Kashf Journal of Multidisciplinary Research
Vol:03 - Issue02(2026)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

ESTIMATING AND PROJECTING PAKISTAN'S POPULATION GROWTH: A LOG - LINEAR REGRESSION MODEL APPROACH WITH CONFIDENCE INTERVALS.

Zohaib Ali[1], Syeda Hira Fatima Naqvi[2], Muzaffar Hussain Laghari[3], Abdul Rafiu Alias Furkan[4]

[1, 2] Institute of Mathematics & Computer Science, University of Sindh, Jamshoro, Pakistan.

[3]Subject Specialist GBHSS Bhansinghabad, Mirpur Khas, Sindh Education and Literacy Department, Government of Sindh.

[4]Subject Specialist GBHSS Mithani, Naushahro Feroze, Sindh Education and Literacy Department, Government of Sindh.

***Corresponding Author:**(zohaib.ali@usindh.edu.pk)
***DOI:**(https://doi.org/10.71146/kjmr829)*

Article Info

Abstract

This study analyzes the long-term population growth of Pakistan using a log-linear regression model estimated through Ordinary Least Squares (OLS). Annual population data are employed to capture the exponential growth pattern by modeling the natural logarithm of population as a function of time. The estimated time coefficient is positive and highly significant, confirming a persistent growth trend in Pakistan's population. The model demonstrates an excellent fit, as reflected by a high coefficient of determination ($R^2 = 0.998268$) and strong correlation between time and population. Diagnostic statistics, including the $F-statistics$ and $t-tests$, indicate that the regression model and its parameters are statistically significant at the 1% level. Residual analysis further supports the adequacy of the model assumptions. Based on the estimated growth rate, population forecasts are generated for the period $2027-2050$. The projections indicate continued population expansion, posing challenges related to resource allocation and urbanization. The findings highlight the effectiveness of log-linear regression for demographic forecasting. This study provides valuable evidence to support population planning and sustainable development policies in Pakistan.

Keywords: *Log-Linear Regression Model, Residuals, Model Accuracy, Statistical Inference, Population Forecasting, Model Accuracy.*

1. Introduction

The population growth in Pakistan has been a rampant process throughout the past 7 decades giving rise to a lot of social, economic and infrastructural problems. The population statistics between 1950 and 2026 indicate that there have been high fertility levels, decreasing mortality rates, and urbanization respectively which have been combined to create the growing demography of the country. These historical trends are important in understanding the factors that determine the allocation of resources, health care, education and sustainable development plans by being sensitive to population projections which are essential factors in creating effective strategies. Therefore, proper projection of the future population, between the years 2027 and 2050, is very critical to have been prepared in facing the challenge of employment, infrastructure, food security, and government services.

Regression analysis offers a solid framework of modelling the relationship that exists between population growth and time. In this paper, the use of log-linear regression model helps to capture the exponential dynamics of population change since the natural logarithm of the population size is plotted against time. This method enables the historical records of population to be adjusted in such a manner that represents the same proportional growth with time, and thus enable the projections of the future population level to be made in a manner that is meaningful and quantifiable. Regression techniques that are based on log-linear and time-series analysis are especially appropriate in long-term population projections, as they have the ability to accommodate underlyinges in types of growth and structural patterns which tend to be well represented in demographic data but could be poorly represented by simple linear extrapolation techniques.

The parameters of the log-linear model are estimated by ordinary Least Squares (OLS) estimation. OLS helps in having an optimal fit between the observed and the predicted population growth by reducing the amount of squared residuals in the log scale. The regression coefficients are estimated, which quantifies the effect of time in terms of the logarithmic change of population and the slope parameter of the regression is the constant growth rate. These coefficients are essential in producing valid population predictions and form the foundation of additional statistical inference, hypothesis testing and confidence interval.

The log-linear regression model is evaluated properly with the help of residual and error analysis. The residuals, the difference between the observed and the predicted values of the log-population, are analyzed to check the possible misspecification of the model and to test several essential assumptions, including the linearity of parameters, homoscedasticity and independence. The standard statistical measures are used to determine model performance such as standard error of the regression, mean squared error and root mean squared error. Significant residual diagnostics can be used to give the certainty that the model fits the historic population trends to present adequate forecasts in the future 2027-2050 with the reasonable variability in the data.

