

**Licensing Statement:** This article is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0) (<https://creativecommons.org/licenses/by/4.0/>).[11]

## STATISTICAL MODELING, ESTIMATION AND FORECASTING OF HEALTH INSURANCE CLAIM AMOUNTS IN INDIA

SANDEEP MISHRA AND SAUMYA GOVIL

**ABSTRACT.** The Indian health insurance sector has experienced consistent expansion in recent years, driven by rising medical costs, greater policy uptake and increasing awareness of healthcare needs[cite: 2]. This growth makes accurate health insurance claim amounts forecasting critical to maintain premium adequacy and ensure reserve sufficiency[cite: 2]. However, the stochastic nature of insurance claims introduces structural complexity that often outperform the conventional forecasting models[cite: 2]. This study examines the evolution of annual health insurance claims in India using secondary data obtained from IRDAI reports[cite: 2]. We begin with multiple linear regression modelling to investigate key determinants of sector performance and extend to the derivation and implementation of a Generalized Bayes estimation framework, incorporating shrinkage to manage parameter uncertainty and improve coefficient stability[cite: 2]. A comparative analysis of Ordinary Least Squares and Generalized Least Squares estimators is conducted under a first-order autoregressive error structure, demonstrating the efficiency gains of GLS in the presence of serial dependence[cite: 2]. Additionally, a suite of univariate time series models, including ARMA(1,2), ARIMA(1,2,2), ARIMA(0,1,2), ARIMA(0,1,1) and ARIMA(1,0,1) are estimated and compared to capture the temporal dynamics and long run growth trajectory of incurred claim amounts[cite: 2]. By integrating Bayesian shrinkage methods with classical and time series estimation approaches, this research offers a more resilient framework for actuarial forecasting and risk management in a volatile insurance environment[cite: 2].

### 1. INTRODUCTION

From actuarial perspective, health insurance presents unique challenges compared to other lines of insurance[cite: 2]. This inherent volatility makes effective claim forecasting a critical function for insurers[cite: 2]. Inadequate estimation of future claims can lead to under reserving, pricing inefficiencies and financial stress while being liberal may result in uncompetitive premiums[cite: 2]. Given the growing scale and complexity of health insurance operations in India, there is an increasing reliance on quantitative methods to support decision-making[cite: 2]. In this context, accurate forecasting of health insurance claims is not merely an academic exercise but a vital business imperative as discussed by Sanyal M. (2025)[cite: 2].

Health insurance claims represent the financial obligation of insurers towards the policyholders to indemnify them against covered medical expenses[cite: 2]. A claim arises when an insured individual seeks reimbursement or cashless settlement for medical expenses incurred due to illness, injury or hospitalization subject to the

---

2000 *Mathematics Subject Classification.* Primary 92C50; Secondary 34A34, 65C05.

*Key words and phrases.* Indian Health Insurance Claim Amounts, Generalized Bayes Estimation, ARIMA, OLS, GLS.

terms and conditions of the policy[cite: 2]. The nature of these claims is inherently stochastic, influenced by a complex interdependence between demographic, medical, economic and behavioral factors[cite: 2]. Unlike premiums, which are collected in advance and are relatively stable over short periods, claims exhibit significant volatility and reflect the realization of insured health risks[cite: 2]. Health insurance claims depend majorly on claim frequency and claim severity which both are influenced by diversity of factors as discussed by Nargunam R., Rathee N. (2025)[cite: 2]. Claim frequency depends on age, gender and behavioural characteristics including access to medical facilities[cite: 2]. Claim severity on the other hand, is influenced by treatment costs, length of hospital stay, etc[cite: 2]. The core underwriting principle of pooling risk relies on collecting sufficient premiums from a large group to pay for the claims of the unfortunate ones[cite: 2]. If aggregate claims consistently exceed premium plus investment income, insurers faces underwriting losses, threatening its solvency and long-term viability[cite: 2].

