Faculty of Natural and Applied Sciences Journal of Mathematical and Statistical Comput
Print ISSN: 3026-8605, e-ISSN 3043-632X
www.fnasjournals.com
Volume 2; Issue 1; September 2024; Page No. 26-38.

Developing a Robust Statistical Model for Predicting Mean Healthcare Costs across Different Sample Size Distributions in the Volta Region of Ghana

[*1]Todoko, C. A. K., [2]Ijomah, M. A., [2]Biu, E. O., [3]Abude, F. M., &. [4]Abledu, G. K.
[1]Department of Mathematics/Statistics, Ignatius Ajuru University of Education, Port Harcourt, Nigeria
[2]Department of Mathematics and Statistics, University of Port Harcourt, Nigeria
[3]Research Department, Bank of Ghana, Accra, Ghana
[4]Koforidua Technical University, Koforidua, Ghana

***Corresponding author email:** cosmosagbe@gmail.com

Abstract
Healthcare cost, particularly positively skewed cost data, modelling is an important area in health policy formulation as it provides policymakers with valuable information on the appropriate distribution, as well as the important covariates to use in cost minimization programmes. Previous studies have attempted to undertake this but with simulated data, inadequate sample size, or different distributions. This study aimed to determine the robustness of some statistical models based on large healthcare cost data. Using real-life healthcare cost data, the study sought to determine how the various statistical models performed with different sample sizes (n=1100 and n=2444). Data for the study was obtained from the Ghana Health Service's (GHS) facilities in the Volta Region of Ghana. We extracted the data from the District Health Information Management System 2 (DHIMS 2) database from 1st January to 31st December 2021 with covariates such as gender of the patients, age, length of stay in the hospital, and cost incurred. We explored both descriptive and inferential statistical techniques to analyze the data. Statistical models such as the ordinary least square (OLS), the OLS log (y), the log-normal (log(y)), the Poisson, the Cox proportional hazard, the Weibull, and the Gamma distributions were employed and the best model(s) were selected based on standard statistical metrics including the Akaike Information Criteria (AIC), the mean average percentage error (MAPE) and the mean squared error (MSE). The OLS log (y) was found to have outperformed all other models across different sample sizes. Policymakers could adopt the OLS log (y) model to predict healthcare costs in order to make a stronger case for adequate budgetary and logistic support.

Keywords: Healthcare Cost, Robustness, Statistical Models, and Estimators.

Introduction

Globally, statistical models have been adopted by policymakers and decision-makers in the health sector have adopted statistical models. This is the result of the validity and reliability provided by these statistical models to predict healthcare costs and expenditures in both developed and developing countries (Khosravi et al., 2024). Statistical models vary widely in terms of the data available for analysis. While one statistical model might prove a model fitting for a specific prediction other show lower signs of achieving model fits. Establishing a strong robust model fitting statistical model has proven to be very important for public health planning and sustainable healthcare delivery. Health care cost is the amount paid by clients or third-party payers for health care services or products. Over the years, healthcare costs and expenditures have been rising in developed and developing countries thereby receiving researchers' attention (Stanmore et al., 2019). The need to ascertain the variability of healthcare costs has been pondered by many health economists and policymakers. Tracking Universal Health Coverage (2023) stated that the recent COVID-19 pandemic has invigorated the argument for an efficient statistical model to predict healthcare costs and expenditures in the case of unforeseen and well-established activities. Countries and economies have to prepare to provide efficient well-measured healthcare delivery systems for their citizens since health is life and has an unmeasurable impact on Gross Domestic Product (GDP) and productivity.