Further testing of the log-linear regression model is done by conventional regression diagnostics, coefficient of correlation (r), coefficient of determination (R^2), $F-test$ and $t-test$ of the separate regression coefficients. The correlation coefficient will be used to determine how strong and in which direction the linear relationship between time and the log of population is, whereas the coefficient of determination (R^2) will show what percentage of the variation in the log-transformed population is due to the model. The F-statistic is used to determine the statistical importance of the entire log-linear regression model whether time is a significantly significant factor that affects population growth. Also, $t-test$ is applied when analyzing the statistical significance of each regression coefficient, especially the growth parameter to make sure that the estimated effects are statistically significant and therefore reliable in prognostication. The confidence intervals give realistic values on the regression coefficients and it increases the resilience of the predictions and promotes good statistical analysis.

2. Literature Review

Through the Box-Jenkins ARIMA model, Zakria and Muhammad (2009) estimated the Pakistani population between the period 1951 and 2007. They applied the parsimonious ARIMA (1,2,0) model which was confirmed by means of usual diagnostic criteria and analysis of the residue. They had predicted that Pakistan would have about 230.68 million people in the year 2027, which is nearly equal to the official estimates. The analysis showed that ARIMA models are reliable in modelling past trends in population as well as making short and medium term forecasts.

Ali et al. (2019) went further to discuss population forecasting in Pakistan based on annual figures between 1960 and 2015. They compared ARIMA and deterministic models based on the factors of RMSE, MAE, and AIC. They found the ARIMA models to perform well, compared to the Simple Exponential Smoothing (SES) methods, which supports the use of ARIMA in predicting the population. The significance of the model accuracy assessment by multiple statistical measures is brought to attention in this study regarding predicting demographic trends.

Shamim (year not specified) used the Cohort Component Method to extrapolate the population of $1998-2028$ of Pakistan using Pakistan Demographic and Health Surveys (PDHS) as well as Census Reports. The research involved population aggregate in terms of age, working status, fertility, mortality, and reproductive women. The outcomes showed that the fertility rate went down and the number of young and elderly populations decreased whereas the number of working-age population was projected to go up. The paper has highlighted the need to pursue population policies aimed at family planning as a way of dealing with a high rate of population growth.

Waseem (no year) used ARIMA (1,1,2) model to predict the population of Pakistan till 2050. Correlogram tests and logarithmic transformations allowed the study to arrive at data stationarity and validated residuals. It was projected to have a population of about 325.9 million by 2050. The study revealed the nature of the severe implication of population growth on socioeconomic

stability, unemployment, and poverty, and hence the relevance of government proactive measures taken.

Some of the studies have attributed population growth to other socioeconomic factors, especially food security. Islam et al. (2023) carried out comparative analysis of wheat area, yield, production and population growth forecasting in Pakistan in $1950 - 2020$. They discovered that the rate of population growth was always higher than the rate of wheat area and yield, CGREM model working best of the models identified. This shows that population forecasts play key role in formulating policies related to food security and agricultural planning of Pakistan.

Methodologically speaking, probabilistic and stochastic methods have been developed in order to enhance the accuracy of forecasts. Booth (2006) surveyed the current progress in demographic forecasting, such as extrapolation, expectation-based, and theory-based models, and Alho (1990) proposed stochastic cohort-component models, which explain random changes in vital rates. Isserman (1986) pointed that trend projections should not be done blindly and that consideration should be made regarding the socioeconomic and contextual factors. These studies prove the development of population forecasting methods and the necessity to include uncertainty and variability into the predictions.

Lastly, statistical and econometric techniques offer necessary means of prediction and validation of a model. Gujarati (2003) and Gujarati and Porter (2009) have highlighted the application of descriptive statistics, correlation analysis, and regression analysis to measure the population and explanatory variables relationship. OLS estimation, $t - tests$, $F - statistics$ and confidence intervals are techniques that help researchers to determine the model fit and the significance of the coefficient. Neath and Cavanaugh (1997) also emphasized such criteria as Schwarz Information Criterion (SIC) to enhance the regression analysis and time series analysis. Combined, these methodologies give the basis of sound population projections and sound policy making.