Claims represent uncertain future liabilities, and their reliable estimation is a foundational business process that directly underpins an insurer's financial stability, regulatory compliance, and competitive positioning[cite: 2]. Without robust forecasting, insurers operate with limited visibility into future cash flow requirements[cite: 2]. Inaccurate forecasts can lead to the under-pricing of policies, where premiums are insufficient to cover subsequent claims, thereby eroding underwriting profit and threatening solvency[cite: 2]. Conversely, systematic overestimation leads to excessively high premiums that reduce market competitiveness and stifle product adoption in a country where coverage is still expanding[cite: 2]. The Insurance Regulatory and Development Authority of India (IRDAI) requires insurers to maintain sufficient buffers to meet all future policyholder obligations[cite: 2]. By applying modern time series and regression forecasting techniques to authentic industry data, this study aims to contribute to a better understanding of claim dynamics and provide practical insights for informed decision-making and financial sustainability in the sector[cite: 2].

The primary objective of this study is to develop, apply and critically evaluate statistical models for forecasting aggregate health insurance claim amounts in India[cite: 2]. Driven by the necessity for accurate predictions, this research seeks to address the need for systematic, data driven approaches within the growing and complex Indian insurance sector[cite: 2]. The objective involves exploratory analysis of historical claim data obtained from the Insurance Regulatory and Development Authority of India (IRDAI)[cite: 2]. The analysis will investigate the empirical relationships between claim amounts and potential variables such as number of policies in force, net premium earned and other relevant financial indicators[cite: 2]. It forms the core analytical component of the study, entailing the specification and estimation of several competing models[cite: 2]. The research will construct and compare multiple regression models including full model with all available predictors and more parsimonious models based on key variables[cite: 2]. It will also develop a suite of time series models to capture the inherent temporal dynamics and autocorrelation structure of claim series[cite: 2]. The study presents the theoretical derivation of the generalized Bayes estimator for regression coefficients and examines the efficiency of OLS and GLS estimators under autocorrelated error structures[cite: 2]. By providing context-specific, evidence-based guidance, this research aims to enhance the understanding of health insurance claim dynamics and

support risk management and financial sustainability in the Indian health insurance industry[cite: 2].

This research is exclusively based on secondary time series data published by the Insurance Regulatory and Development Authority of India (IRDAI), where the primary variable of interest is the aggregate health insurance claim amount incurred by the industry or a significant segment thereof[cite: 2]. Companion variables, such as the number of policies in force, net premium earned and other related financial metrics from the same source are used as potential predictors in regression models[cite: 2]. By focusing on nationally aggregated annual data, the study aims to capture the broader evolution and long-term trends of health insurance claims in the Indian context, providing a macro-level perspective rather than insights into individual insurer portfolios or policyholder-level dynamics[cite: 2]. Such aggregated data are particularly relevant for applications such as reserve planning, solvency assessment and strategic decision-making, where insurers are concerned with overall claim outgo[cite: 2].

The implementation of these models is carried out using the R statistical programming environment[cite: 2].

## 2. LITERATURE REVIEW

The Indian insurance landscape has been a subject of rigorous academic study, particularly concerning the efficiency of claim settlement and the factors influencing claim behaviour[cite: 2]. A foundational study by Bhoi and Satpathy (2025) examined long-term trends in claim settlement ratios, repudiation rates, pending claims, and unclaimed amounts using official regulatory and insurer-level data[cite: 2]. Employing Time Series Regression and Data Envelopment Analysis (DEA), the research underscored the importance of consistent definitions and regulatory data sources when analysing claim behaviour over time[cite: 2]. Complementing this, Sen and Chaudhuri (2017) established the critical need for structured time-series methodologies in the volatile Indian healthcare sector, proposing analytical frameworks that utilise decomposition techniques to isolate long-term trends and seasonal variations for strategic planning[cite: 2]. Their work also provides methodological guidance on implementing models such as ARIMA, involving a disciplined workflow of stationarity testing, parameter identification via ACF and PACF plots, and rigorous residual validation[cite: 2].

A consistent theme in the literature emphasises the systemic importance of accurately forecasting claim reserves to ensure insurer solvency and maintain policyholder trust[cite: 2]. Brotosaputro et al. (2025) directly addressed the forecasting of claim fund reserves through the ARIMA (Autoregressive Integrated Moving Average) model, motivated by the high variability of claim amounts and timings that threaten corporate liquidity[cite: 2]. Following the structured Box-Jenkins framework – encompassing data pre-processing, stationarity testing, and model parameter identification – the study demonstrated that while short-term monthly forecasts can be volatile, ARIMA provides relatively accurate and stable predictions for annual aggregate claim requirements, establishing it as a robust procedure for reserve estimation[cite: 2]. The general ARMA equation applied in this context takes the form:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + e_t + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q}$$

capturing both autoregressive and moving average dynamics in claim series[cite: 2].

