. The purpose of this article is to verify whether the application of regression analysis could be a useful tool that supplements quantitative research techniques in assessing the causal effect relationships between urban form and socioeconomic indicators of urban life.

Methodology

For the purposes of the research, a case study of the tool's usefulness [5] is employed. The methodology is divided into two steps, which consider the depiction of regression analysis and the review of conducted studies that apply statistical inference to assess the causal relationships between the urban environment and socioeconomic indexes. In the first step, a comprehensive depiction of the step-by-step algorithm of regression analysis will be provided. The focus will be on the possibilities of employing the tool's features for assessing urban regeneration policy outcomes and interpreting the statistical modeling results. The second step will include a review of selected studies that apply regression analysis to evaluate relationships between the urban environment and socioeconomic indexes. This part will focus on interpreting the results of regression modeling.

In the article's conclusion, the possibilities and limitations of applying regression analysis in urban planning will be summarized. The authors will address the importance of applying numerical methods in planning research and the necessity to align with qualitative expertise.

The depiction of the regression analysis methodology

Regression analysis can be depicted as an iterative algorithm consisting of five steps (fig. 1.). Mathematically, a regression model can be represented in the form of an equation [6]:

$Y = \alpha+\beta X+\varepsilon$ (1)

where:

Y – dependent variable,

X – explanatory variable (s),

α – intercept,

β – slope coefficient(s),

ε – random component

The Latin letters in equation (1) correspond to the data that researchers should collect before running regression analysis. Greek letters (α and β), on the other hand, are called estimates. These parameters could be calculated by inserting the collected data into the formula of estimators. The last component of the regression model is ε – the random component. After the calculating of the estimates (α and β) and having the data, the random component can be calculated as the difference between the dependent variable and other known components of the equation ($\varepsilon = Y - \alpha - \beta X$). It is worth noting that equation (1) without random component ε is a regular equation that defines a linear function and by adding the ε, the deviation from the line is specified for each data point. Consequently the addition of the random component transforms the linear function into the regression model.

At the first step of the regression analysis algorithm, the researcher prepares and analyzes the data and states the research hypothesis. The hypothesis concerns confirming the functional dependency between the dependent variable and explanatory variable(s). In this phase, graphical analysis is employed to assume whether this dependency tends to be linear or nonlinear (fig. 2.). This study concerns only linear models estimated using the Least Squares family of estimators [7], which are the most popular regression analysis techniques. Nonlinear models estimation is beyond the scope of this elaboration.

The second step in the regression analysis algorithm concerns estimating the structural parameters of the regression equation. For this purpose, least squares estimators are employed. The ordinary least squares (OLS) estimator formula is presented below:

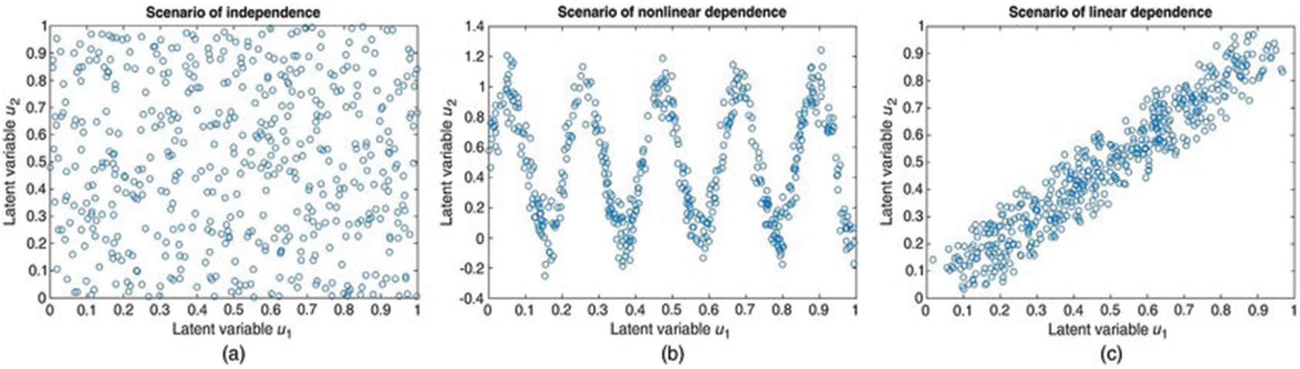

Fig. 2. An example of non-linear(b) and linear (c) dependency; source: [8]

$$\hat{\beta} = (X^T X)^{-1} X^T y \qquad (2)$$

where:

β – slope coefficient(s)

Y – dependent variable,

X – explanatory variable (s),

After the calculating the slope coefficient ($\hat{\beta}$) the intercept coefficient is calculated as the difference between the mean value of the dependent variable in empirical data and the slope coefficient times the mean value of the independent variable ($\hat{a} = \bar{y} - \hat{\beta}\bar{x}$). After the calculation of residuals for each data point, the model is constructed.