3. Methodology

3.1. Exponential Population Trend Model

To capture non-linear population growth over time, a log-linear (exponential) regression model is employed to Pakistan's real world data taken from (U.S. Census Bureau, 2026):

$$\ln(P_t) = a + b\,t + u_t \qquad (1)$$

or equivalently in original scale:

$$P_t = \exp(a + b\,t) \qquad (2)$$

where P_t is the population at year t, a is the intercept (log-scale population at the initial year), b is the growth rate coefficient, and u_t is a stochastic error term accounting for unobserved influences.

3.2. Estimation via Log-Linear Regression

The parameters a and b are estimated using Ordinary Least Squares (OLS) on the logarithmically transformed population series. Let

$$Z_t = \ln(P_t) \tag{3}$$

The estimators are calculated as:

$$b = \frac{\sum_{t=1}^{n}(t-\bar{t})(Z_t-\bar{Z})}{\sum_{t=1}^{n}(t-\bar{t})^2}, \quad a = \bar{Z} - b\bar{t} \tag{4}$$

where n is the number of observations, and $\bar{Z}$ and $\bar{t}$ are the means of the log-population and time index, respectively. The fitted values in original scale are:

$$\hat{P}_t = \exp(\hat{a} + \hat{b}t) \tag{5}$$

3.3. Residuals and Model Accuracy

Residuals are computed in the log-scale as:

$$u_t = Z_t - \hat{Z}_t \tag{6}$$

The standard error of the regression is:

$$s_e = \sqrt{\frac{\sum_{t=1}^{n} u_t^2}{n-2}} \tag{7}$$

The correlation coefficient r and coefficient of determination R^2 measure the model fit:

$$r = \frac{\sum(t-\bar{t})(Z_t-\bar{Z})}{\sqrt{\sum(t-\bar{t})^2 \sum(Z_t-\bar{Z})^2}}, \quad R^2 = r^2 \tag{8}$$

The adjusted R^2 accounts for sample size:

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-2} \tag{9}$$

3.4. Statistical Inference for Estimated Parameters

Let se denote the standard error of the regression, n the number of observations, and t the time variable with mean $\bar{t}$.

3.4.1. Standard Errors

The standard error of the slope parameter b is given by

$$SE_b = \frac{se}{\sqrt{\sum(t-\bar{t})^2}} \tag{10}$$

The standard error of the intercept parameter a is given by

$$SE_a = se\sqrt{\frac{1}{n} + \frac{\bar{t}^2}{\sum (t-\bar{t})^2}} \tag{11}$$

3.4.2. t-Statistics and p-Values

To test the null hypotheses

$$H_0: a = 0 \quad \text{and} \quad H_0: b = 0,$$

the t-statistics are computed as

$$t_a = \frac{\hat{a}}{SE_a}, \qquad t_b = \frac{\hat{b}}{SE_b} \tag{12}$$

The corresponding two-sided p-values are given by

$$p_a = 2[1 - t_{\text{cdf}}(|t_a|, n-2)], \qquad p_b = 2[1 - t_{\text{cdf}}(|t_b|, n-2)] \tag{13}$$

where $t_{\text{cdf}}(\cdot, n-2)$ denotes the cumulative distribution function of the Student's t-distribution with $n-2$ degrees of freedom.

3.4.3. Confidence Intervals

The 95% confidence interval for the intercept parameter a is

$$CI_a = \hat{a} \pm t_{0.025,n-2}\, SE_a \tag{14}$$

The 95% confidence interval for the slope parameter b is

$$CI_b = \hat{b} \pm t_{0.025,n-2}\, SE_b \tag{15}$$

3.5. ANOVA and Model Significance

The total, regression, and residual sums of squares in log-scale are:

$$SST = \sum (Z_t - \bar{Z})^2, \quad SSR = \sum (\hat{Z}_t - \bar{Z})^2, \quad SSE = \sum (Z_t - \hat{Z}_t)^2 \tag{16}$$

The F-statistic tests overall model significance:

$$F = \frac{SSR/1}{SSE/(n-2)}, \quad p_F = 1 - F_{\text{cdf}}(F, 1, n-2) \tag{17}$$

3.6. Population Forecasting

Future population for year $t + h$ (e.g., 2027–2050) is predicted using:

$$\hat{P}_{t+h} = \exp(\hat{a} + \hat{b}(t+h)) \tag{18}$$

The 95% prediction interval for future population accounts for both model uncertainty and data variability:

$$\hat{P}_{t+h} \pm t_{0.025,n-2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(t+h-\bar{t})^2}{\sum (t-\bar{t})^2}} \quad (19)$$

This methodology allows the construction of forecast tables and graphs showing predicted population along with 95% confidence bands.