Comparative methodological studies evaluate the performance of classical and modern forecasting approaches[cite: 2]. Albeladi et al. (2023) pitted ARIMA against deep learning-based LSTM (Long Short-Term Memory) networks, finding that while LSTM excels at capturing complex, non-linear sequential patterns, ARIMA retains significant advantages in terms of statistical transparency, reliability, and stability – particularly in scenarios with limited historical data, a common constraint in the Indian insurance context[cite: 2]. This reinforces ARIMA’s role as a highly effective baseline for claim behaviour forecasting[cite: 2]. In parallel, Ibrahimoh (2024–25) documented a significant technological shift in claim evaluation, demonstrating that machine learning algorithms – particularly Random Forest models – achieve high predictive accuracy in determining individual claim outcomes, substantially outperforming simpler decision-tree methods[cite: 2]. While this micro-level approach differs from aggregate forecasting, it underscores the complexity of claim behaviour and positions artificial intelligence as a powerful tool for enhancing the precision and efficiency of insurance claim workflows[cite: 2].

The literature also identifies several important limitations and research gaps[cite: 2]. The analyses of Bhoi and Satpathy (2025) are largely descriptive and retrospective, focusing on historical aggregate metrics rather than generating explicit, forward-looking predictive models for reserve planning[cite: 2]. Their reliance on national-level data may further obscure regional variations or challenges specific to particular policy types[cite: 2]. Brotoaputro et al. (2025) acknowledge that ARIMA’s accuracy diminishes for shorter-term predictions and that existing approaches often inadequately account for the simultaneous occurrence of multiple large claims – a key insolvency risk[cite: 2]. Albeladi et al. (2023) highlight persistent uncertainty regarding optimal model selection, noting that deep learning models are constrained by data availability, while Sen and Chaudhuri (2017) caution that classical models like ARIMA, reliant on stationarity assumptions, may not adapt well to structural breaks or regulatory shifts[cite: 2]. Finally, Ibrahimoh (2024–25) points to a gap between micro-level AI precision and the macro-level, time-dependent task of forecasting aggregate claim volumes critical to actuarial functions[cite: 2]. Collectively, these gaps motivate the need for forecasting approaches that effectively balance accuracy, interpretability, and feasibility under conditions of data constraint within the Indian insurance sector[cite: 2].

### 3. DATA DESCRIPTION AND METHODOLOGY

**3.1. Description of Health Insurance Data Used.** The empirical foundation of this research is built upon secondary data that captures the aggregate performance of the health insurance sector in India[cite: 2]. To ensure the analytical reliability and official standing of all findings, the dataset was procured directly from the Annual Reports and the Handbook on Indian Insurance Statistics issued by the Insurance Regulatory and Development Authority of India (IRDAI)[cite: 2]. The specific data utilized consists of annual records of health insurance claim amounts, number of policies and gross premium amount[cite: 2]. Incurred claim amounts serve as the primary variable under observation, effectively capturing the aggregate financial liability, the combined outcome of claim frequency and severity, which insurers are obligated to cover[cite: 2]. Gross Premium is defined as the total premium income collected by insurers before any deductions for reinsurance[cite: 2].

2]. From an actuarial standpoint, it is intrinsically linked to expected claim costs, as premiums are fundamentally calculated based on anticipated future claims[cite: 2]. Number of Policies quantifies the volume of health insurance contracts in force and acts as a proxy for the breadth of coverage[cite: 2].