This allows moving to the third stage of regression analysis – model testing. This includes both testing the statistical significance of the slope coefficient (β) by applying t-tests and testing the statistical properties of residuals (e.g., normality, heteroskedasticity, and autocorrelation tests). Each of these tests is designed as a statistical test with null and alternative hypotheses. After calculating the test statistic, which is a combination of the regression model components, the researcher checks if the test statistic value falls within the confidence interval (fig. 3.) for the specific test. If the test statistic fits within the confidence interval, the results confirm the statistical hypothesis of the test. The most important test from the perspective of applied analysis is the t-test of the slope coefficient. The null hypothesis of this test is that the slope coefficient (β) equals 0, and the alternative hypothesis is that it does not equal 0. If the researcher calculates the test statistic at a given confidence level and finds that it falls outside the confidence interval for the t-test, it is stated that the null hypothesis is rejected. Consequently, the slope coefficient does not equal 0, and the functional relationship is confirmed.

The linear regression model, estimated via OLS, includes the calculation of the coefficient of determination called R-squared [9]. The value of the coefficient spans from 0 to 1, where 1 means 100% explanation of empirical data by the model, and 0 means no explanation of empirical processes by the model. The third and fourth steps of the linear regression model analysis are designed for assessing the model's goodness of fit to empirical data. Naturally, if the model does not fit the data, or some problems with the model specification are discovered as a result of statistical tests, the researcher returns to the previous phases of modeling to improve the model by choosing a different estimator or different data.

The final step of regression modeling involves interpreting the results [10]. Applied regression analysis is mostly concerned with interpreting the slope coefficient of the regressors (explanatory variables) and the determination coefficient. Moving from general to specific, the researcher addresses the issue of the goodness of fit of the model by interpreting the R-squared coefficient. It is widely considered that an R-squared coefficient greater than 0.6 indicates a sufficient level of explanation of the determination part of the model. The next step is interpreting the slope coefficients, which in models estimated via Least Squares Estimators represent the ratio of change in the dependent variable Y in response to an increase of one unit in the regressors, ceteris paribus (with other components of the equation held constant). This allows measuring the average influence of one factor on another and estimating the strength of this influence. It is worth noting, that the regression analysis is an iterative process, that involves repetitions of the aforementioned steps in order to receive the best model specification and complete theoretical explanation of the random process, that is depicted with a data.

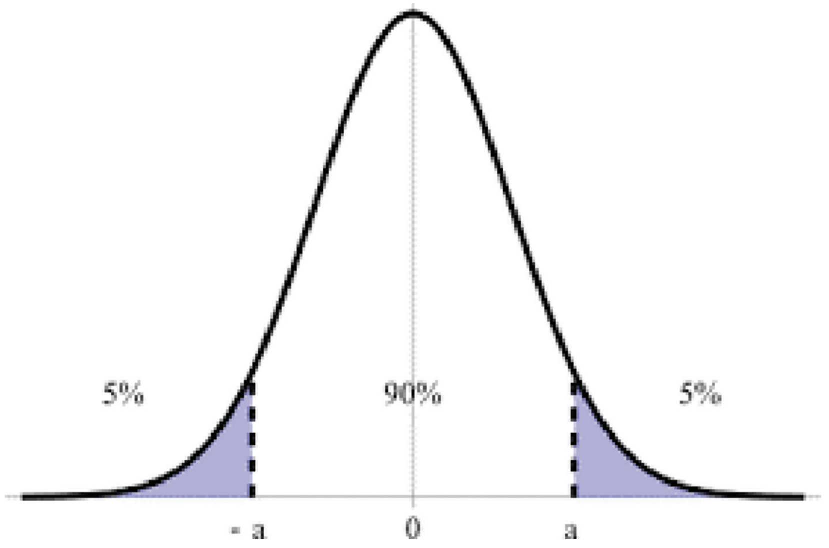

Fig. 3. Confidence interval; Source: https://www.coursesidekick.com/statistics/study-guides/boundless-statistics/confidence-intervals; Access on 30.07.2024

Sample empirical studies

One of the most popular issues related to the impact of urban form on economic variables is the aspect of the real estate market. Wang et al. [11] estimate a regression model in which the dependent variable (housing prices) is explained by mixed land use, public service level, high-rise buildings, and air pollution. The model results suggest that mixed land use, high-rise buildings, and air pollution decrease the value of properties, while proximity to public service buildings increases it (all variables are statistically significant at least at the 5% confidence level). The results of the study are somewhat contradictory to the common understanding of sustainable urban form development, which assumes a shift from typically multifamily districts to multifunctional urban areas, as inhabitants are ready to pay more for apartments in areas with less functional variation. Nevertheless, it provides valuable evidence of real market response forces to urban development processes and allows local authorities to adjust their policies in the future.