Healthcare data comprises the cost and expenditures incurred in the sector within a specific timeframe (Almanie, 2024; Gold et al., 2022). It is a major hurdle for governments, and taxpayers and in various ways is instituted to

Cite this article as:
Todoko, C.A.K., Ijomah, M.A., Biu, E.O., Abude, F.M., &. Abledu, G.K. (2024). Developing a robust statistical model for predicting mean healthcare costs across different sample size distributions in the Volta region of Ghana. *FNAS Journal of Mathematical and Statistical Computing*, *2*(1), 26-38.

ensure that the cost is held stable or reduced. In many situations, the desirability of a reduced healthcare cost has not been achieved by many health economists (Ravangard et al., 2014). This is due to the variance of increase in healthcare delivery, infrastructure to support healthcare, personnel remuneration, infrastructure costs, and healthcare facilities incurring high medical cost bills, financial malfeasance, expenditure into non-productive related issues, and non-related healthcare delivery expenditure. Healthcare cost and expenditures forms a large majority of healthcare data, thus, taking into consideration health-related issues and diseases that are not taken care of by the National Health Insurance Scheme (NHIS), but not limited to per episode or lifetime costs of diseases, specific disease incident cases (Russel, 2004).

Healthcare costs may also come from unaccounted healthcare services and deliveries. Some researchers opined that healthcare cost response may change by level of consumption (e.g., outpatient versus inpatient, or low versus high levels) (Sturmberg & Bircher, 2019). Hence, some different parameters and factors affect the estimation of the mean population of healthcare costs. Thus, an efficient statistical model needs to be implemented to correct variations in costs and estimates in the short- or long-term for a sustainable healthcare delivery system. Given that cost and expenditure form a huge chunk of the healthcare data, there is a need to evaluate the efficiencies of the statistical models applicable for model-fitting predictions.

Skewed data are those data that create asymmetrical, or skewed curves on a graph. These data from healthcare costs and expenditures are noticeable from various factors and a few of them are patient's inability to continue with healthcare after defaulting payments which might be a result of death (Malehi et al., 2015). Another is due to few patients having peculiar ailments which shore up the cost of health delivery as against a large majority of patients who reported ailments with much smaller or negligible costs This inherently results in heavy right tails of the healthcare cost curve showing that there is disproportionality in the distribution curve. These factors have rendered the need to envisage and establish the most efficient strategies to calculate and predict healthcare costs and expenditures. Using standard statistical analysis to estimate the mean healthcare cost has been derailed by these characteristics. Hence, the conventionally acceptable strategy or analytics to solve highly positive skewed data is to adopt the use of the logarithmic transformation approach. This approach takes the log value of the dependent variables or covariates to match against the healthcare cost outcome variable (Khosravi et al., 2024). This is done primarily to achieve normal distribution of the data set and most importantly to reduce drastically the level of skewness in the data for achieving the reduction in positive skewness in healthcare data. There is a need to implement efficient statistical models to achieve predictability of healthcare data and expenditures through linear regression. Linear regression has the capacity for easy modelling. Moreover, they are easy to interpret the relationship between the dependent and independent variables, thus the healthcare cost and the covariates.

Regression modelling has been the most efficient statistical model used in predicting healthcare costs and expenditures, in addition, linear regression has been adopted as one of the critical tools used in prediction models of healthcare cost. With this analytical tool, there have been major challenges that have resulted not from the statistical models per se, but from attributes of healthcare data (Dash et al., 2019). Delving into the composition of healthcare data discloses the inefficiencies of using basic linear regression for prediction models. This is because healthcare data are characterized by high positive skewness and in some cases large number of resources with zero costs (Malehi et al., 2015). The Ordinary Least Squares (OLS) regression with logarithmic transformation has been adopted as the appropriate approach to solve the skewness in healthcare data. However, a few caveats have to be taken into consideration. The transformed scale data are not robust enough to give a strong model-fitting prediction of healthcare cost. The lack of robustness comes from the fact that it becomes imperatively difficult to determine heteroscedasticity within the variables. Hence, using just the logarithmic transformed OLS regression means having the propensity to lead to biased and imprecise estimates of the mean. Scholarly works in finding befitting statistical models for predicting healthcare costs have encountered challenges in estimating the population mean of healthcare costs. Also, the need to find a near-exact relationship between cost and covariates through regression modelling is another challenge in finding befitting statistical models for healthcare cost prediction.