TABLE 1. Policies premium and claim amount in India\* (all data in crores)

| Year | Number of Policies | Gross Premium | Incurred Claim Amounts |
|------|--------------------|---------------|------------------------|
| 2010 | 0.61               | 8109.18       | 6394.04                |
| 2011 | 0.70               | 11031.00      | 7744.15                |
| 2012 | 0.82               | 13069.00      | 9013.42                |
| 2013 | 0.89               | 15453.00      | 10834.29               |
| 2014 | 1.00               | 17495.00      | 14007.22               |
| 2015 | 1.09               | 20096.00      | 18222.93               |
| 2016 | 1.18               | 24448.00      | 21758.91               |
| 2017 | 1.31               | 30392.00      | 27545.60               |
| 2018 | 1.47               | 37029.00      | 30244.04               |
| 2019 | 2.07               | 44873.00      | 34982.82               |
| 2020 | 1.79               | 50758.00      | 40025.59               |
| 2021 | 2.37               | 58238.00      | 43534.60               |
| 2022 | 2.26               | 73052.00      | 69498.48               |
| 2023 | 2.26               | 89492.00      | 70929.82               |
| 2024 | 2.68               | 107681.00     | 83493.17               |
| 2025 | 2.65               | 117505.00     | 94247.60               |

**3.2. Regression Modelling Framework.** This study employs a multiple linear regression framework as an initial analytical step to investigate the fundamental relationships between the core financial and operational variables within the Indian health insurance sector[cite: 2]. Regression analysis serves a critical dual purpose: first, to quantify the influence of key factors on sector performance and second, to establish an interpretable actuarial benchmark against which the dynamic time series models can later be evaluated[cite: 2].

To examine hypotheses, model specifications were developed and estimated using the Ordinary Least Squares (OLS) method[cite: 2]. Model performance and validity are rigorously evaluated[cite: 2]. Goodness-of-fit is assessed using the coefficient of determination ( $R^2$ ) and the adjusted  $R^2$ [cite: 2]. Furthermore, standard regression assumptions including linearity, independence of errors, homoscedasticity and normality of residuals are examined through diagnostic checks such as residual plots as discussed by Aiken L. S. (1991) and James G. et al. (2021)[cite: 2].

**3.3. Time Series Modelling Framework.** Time series modelling framework fundamentally differs from the regression framework by treating the claim series not as a function of other variables, but as a stochastic process whose future values depend on its own past behaviour, trends and random shocks[cite: 2]. Given that health insurance claims exhibit clear persistence, long-term growth and sequential dependence, autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models provide a theoretically sound and practically

effective structure for generating forecasts based solely on historical claim experience, a critical capability for actuarial reserve planning when future covariates are uncertain as discussed by Chatfield C. (5th edition)[cite: 2].

The initial step involves assessing the stationarity of the incurred claims series, a prerequisite for ARMA modelling[cite: 2]. To address this, differencing is applied; the order of integration ( $d$ ) is chosen to transform the series into a stationary one by subtracting previous observations  $Y'_t = Y_t - Y_{t-1}$ [cite: 2]. Following stationarity transformation, the model identification stage utilizes autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the differenced series to identify potential orders for the autoregressive ( $p$ ) and moving average ( $q$ ) components as discussed by Chatfield C. (5th edition) and Sen J. (2017)[cite: 2].

A critical subsequent stage is diagnostic checking, where the residuals of the fitted models are scrutinized to ensure they resemble white noise, random and uncorrelated, confirming that the model has successfully captured all systematic patterns in the data[cite: 2].

**3.4. Generalized Bayes Estimator.** To improve the stability of parameter estimation in regression models, this study considers a Bayesian estimation framework for the regression coefficients[cite: 2]. Bayesian approaches are particularly useful in actuarial and insurance applications where data may exhibit volatility or limited sample sizes, leading to unstable parameter estimates under classical estimation methods as discussed by Chaturvedi A. and Mishra S. (2019)[cite: 2]. The generalized Bayes estimator incorporates prior information about the parameters and introduces a shrinkage mechanism that improves estimation stability[cite: 2]. Consider the standard linear regression model

$$Y = X\beta + u$$

Under the classical framework, the regression coefficients are estimated using the ordinary least squares estimator

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Within Bayesian framework, regression coefficients are treated as random variables and a prior distribution is assigned to them[cite: 2]. In this study, a normal prior is assumed[cite: 2]

$$\beta \sim N(0, \sigma^2 g(X'X)^{-1})$$

Where ' $g$ ' represents a shrinkage parameter[cite: 2]. A hyper-prior distribution is also specified[cite: 2]

$$g = \frac{1 - \lambda}{\lambda}, \quad 0 < \lambda < 1$$

Which allows the shrinkage intensity to vary depending on the data[cite: 2]. By combining likelihood function of regression model with prior distribution, joint posterior distribution of the parameters can be obtained[cite: 2].