Research by De Vries et al. [12] examines the impact of the quality and quantity of urban greenery on the perceived general health of city inhabitants. Guided by intuition based on mounting evidence in this relationship, the researchers construct a regression model in which perceived general health level (measured from 1 to 5) is explained by both the quality and quantity of urban greenery as well as other health indicators such as stress level, social cohesion, and green activity. The empirical results provide evidence that health levels depend on the quality of urban greenery at the 5% confidence level. According to the model, better quality of urban greenery raises the likelihood of inhabitants perceiving themselves in better health by 9%, ceteris paribus. In this way, the authors provide confirmation for the qualitative hypothesis and support current trends in urban planning, offering quantifiable evidence of the dependency.

Research conducted [13] for Beijing addresses the issue of renovating urban villages, parts of the Chinese capital developed chaotically during rapid urbanization and now being redeveloped due to the City Strategic Development Act. The study employs regression analysis to assess the effects of urban village renewal on nearby housing prices. The authors create dummy variables indicating different proximities of housing from renewed urban villages. The results indicate a statistically significant impact of proximity to urban villages on housing prices. Housing prices within 1km of renewed urban villages decrease by 4.8% with a 1% confidence interval, and prices of houses 1-2km away decrease by an average of 2.9% (5% confidence interval). The authors state that such a decrease in housing prices near urban villages could lead to the gentrification of the urban villages themselves.

Another study [14] considers the impact of neighborhood renewal (NR) programs in Northern Ireland on household fuel poverty (an index that detects whether a household spends more than 10% of its income on heating). The authors refer to both NR interventions, such as fostering entrepreneurship and education of NR inhabitants, which should potentially increase inhabitants' incomes, and the improvement of buildings in NR areas, such as insulation and heating system modernization. The applied difference-in-differences estimates reveal a statistically significant (95% confidence interval) reduction of 3% in the likelihood of respondents in program areas being in fuel poverty compared to the rest of Northern Ireland.

A study [15] conducted for Glasgow TRA (Transformational Regeneration Areas) assesses the impact of multifamily housing regeneration on crime rates in Glasgow. Crime data was used as a dependent variable explained by two independent variables: proximity rings to TRAs and a year dummy to explain the change over time. The study results indicate a statistically significant decrease in crime rates in the closest distance rings to renewed TRAs over a 14-year period.

Conclusion

In recent years, more statistical data has been gathered at a level of detail appropriate for urban planning analysis (e.g., Open Street Map, local authority web sources). This opens new possibilities in urban planning research. Quantitative assessment of urban form relationships with socioeconomic variables creates the possibility to adjust urban development policies based on urban planning as a discipline that plans and designates specific metrics of urban form.

It is worth noting that disciplines such as economics developed quantitative reasoning to support theories in the 20th century. The employment of probability theory and regression analysis allowed the rapid development of quantitative economic theory in the late 20th century.

Moreover, the studies referenced in this article prove that it is possible to apply regression analysis to assess the impact of urban form on social life indicators in urban areas. It should also be noted that each of these studies used qualitative domain expertise to formulate hypotheses and interpret results. This supports the hypothesis of this elaboration that in planning research, as well as in social sciences, the employment of quantitative research should be strictly complemented by qualitative research.

There are several limitations to applying regression analysis in planning research. The first limitation concerns the pioneering nature of this methodology in the urban planning domain. Due to the lack of sufficient data, the research conducted so far has not significantly contributed to a paradigm shift in planning research, which has been dominated by qualitative methods. The second limitation arises from the statistical properties of regression estimators, which are based on the Law of Large Numbers. Consequently, the dataset used for regression analysis should include a sufficiently large sample size, which is not always feasible.

Nevertheless, statistical inference is a very broad and continuously developing field. Regression analysis encompasses not only linear dependencies but also non-linear models that are built on the assumption of a non-linear trend function. Additionally, temporal and spatial models address the issues of time lags and spatial spillovers of random processes. Furthermore, the econometric concept of endogeneity assumes the existence of factors that influence both the cause and effect and allows us to discover and analyze those effects. For these reasons, regression modeling, despite the challenges related to data availability, is a flexible tool that can significantly enhance scientific analysis in the domain of urban planning.

BIBLIOGRAPHY

[1] Timothy Malcolm Baynes, 2009, Complexity in urban development and management: Historical overview and opportunities, Journal of Industrial Ecology, Doi: 10.1111/j.1530-9290.2009.00123.x.

[2] Marıa-Dolores Ramón, Jose Miguel Martínez-Carrión, 2011, The relationship between height and economic development in Spain, 1850–1958, Economics & Human Biology 9.1, Doi:10.1016/j.ehb.2010.07.001.