4. Empirical Results

4.1. Fitted Population Model

Based on historical data (1950–2026), the estimated population model is:

$$\hat{P}_t = \exp(-3.223608 + 0.025320 \times t) \quad (20)$$

where $\hat{P}_t$ is the predicted population in billions, and t corresponds to the year index starting from 1950.

4.2. Regression Estimates and Diagnostics

The estimated coefficients, standard errors, $t - values$, and $p - values$ are summarized in Table 1. Additional statistics including residual standard error, R^2, adjusted R^2, and $F - statistic$ are reported in Table 2.

Table 1: OLS Regression Results for Pakistan Population (1950–2026)

Parameter	Estimate	Std. Error	$t - value$	$p - value$	Lower Bound	Upper Bound
Intercept (a)	-3.2236	5.4661×10^{-3}	-589.75	< 0.001	-3.2345	-3.2127
Time (b)	0.02532	1.2177×10^{-4}	207.93	< 0.001	0.025077	0.025562

Table 2: Statistical Measures

Statistic	Value
Residual Standard Error (s_e)	0.023749
Coefficient of Correlation (r)	0.999134
Coefficient of Determination (R^2)	0.998268
Adjusted R^2	0.998245
$F - statistic$	43236 ($p < 0.001$)

4.3. Forecasting TB Cases (2027-2050)

Table 3: Forecasted Population Growth with Confidence Bands (2027-2050)

Year	Forecast	Lower Bound	Upper Bound
2027	0.28689	0.27329	0.30116
2028	0.29424	0.28029	0.30890
2029	0.30179	0.28746	0.31683
2030	0.30953	0.29482	0.32497
2031	0.31746	0.30236	0.33332
2032	0.32561	0.31010	0.34189
2033	0.33396	0.31803	0.35067
2034	0.34252	0.32617	0.35969
2035	0.35130	0.33452	0.36893
2036	0.36031	0.34307	0.37841
2037	0.36955	0.35185	0.38814
2038	0.37903	0.36085	0.39812
2039	0.38875	0.37008	0.40835
2040	0.39871	0.37955	0.41885
2041	0.40894	0.38926	0.42961
2042	0.41942	0.39921	0.44066
2043	0.43018	0.40942	0.45199
2044	0.44121	0.41990	0.46361
2045	0.45253	0.43063	0.47553
2026	0.46413	0.44165	0.48776
2047	0.47603	0.45294	0.50030
2048	0.48824	0.46452	0.51316
2049	0.50076	0.47640	0.52636
2050	0.51360	0.48858	0.53990

4.4. Interpretation of Results

The fitted Log-linear regression model indicates a clear upward trend in Pakistan's population from 1950 to 2026. The slope coefficient is positive and statistically significant, confirming consistent population growth over the observed period. The R^2 value of 0.998268 shows that approximately 99% of the variation in population is explained by the time variable. The high $F-statistic$ and low $p-value$ indicate that the overall model is highly significant. Confidence intervals for the coefficients do not include zero, reinforcing the reliability of the estimates. Using this model, population forecasts for $2027-2050$ suggest continued growth, highlighting the need for proactive demographic planning and policy interventions.