Under certain conditions on the prior parameters, the estimator can be shown to be admissible and minimax as discussed by Chaturvedi A. and Mishra S. (2019)[cite: 2]. These properties imply that the estimator provides improved risk performance relative to traditional estimators under quadratic loss[cite: 2]. In context of insurance claim modelling, the shrinkage effect of the generalized Bayes estimator helps stabilize regression coefficients[cite: 2].

### 3.5. Comparison of Ordinary Least Squares and Generalized Least Squares Estimators.

In the classical linear regression framework, the parameters of the model are typically estimated using the Ordinary Least Squares (OLS) estimator[cite: 2]. One of the key assumptions underlying OLS estimation is that the error terms are independently distributed with constant variance and zero covariance[cite: 2]. However, in many financial and actuarial datasets, particularly those involving time series observations such as annual insurance claims, the assumption of independence among error terms may not hold[cite: 2]. When regression errors are correlated over time, the model is said to exhibit autocorrelation or serial correlation[cite: 2]. In the presence of autocorrelated errors, the OLS estimator remains unbiased but no longer retains its property of minimum variance among linear unbiased estimators as discussed by Chatfield C. (5th edition)[cite: 2].

To address this limitation, the Generalized Least Squares (GLS) estimator is employed[cite: 2]. The GLS method modifies the estimation procedure by explicitly incorporating the covariance structure of the error terms into the regression model[cite: 2]. Consider the linear regression model

$$Y = X\beta + u$$

If the covariance matrix of errors is denoted by[cite: 2]

$$\text{Var}(u) = \sigma^2\Omega$$

Where  $\Omega$  a matrix capturing correlation among the disturbances, the GLS estimator is[cite: 2]

$$\hat{\beta}_{\text{GLS}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

Unlike the OLS estimator, the GLS estimator accounts for the non-spherical structure of the error covariance matrix and therefore provides efficient parameter estimates in the presence of autocorrelation[cite: 2]. In this study, particular attention is given to the case where the regression errors follow a first-order autoregressive process, denoted as AR(1)[cite: 2]. Under this structure, the covariance matrix  $\Omega$  depends on the autocorrelation parameter  $\rho$ , which measures the strength of serial dependence between successive error terms[cite: 2].

## 4. EMPIRICAL ANALYSIS

**4.1. Regression Model Estimation and Results.** This section presents the empirical findings from the regression analysis conducted to identify the determinants of the current year's incurred claim amounts in the Indian health insurance sector[cite: 2]. Utilizing historical aggregate data, model specifications were estimated via the functions in R programming[cite: 2].

The model was estimated as a comprehensive specification, incorporating all available prior-year explanatory variables: Gross Premium and Number of Policies[cite: 2]. This model achieved an exceptionally high goodness of fit, with an adjusted  $R^2$  of 0.9847 and  $R^2$  of 0.9867[cite: 2]. Individual p-values for number of policies is quite high, making it statistically insignificant but gross premium is significant at all conventional level of significances[cite: 2].

**4.2. Time Series Model Estimation.** This section details the estimation and interpretation of the time series models applied to the annual aggregate health insurance claim series in India[cite: 2]. Motivated by the limitations of regression approaches, particularly their inability to adequately model temporal dependence,

these models serve as the primary forecasting framework[cite: 2]. The objective is to capture the dynamic structure of claim amounts, characterized by a strong upward trend, persistence and random fluctuations, and to identify the most suitable ARMA and ARIMA specifications for future forecasting[cite: 2]. Five distinct models were fitted using R programming, providing a comparative landscape from a simple baseline to a complex trend-correcting framework[cite: 2].

#### ARMA(1,2) Model

This model combines one autoregressive term and two moving average terms, expressed as[cite: 2]

$$Y_t = c + \phi_1 Y_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}$$

The estimated parameters revealed autoregressive effect  $\phi_1 = 0.9633$  and moving average components  $\theta_1 = -0.1455$ ,  $\theta_2 = 0.732$  with intercept  $c = 48087.14$ [cite: 2]. This indicates that while current claim amounts are influenced by past levels, recent random shocks or unexpected fluctuations have a significant and persisting impact on the series[cite: 2].