[3] Dean Jolliffe, 2011, Overweight and poor? On the relationship between income and the body mass index, Economics & Human Biology 9.4,doi:10.1016/j.ehb.2011.07.004.

[4] Halvor Mehlum, Edward Miguel, Ragnar Torvik, 2006, Poverty and crime in 19th century Germany, Journal of Urban Economics 59.3, doi:10.1016/j.jue.2005.09.007.

[5] Barbara Kitchenham, Lesley Pickard, Shari Lawrence Pfleeger, 1995, Case studies for method and tool evaluation, IEEE software 12.4, pp. 52–62.

[6] Niemiro W., Statystyka I, Wydział Matematyki i Informatyki, Uniwersytet Mikołaja Kopernika, Toruń.

[7] Strejc V., 1981, Least squares and regression methods, Trends and progress in system identification, Pergamon, doi:10.1016/B978-0-08-025683-2.50009-0.

[8] Wenxing Hu, Aiying Zhang, Biao Cai, Vince Calhoun, Yu-Ping Wang, 2019, Distance canonical correlation analysis with application to an imaging-genetic study, Journal of Medical Imaging 6.2 doi: 10.1117/1.JMI.6.2.026501.

[9] Rawlings, J. O., Pantula S.G., Dickey D.A., Applied regression analysis: a research tool, Springer New York, New York.

[10] Silva, E. A., Healey, P., Harris, N., Van den Broeck, P., The Routledge handbook of planning research methods, Routledge.

[11] Yang Wang, Kangmin Wu, Yabo Zhao, Changjian Wang, Hong'ou Zhang, 2022, Examining the effects of the built environment on housing rents in the Pearl River Delta of China, Applied Spatial Analysis and Policy, Doi: 10.1007/s12061-021-09412-4.

[12] Sjerp de Vries, Sonja van Dillen, Peter Groenewegen, Peter Spreeuwenberg, 2013, Streetscape greenery and health: Stress, social cohesion and physical activity as mediators, Social science & medicine 94, Doi: 10.1016/j.socscimed.2013.06.030.

[13] Wenjie Wu, Jianghao Wang, 2017, Gentrification effects of China's urban village renewals, Urban Studies, Doi: 10.1177/0042098016631905usj.sagepub.com.

[14] Gretta Mohan, Alberto Longo, Frank Kee, 2018, The effect of area based urban regeneration policies on fuel poverty: Evidence from a natural experiment in Northern Ireland, Energy policy 114, Doi: 10.1016/j.enpol.2017.12.018.

[15] Borbely D., Rossi G., 2023, Urban regeneration projects and crime: evidence from Glasgow, Journal of Economic Geography, Doi: 10.1093/jeg/lbad021.

DOI: 10.5604/01.3001.0054.6954

PRAWIDŁOWY SPOSÓB CYTOWANIA
Polishchuk Andrii, Cysek-Pawlak Monika Maria, Krzysztofik Sylwia, 2024, Regression Analysis as a Tool for Assessing the Causal Relationships Between Urban Form and Socioeconomic Indicators of Urban Life, „Builder" 09 (326).
DOI: 10.5604/01.3001.0054.6954

ABSTRACT:

The article addresses the issue of the functional dependency between the urban environment and social life in the city. The research hypothesis concerns that the regression analysis, the tool that is designed and dedicated to extract deterministic dependencies from random data is the method that could be applied in planning research in order to access the impact of the urban form on the socioeconomic variables in the city. Authors briefly describe the possibilities of regression analysis and provide example of its application in urban planning.

KEYWORDS:

econometrics, urban planning, regression analysis

STRESZCZENIE:

ANALIZA REGRESJI JAKO NARZĘDZIE OCENY ZWIĄZKÓW PRZYCZYNOWYCH MIĘDZY FORMĄ URBANISTYCZNĄ A WSKAŹNIKAMI SPOŁECZNO-EKONOMICZNYMI ŻYCIA MIEJSKIEGO. Artykuł porusza kwestię zależności funkcjonalnej między środowiskiem miejskim a życiem społecznym w mieście. Hipoteza badawcza zakłada, że analiza regresji, narzędzie zaprojektowane i dedykowane do wyodrębniania deterministycznych zależności z losowych danych, jest metodą, która może być zastosowana w badaniach planistycznych w celu oceny wpływu formy urbanistycznej na zmienne społeczno-ekonomiczne w mieście. Autorzy krótko opisują możliwości analizy regresji i podają przykład jej zastosowania w planowaniu urbanistycznym.

SŁOWA KLUCZOWE:

ekonometria, planowanie urbanistyczne, analiza regresji

This article has been completed while author (ANDRII POLISHCHUK) was the Doctoral Candidate in the Interdisciplinary Doctoral School at the Lodz University of Technology, Poland.