One of the critical solutions to the problem not being solved by the logarithmic transformation approach is the adoption of General Linear Models (GLMs) which result from the Exponential Conditional Mean (ECM). This is an important framework because the analytical model in ECM considers non-normal distributions of the variables. Healthcare systems have adopted the use of log-link modelling of GLMs since the function truncates the weaknesses and problems of the OLS regression. Alternatively, other models from non-parametric and fully parametric modelling analytics have proved to have robust results that can solve the issues when dealing with healthcare data with all its accompanied attributes of positive skewed data and leptokurtosis. Robust model output and forecasts are consistently accurate.

Cite this article as:
Todoko, C.A.K., Ijomah, M.A., Biu, E.O., Abude, F.M., &. Abledu, G.K. (2024). Developing a robust statistical model for predicting mean healthcare costs across different sample size distributions in the Volta region of Ghana. *FNAS Journal of Mathematical and Statistical Computing*, *2*(1), 26-38.

Scholarly work underpinned the need for testing statistical models when dealing with healthcare data as well as the accompanied estimation of the population mean of the healthcare cost. (Malehi et al., 2015) opined that while the covariates are very important to understand the impact of estimation on healthcare delivery, there is also a need to look at the relationship between the covariates of predicting healthcare cost and expenditures. With the assertion of a biased normal distribution which has been adjusted and corrected over some time, there is the need for statistical models to be given enough robust tests to ascertain their unbiasedness and accurate predictability of the models implemented (Sendi et al., 2021).

Statement of the problem

Few studies have also established the strength of covariates within the estimators that affect healthcare costs. However, not enough studies have been conducted on the need to test the variables and parameters that have strong tendencies to impact the statistical techniques in achieving model fits in context to varied sample sizes. The need to test the statistical robustness in estimating the mean of healthcare cost as well as the relationships between the covariates which determines the impact of the factors affecting healthcare cost and expenditure therefore becomes crucial. The study adopts a systematic approach study to compare the various statistical models to achieve the goals of reducing skewness in data, limiting bias in estimators, and improving on precision of the statistical models best fit for specific sample sizes. One of the problems of the study has to do with the lack of studies concerning comparative analysis between statistical models with extensive literature such as OLS log, Cox proportional hazard models, and GLMs which have not been tested against various sample sizes to ascertain their efficiencies in biases elimination and improved precision. This study attempts to solve the problem for various health economists irrespective of the size of their sample size. Moreover, solving a knowledge gap problem that exists in adjudicating the best models for estimating means of healthcare costs based on varied sample sizes.

Objectives

This study aims to determine the robustness of some statistical models as a method of estimating the mean of healthcare costs with large sample data from the Volta Region of Ghana. The objectives of this study are to:

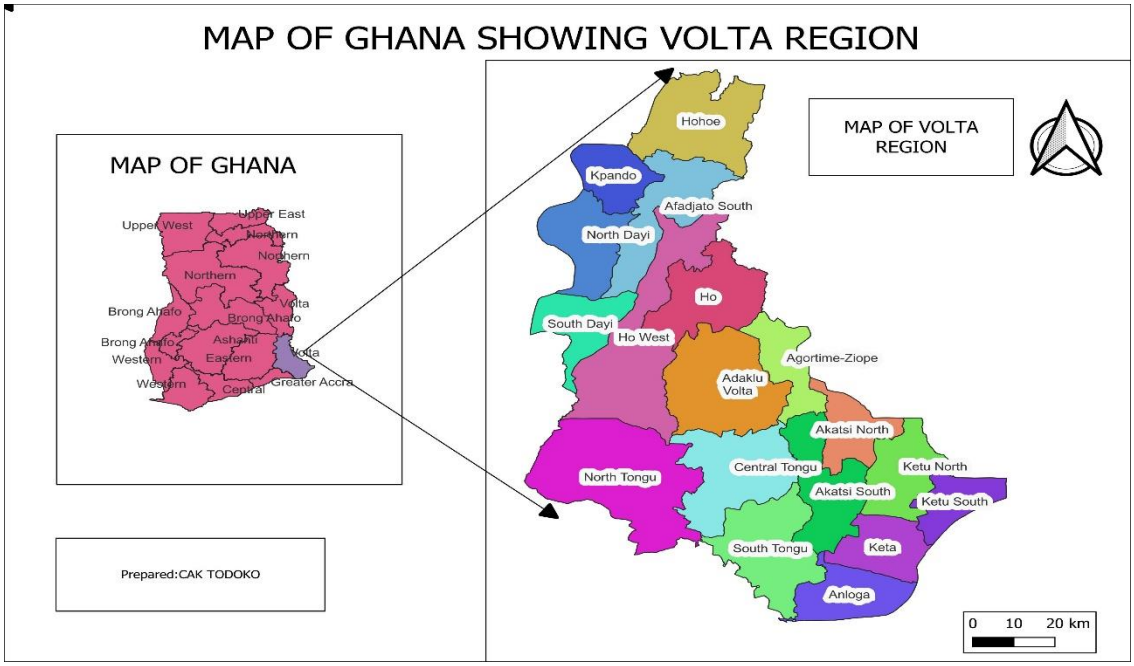
1. Determine the most appropriate estimator of skewed data in healthcare cost across various large sample sizes of 1100 and 2444.
2. Determine whether the estimated mean of the population or covariates effects is the best fit for statistical models.
3. Determine the impact of covariate effects on all conditions of the statistical models.
4. Identify the most efficient statistical models for estimating healthcare costs with varied sample sizes of 1100 and 2444.