#### ARIMA(1,2,2) Model

To directly address the pronounced non-stationarity and accelerating growth trajectory of the claim data, this model was estimated[cite: 2]. This specification applies second-order differencing ( $d = 2$ ), effectively modelling the "change in the growth rate" of claims to achieve stationarity before applying autoregressive and moving average components[cite: 2]. The estimated parameters are  $AR_1 = 0.5013$ ,  $MA_1 = -1.7015$  and  $MA_2 = 0.9754$ [cite: 2].

#### ARIMA(0,1,2) Model

This model was estimated as a purely moving average specification with first-order differencing ( $d = 1$ )[cite: 2]. The structure excludes an autoregressive component, focusing entirely on error correction mechanisms to model the differenced series, making it suitable for series where short-term irregular movements dominate[cite: 2]. The estimated coefficients are  $MA_1 = -0.1335$  and  $MA_2 = 0.6886$ [cite: 2].

#### ARIMA(0,1,1) Model

This model, statistically equivalent to simple exponential smoothing, serves as a parsimonious and stable baseline as discussed by Munarsih E., Sauza I. (2020)[cite: 2]. The estimated moving average term is  $MA_1 = 0.2150$ [cite: 2].

#### ARIMA(1,0,1) Model

This model was chosen by observing the autocorrelation function (ACF) plot and partial autocorrelation function (PACF) plot of the incurred claim amounts[cite: 2]. Since there was only 1 significant spike in both the plots, we decided the  $p$  and  $q$  values as 1[cite: 2]. The estimated terms for this model are  $AR_1 = 0.9692$  and  $MA_1 = 0.2235$ [cite: 2].

TABLE 2. 3 step ahead forecast values using ARIMA (1, 0, 1)

| Year            | 2026        | 2027        | 2028        |
|-----------------|-------------|-------------|-------------|
| Predicted claim | 94770.82 Cr | 93333.04 Cr | 91939.56 Cr |

**4.3. To Compare between OLS & GLS Estimators.** To analyze the relationship between incurred claim amounts, number of policies and gross premium, a

multiple linear regression model was considered[cite: 2]:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t$$

Where[cite: 2]:  $y_t$  = incurred claim amount

$x_{1t}$  = number of policies

$x_{2t}$  = gross premium

$u_t$  = random error terms

Since observations are time ordered, error terms were assumed to follow a first order autoregressive process AR(1)[cite: 2]:

$$u_t = \rho u_{t-1} + e_t$$

First, the model was estimated using ordinary least square method[cite: 2]:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$$

Where  $X$  is a matrix consisting of a column of ones (intercept), number of policies and gross premium amounts[cite: 2].

The estimated coefficients obtained were[cite: 2]:  $\beta_0 = -79.7327$ ,  $\beta_1 = 539.9535$ ,  $\beta_2 = 0.7933$

Thus the regression equation obtained was[cite: 2]:

$$y_t = -79.7327 + 539.9535x_{1t} + 0.7933x_{2t}$$

The residuals from OLS were then obtained and AR(1) parameter  $\rho$  was estimated[cite: 2]:

$$\rho = \frac{\sum_{t=2}^n u_t u_{t-1}}{\sum_{t=2}^n u_{t-1}^2}$$

Using the  $\rho$ , the covariance matrix of error terms was constructed as[cite: 2]:

$$\Omega_{ij} = \sigma^2 \rho^{|i-j|}$$

Since the errors exhibit autocorrelation, generalized least square estimator was used[cite: 2]:

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

The estimated coefficients obtained were[cite: 2]:  $\beta_0 = -646.0701$ ,  $\beta_1 = 1332.5115$ ,  $\beta_2 = 0.7782$

Thus the GLS regression equation obtained was[cite: 2]:

$$y_t = -646.0701 + 1332.5115x_{1t} + 0.7782x_{2t}$$

The GLS estimator account for the autocorrelation present in error terms, whereas OLS ignore this structure[cite: 2]. Although OLS estimates remain unbiased, they are inefficient in the presence of autocorrelated errors[cite: 2]. GLS corrects for this by incorporating the covariance structure of errors, resulting in more efficient parameter estimates[cite: 2].