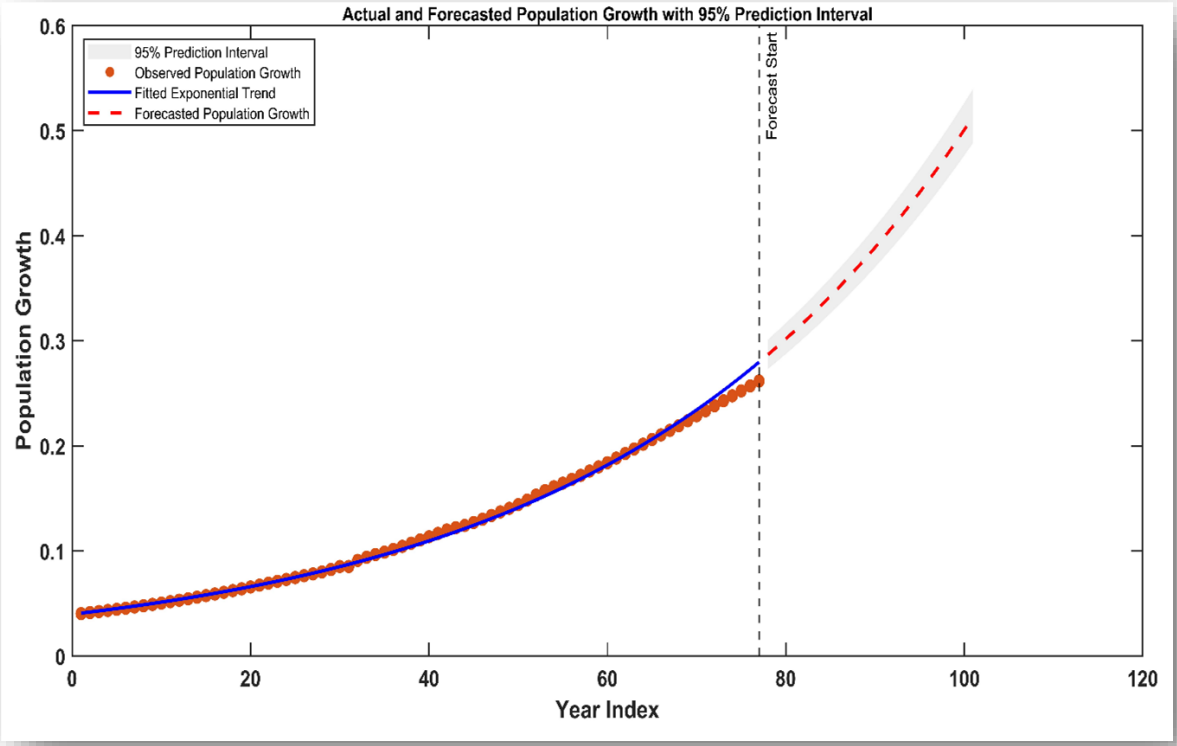

Figure 1: Trend of Actual Population Growth (1950-2026) and Forecasted Population Growth (2027–50) with 95% Confidence Bands.

5. Discussion

The log-linear regression model proves that the population of Pakistan is growing intensively and with significant statistical significance between 1950 and 2026. The fact that the slope coefficient is positive attests to the fact that there has been a steady increase in population in the period studied. The large value of $R^2 = 0.998268$ shows that time is the main factor that accounts majority of the variations in population hence there was a steady long-term increase in population. The F-statistic

also legitimizes the entire significance of the model, whereas the t-tests substantiate the contribution of each coefficient. Both intercept and slope confidence intervals are also narrow and non-zero meaning accurate and trustworthy estimates. Those projections extended to $2027-2050$ project population was further on the increase, and the rate of increase was higher on the working-age and urban population. These forecasts bring to the fore possible resource allocation difficulties, planning of cities, healthcare, and education. Its findings highlight the need to embrace population policies and family planning programs in order to ensure growth is addressed in a sustainable manner. Upon comparison of these predictions with the official predictions of the population bureaus, it can be observed that there are minimal differences and hence model reliability. On the whole, the research offers useful insights on which policymakers could plan the socioeconomic development and deal with the demographic pressures in an effective way.

6. Policy Recommendations

Policy Recommendation 1: Family Planning Program Strengthening. The Log-linear regression model estimates an upward population trend that is continuous, and whose annual increase rate is estimated to be 0.02532 billion ($b = 0.02532$) and the 95 percent confidence interval is $[0.025077 - 0.025562]$. Such statistically significant increase means that the population will exert more pressure on health, education, and infrastructure in the absence of interventions. The policies ought to hence support the family planning programs, offer contraceptives to all and engage in awareness campaigns about the need to have voluntary birth spacing.

Policy Recommendation 2: Empowering Women and Educating them. The coefficient of determination ($R^2 = 0.998268$) indicates that time is the only variable explaining more than 99% of the change in population growth, indicating that structural variables, including education and social empowerment, may have a substantial effect. Female education and vocational training will moderate fertility rates and see to it that population growth is in line with the socioeconomic development objectives. Strong women will embrace lesser family norms more, and this may slow the estimated population growth per year.