Methods and Materials

The study employed a retrospective research approach by extracting secondary data (admission data) from DHIMS2. Second, the Event Report for 2021 and the client's covariates such as sex, age, length of stay, and number of time visits, were also downloaded. Those clients were selected with the hospital admission number traced to the hospital account units which provided the amount paid by each client. The real data of patients were gathered from hospitals and physicians' visits within one year. The total inpatient data for the year was 116220 observations on patients who at least visited a doctor in the respective hospitals in 12 months (January to December 2021) and the data obtained was 110000. The records of patients' healthcare cost data were grouped into different sample sizes. The study aims at having the power of 80 percent with a 95 percent confidence interval and considering that secondary data will be used for the research all patient morbidity and amount paid data entered into DHIM2 2021 was used for the study.

The study was conducted in the Volta Region of Ghana. The Volta region is among the sixteen regions of Ghana and has a population of 1649523 based on the 2020 population census with an annual growth rate of 2.1%. The region had a total of 18 which include districts/municipalities, 525 health facilities comprising: 322 CHPS, 39 Clinics, 24 hospitals, 120 health centres, 10 maternity homes, 4 polyclinics, a regional hospital, and a teaching hospital. Volta Region was selected for this research through simple random sampling techniques.

Cite this article as:
Todoko, C.A.K., Ijomah, M.A., Biu, E.O., Abude, F.M., &. Abledu, G.K. (2024). Developing a robust statistical model for predicting mean healthcare costs across different sample size distributions in the Volta region of Ghana. *FNAS Journal of Mathematical and Statistical Computing, 2*(1), 26-38.



Esthetical consideration was granted by Ghana Health Service with the number GHS:008/03/2023. The sample size determination for this work focuses on different variations. In this regard, the sample size calculation was evaluated taking into consideration varying precision levels of computations. The precision level could be structured into $\pm 2, \pm 3\%$. Thus, ensuring that the confidence level of 95% has been maintained across the sample size determination. The sample size was determined using Slovin's Formula with parameters of population size (N) and the margin of error (e). It is specified as follows:

$$n = \frac{N}{1 + N\,(e)^2}$$

N-population, n = sample size, e - precision level

The formula is universally accepted for determining the minimum sample size for scientific and health research. It gives a researcher an idea of how large the sample size needs to be to ensure a reasonable accuracy of results. To determine the sample size for the study, several scenarios were considered (simulation) based on the level of accuracy (level of precision) and the performance and robustness of the various distributions that were used in this study.