## 5. CONCLUSION

This study set out to develop and evaluate statistical models for forecasting aggregate health insurance claim amounts in India using secondary data sourced from the Insurance Regulatory and Development Authority of India (IRDAI)[cite: 2]. The research pursued three interrelated analytical objectives: examining the regression-based determinants of incurred claim amounts, identifying an appropriate time series model for claims forecasting, and comparing the efficiency of OLS

and GLS estimators in the presence of autocorrelated errors[cite: 2]. The empirical findings across all three frameworks yield a coherent and practically meaningful picture of health insurance claim dynamics in India[cite: 2].

The regression analysis demonstrated that gross premium is a statistically significant and economically meaningful predictor of incurred claim amounts, with an adjusted  $R^2$  of 0.9847 confirming that the model accounts for nearly all systematic variation in the dependent variable[cite: 2]. The number of policies in force, while intuitively relevant as a proxy for exposure, was found to be statistically insignificant, suggesting that at the aggregate national level, premium income subsumes much of the explanatory variation attributable to policy volume[cite: 2]. The exceptionally high goodness of fit affirms that the macro-level financial indicators tracked by the IRDAI are internally consistent[cite: 2].

The time series analysis evaluated five competing specifications – ARMA(1,2), ARIMA(1,2,2), ARIMA(0,1,2), ARIMA(0,1,1), and ARIMA(1,0,1) – to capture the dynamic structure of the incurred claim series[cite: 2]. The first four models were considered in light of their prominence in existing insurance forecasting literature; however, their primary purpose in this study was comparative and contextual[cite: 2]. The ARIMA(1,2,2) model, through second-order differencing, addressed the accelerating growth trajectory of claims but at the cost of parsimony[cite: 2]. The ARIMA(0,1,2) and ARIMA(0,1,1) models offered simpler moving-average structures, with the latter being equivalent to exponential smoothing and serving as a useful benchmark as discussed by Munarsih E., Sauza I. (2020)[cite: 2]. The ARMA(1,2) model captured the high persistence in claim levels through a near-unit autoregressive coefficient[cite: 2].

The ARIMA(1,0,1) model emerged as the most appropriate specification for the Indian health insurance claim series[cite: 2]. This selection was grounded not in methodological convention but in empirical evidence: the ACF and PACF plots of the incurred claim series each exhibited a single significant spike, directly suggesting an ARIMA(1,0,1) structure[cite: 2]. Crucially, the residuals of the fitted ARIMA(1,0,1) model were subjected to the Ljung-Box test, and the results confirmed that the residuals behave as white noise – indicating that the model has successfully captured all systematic temporal patterns in the data and that no predictable structure remains unexplained[cite: 2]. With estimated parameters  $AR_1 = 0.9692$  and  $MA_1 = 0.2235$ , the model reflects both the strong persistence in annual claim levels and the influence of recent random shocks, making it well-suited for producing reliable short-to-medium-term forecasts of aggregate claim amounts in the Indian context[cite: 2].

The comparison between OLS and GLS estimators addressed the practical consequences of autocorrelated errors in time-ordered insurance data[cite: 2]. The OLS estimates are unbiased but inefficient under serial correlation[cite: 2]. The AR(1) autocorrelation structure detected in the OLS residuals was explicitly modelled to construct the error covariance matrix, and the resulting GLS estimates provide a more efficient characterisation of the underlying relationships[cite: 2].

## REFERENCES

1. L. S. Aiken, *Multiple Regression: Testing and Interpreting Interactions*, 1991.
2. K. Albeladi et al., *Time Series Forecasting Using LSTM and ARIMA*, 2023.
3. G. Bhoi and B. Satpathy, *Assessing Claim Settlement Dynamics in Indian Life Insurance with Specific Reference to LIC of India Using Time Series and DEA*, 2025.