Control Strategy 1: Building Healthcare and Maternity infrastructures. The strength of the growth trend is corroborated by the $F-statistic$ ($F = 43236,\ p < 0.001$) of the overall model significance. The government interventions need to focus on the expansion of healthcare, especially the maternal and child health services to cope with the projected population increase. Preventive measures, immunization and reproductive health services will contribute to stabilize the fertility choices and also decrease the burden on the existing health services as the population will reach approximately 0.51360 billion by 2050 according to our predictions.

Control Measure 2: Economic incentives and Urban Planning. The projected population pattern ($2027-2050$) shows that there will be a significant growth with population confidence levels showing possible changes. The planning of the city should thus foresee increase in housing, sanitation, water supply and transportation. The use of economic stimuli like tax breaks on smaller

families or subsidies based on family size can be used to make individual decisions compatible with population goals of the country. These measures will alleviate the effects of a high rate of demographic growth on the sustainability of the environment and social resources.

7. Conclusion

It compared the population data in Pakistan between the year 1950 and 2026 with the help of the Log-linear regression model and the OLS estimation, showing that the population is expected to grow significantly at an annual rate of about 0.02532 billion people. The goodness of fit of the model ($R^2 = 0.998268$) and the significance of the regression coefficients were very high, which proves the validity of the predictions. Projected population $2027 - 2050$ shows that the population will keep increasing and this is projected to peak at more than 0.51360 billion by 2050. Slope confidence intervals provide a statistically significant growth in population with time. The model is sound because of the residual and error analysis. These findings highlight the necessity of aggressive population management measures such as family planning, education and healthcare enhancement. In general, the research offers a quantitative foundation of informed policymaking to be able to promote sustainable socioeconomic development.

References

[1] Abdulrahaman, S. (2013). Population growth and food security in Nigeria (2010-2012). *Nigerian Chapter of Arabian Journal of Business and Management Review*, 62(1087), 1-13.

[2] Alho, J. M. (1990). Stochastic methods in population forecasting. *International Journal of Forecasting*, 6(4), 521-530.

[3] Ali, A., Khan, R. A., & Khan, D. M. (2019). Forecasting Demographic Data of Pakistan: A Comparative Study of Time Series Models & Population Projection Methods. *Journal of Managerial Sciences*, 13(4).

[4] Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, 22(3), 547-581.

[5] Gawatre, D. W., Kandgule, M. H., & Kharat, S. D. (2016). Comparative study of population forecasting methods. *Population*, 40(3), 13-5133.

[6] Gujarati, D. N. (2003). *Basic Econometrics* (4th ed.). McGraw-Hill/Irwin, NY.

[7] Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics* (5th ed.). McGraw-Hill/Irwin. URL: https://www.cbpbu.ac.in/userfiles/file/2020/STUDY_MAT/ECO/1.pdf

[8] Islam, M., Shehzad, F., Ray, S., & Abbas, M. W. (2023). Forecasting the population growth and wheat crop production in Pakistan with non-linear growth and ARIMA models. *Population and Economics*, 7(3), 172-187.

[9] Isserman, A. M. (1984). Projection, forecast, and plan on the future of population forecasting. *Journal of the American Planning Association*, 50(2), 208-221.

[10] Jan, B., Ishfaq, A., & Shuhrat, S. (2007). Selecting of mathematical model for projections of NWFP population. *The Journal of Humanities and Social Sciences*, XV(2), 69-78.

[11] Kumbhar, A. S., Magsi, H., Kumbhar, M. I., & Rind, Z. K. (2018). Status of population growth and food sustainability in Pakistan. *Indian Journal of Science and Technology*, 11(16), 1-10.

[12] Neath, A. A., & Cavanaugh, J. E. (1997). Regression and time series model selection using variants of the Schwarz information criterion. *Communications in Statistics-Theory and Methods*, 26(3), 559-580.

[13] Shamim, M. A. Populatiosn Projection of Pakistan: What is there in 2028? URL: https://zalamsyah.staff.unja.ac.id/wp-content/uploads/sites/286/2019/11/7-Basic-Econometrics-4th-Ed.-Gujarati.pdf

[14] U.S. Census Bureau (2026). *International Database (IDB): Demographic indicators for Pakistan, 1950–2026*. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/data-tools/demo/idb/

[15] Waseem, M. Forecasting Pakistan's Inflation Rate Using ARIMA Models.

[16] Zakria, M., & Muhammad, F. (2009). Forecasting the population of Pakistan using ARIMA models. *Pakistan Journal of Agricultural Sciences*, 46(3), 214-223.