Model specification

Generalised Linear Models

The Generalised Linear Models (GLMs) are broad classes of statistical models that help with non-normal dependent variables to linear combinations of predictor variables. Taking into consideration Y_i which denotes healthcare expenditures for the person *i*, and X_i denoting the covariates which also goes a long way to include the intercepts. Using an invertible link function (g $(y_i.)$) included the expectation of the response variable *E* (Y_i) to the linear predictor which has been ascertained in equation (1) below.

$$g\big(E(y_i)\big) = g(\mu_i) = x_i\beta \qquad (1)$$

The ECM works with the log link function to produce a non–linear regression model as shown in equations (2a) and (2b) below:

$$\ln(E(y|x)) \;=\; x\beta \qquad (2a)$$

$$E(y) = \exp(x\beta) = \mu\,(x\beta) \qquad (2b)$$

One of the key attributes of these stochastic functions is the existence of the respective conditional mean-variance relationship. One of the main general structures for this function is;

$$Var\,(y) = \sigma^2 v(x) \qquad (3)$$

Cite this article as:
Todoko, C.A.K., Ijomah, M.A., Biu, E.O., Abude, F.M., &. Abledu, G.K. (2024). Developing a robust statistical model for predicting mean healthcare costs across different sample size distributions in the Volta region of Ghana. *FNAS Journal of Mathematical and Statistical Computing*, *2*(1), 26-38.

In relation to equation (2), the $v(x)$ represents $Var\ (y(x))$.
There is one function which has the attribute to the gamma structure with $v(x) = k_2(\mu(x))^2$, where $k_2 > 0$;
The standard deviation is proportional to the mean. Within this class of power–proportional variance functions, it is useful to think more generally of the variance function $v\ (x)$ being.

$$v(x) = k(\mu(x\beta))^{\lambda} \quad (4)$$

In the case of equation (4) the λ must be finite or non-negative. In situations where the $\lambda = 0$, then the position for the nonlinear least squares estimator is achieved.

Ordinary Least Squares (OLS) Models

Conventionally, the most adopted model is OLS-based model with logarithmic transformation of the dependent variable, $\ln(y)$. The log transformation was used to decrease the skewness in healthcare data. In the case of applying the ordinary least squares approach, the assumed regression model which was adopted to estimate the mean of healthcare cost data was.

$$y = \exp(x\beta) + \varepsilon \quad (5)$$

When the ε was taken to be homoscedastic, then there was the need to ascertain that the natural function $Var[y|x]$ was proportional to the estimated mean value of $E[y|x]^2$. Thus, just as the homoscedastic linear model in equation (6);

$$y = x\beta + \varepsilon \quad (6)$$

$$\ln(y) = x\delta + \varepsilon \quad (7)$$

For this model to achieve its robustness, it was assumed that $E(x\varepsilon) = 0$ and $E(\varepsilon) = 0$, in order to predict cost on the original scale of the data.
Moreover, in case the error term was heteroscedastic in x – meaning, the $E(exp(e))$ is some function $f(x) - then\ E(y) = f(x)\ x\ exp(x\delta)$, or, equivalently,

$$\ln\big(E(y)\big) = x\delta + \ln(f(x)) \quad (8)$$

In the log-normal case,

$$\ln\big(E(y)\big) = x\delta + 0.5\delta^2(x) \quad (9)$$

Where the last variance term is the error variance on the log scale.
These biases can be eliminated by including an estimate of the variance function, $v(x)$, if the error is log-normal, or more generally, of $E(\exp(e)|x)$.

$$y = \exp(x\beta + \varepsilon) \quad (10a)$$

$$E(y|x) = E\ (\exp(\varepsilon)\,|x)\ exp(x\beta) \quad (10b)$$

However, if the error term was normally distributed $N(0, \sigma_\varepsilon^2)$, then the log normal case is applied as shown below in 11.

$$E(y|x) = \exp(x\beta + 0.5\sigma_\varepsilon^2) \quad (11)$$

Cox Proportional Hazard Model

In simple terms, the hazard function could be expressed as the risk of dying at time t. which is estimated below.