4. R. Bhouradia and N. Paraksh, *The Role of Insurance Ombudsman in Enhancing Consumer Protection in India*, 2024.
5. G. Brotosaputro et al., *Prediction of Claim Fund Reserves in Insurance Companies Using the ARIMA Method*, 2025.
6. C. Chatfield, *Analysis of Time Series An Introduction*, 5th edition.
7. A. Chaturvedi and S. Mishra, *Generalized Bayes Estimation of Spatial AR Model*, Statistics in Transition New Series **20** (2019), no. 2, 15–32.
8. C. Datta and J. Sen, *An Alternative Framework for Time Series Decomposition and Forecasting and its Relevance for Portfolio Choice: A Comparative Study of Indian Consumer Durable and Small Cap Sectors*, 2016.
9. C. Dimas et al., *Legal Protection for Insurance Companies Against Abuse of Insurance Claims*, 2025.
10. R. Ibrahimoh, *Using artificial intelligence to improve insurance claim evaluation*, 2025.
11. IRDAI, *Handbook on Indian Insurance Statistics 2007-08*, 2008.
12. IRDAI, *Handbook on Indian Insurance Statistics 2008-09*, 2009.
13. IRDAI, *Handbook on Indian Insurance Statistics 2009-10*, 2010.
14. IRDAI, *Handbook on Indian Insurance Statistics 2010-11*, 2011.
15. IRDAI, *Handbook on Indian Insurance Statistics 2011-12*, 2012.
16. IRDAI, *Handbook on Indian Insurance Statistics 2012-13*, 2013.
17. IRDAI, *Handbook on Indian Insurance Statistics 2013-14*, 2014.
18. IRDAI, *Handbook on Indian Insurance Statistics 2014-15*, 2015.
19. IRDAI, *Handbook on Indian Insurance Statistics 2015-16*, 2016.
20. IRDAI, *Handbook on Indian Insurance Statistics 2016-17*, 2017.
21. IRDAI, *Handbook on Indian Insurance Statistics 2017-18*, 2018.
22. IRDAI, *Handbook on Indian Insurance Statistics 2018-19*, 2019.
23. IRDAI, *Handbook on Indian Insurance Statistics 2019-20*, 2020.
24. IRDAI, *Handbook on Indian Insurance Statistics 2020-21*, 2021.
25. IRDAI, *Handbook on Indian Insurance Statistics 2021-22*, 2022.
26. IRDAI, *Handbook on Indian Insurance Statistics 2022-23*, 2023.
27. IRDAI, *Handbook on Indian Insurance Statistics 2023-24*, 2024.
28. IRDAI, *Industry Trends*, 2024.
29. G. James et al., *An Introduction to Statistical Learning, with applications in R*, 2021.
30. M. Jia et al., *Comparison Analysis of ARIMA and Machine Learning Methods for Predicting Trend of US Semiconductor Stocks*, 2022.
31. E. Munarsih and I. Sauza, *Comparison of Exponential Smoothing Method and Autoregressive Integrated Moving Average (ARIMA) Method in Predicting Dengue Fever Cases in City of Palembang*, 2020.
32. R. Nargunam and N. Rathee, *An Actuarial Perspective on Discrepancies in Health Insurance Pricing: Evidence from India*, 2025.
33. M. Sanyal, *Determinants of Health Insurance Choice and Ownership in India-Evidence from National Sample Survey Data*, 2025.
34. V. Selvakumar et al., *Modelling and Prediction of Third Party Claim Using a Machine Learning Approach*, 2020.
35. J. Sen, *Time Series Analysis-Based Forecasting Framework for the Indian Healthcare Sector*, 2017.
36. Z. Uulu et al., *AI Based Health Insurance Claim Forecasting*, 2022.
37. A. Witasari et al., *Model and Implication Insurance Claim Payment Failure in Indonesia*, 2025.
38. M. Yamada et al., *Epidemiology and Recurrence of Sigmoid Volvulus: Analysis of Health Insurance Claims Data in Japan*, 2025.

DEPARTMENT OF STATISTICS, SHAHEED RAJGURU COLLEGE OF APPLIED SCIENCES FOR WOMEN,  
UNIVERSITY OF DELHI, NEW DELHI, INDIA  
E-mail address: sandeep.mishra@rajguru.du.ac.in

DEPARTMENT OF STATISTICS, SHAHEED RAJGURU COLLEGE OF APPLIED SCIENCES FOR WOMEN,  
UNIVERSITY OF DELHI, NEW DELHI, INDIA  
E-mail address: saumya.govil29@gmail.com