$$h(t) = h_0(t)\ \times \exp(b_1x_1 + b_2x_2 + \cdots + b_px_p) \quad (12)$$

$$h(y|x) = h_0(y)\exp(x\beta) \quad (13)$$

Poisson Regression Model

Taking into consideration that the expected value of y_i is given by equation 14 below.

$$E\{y_i|x_i\} = \exp\{x_i^T\beta\} \quad (14)$$

Cite this article as:
Todoko, C.A.K., Ijomah, M.A., Biu, E.O., Abude, F.M., &. Abledu, G.K. (2024). Developing a robust statistical model for predicting mean healthcare costs across different sample size distributions in the Volta region of Ghana. *FNAS Journal of Mathematical and Statistical Computing*, *2*(1), 26-38.

In count data models which focus on estimating positive skewness data, one of the fundamental assumptions is that for a given x_i, the count variable y_i has a Poisson distribution with expectation $\lambda_i = \exp\{x_i^T \beta\}$. Thus, after ensuring log functions are applied the probability mass function of y_i conditional upon x_i is given by the equation below.

$$P\{y_i = y|x_i\} = \exp\{-\lambda_i\}\,\lambda_i^y |\, y!, y = 0, 1, 2, \ldots\ldots \quad (15)$$

Log-Normal Distribution Data

To start with, if the log scale error ε is normally distributed with $mean = 0$ and $variance\ v$, then the raw scale skewness (S) for this data generating mechanism is provided as.

$$S_{raw} = (w + 2)((w - 1)^2) \quad (16)$$

In reference to the equation 3.17, w = $exp(v)$. Using $N(0, v)$ normal distribution deviate, then let the log scale variance range from 0.5 to 2.0 in steps of 0.5 (i.e: 0.5, 1, 1.5, 2.0…). The true model thereof;

$$\ln(y) = \beta_0 + \beta_1 x + \varepsilon. \quad (17)$$

From 16 where x is uniform (0,1), e is N (0, v) with a variance of v = 0.5, 1.0, 1.5, or 2.0, Expected outcome $E(x`e) = 0$. B = 1.0. The value for the intercept B is selected so that $E(y) = 1$. However, for this data mechanism, the expectation of y is:

$$\mathrm{E(y)} = \mathrm{e}^{(\beta_0 + \beta_1 x + 0.5v)} \quad (18)$$

The slope of $E(y)$ with respect to x equals $\beta\ exp\ (\beta + \beta x + 0.5v)$.

Gamma Distribution Data

In a contest of the probability density function (pdf) of Gamma distribution, the model is ascertained or established as;

$$f(y) = \frac{1}{\Gamma(\alpha) b^{\alpha}} y^{\alpha - 1} e^{-y/b} \quad (19)$$

In equation (19) above, the α represents the shape parameter and b is the scale parameter. $b = \exp(\beta_0 + \beta_1 x)$ and α are the scale and shape parameters, respectively. The mean = αb and the skewness is a decreasing function of the shape parameter as $\frac{2}{\sqrt{\alpha}}$.

Weibull Distribution Data

The Weibull data predicted mechanism has some portion of the proportional hazard properties. One of the assumptions for the Weibull data predicted has been indicated below.

$$\mathrm{f(y)} = \frac{\alpha}{\mathrm{b}}\left(\frac{\mathrm{y}}{\mathrm{b}}\right)^{\alpha+1} \mathrm{e}^{(-\mathrm{y/b})^{\alpha}} \quad (20)$$

This equation means that $b = exp\ (\beta_0 + \beta_1 x)\ and\ \alpha$ represents the scale and shape parameters respectively. However, the mean; $b\Gamma(1 + \frac{1}{\alpha})$. The skewness is most importantly a decreasing function of the shape parameter, as depicted in equation (21).

$$b^3\Gamma\left(1 + \frac{3}{\alpha}\right) - 3\Gamma\left(1 + \frac{1}{\Gamma}\right)\Gamma\left(1 + \frac{2}{\alpha}\right) + 2\left(\Gamma\left(1 + \frac{1}{\alpha}\right)\right)^3 \quad (21)$$

The data-generating mechanism ascertains the shape parameter to be 0.5, 1 and 5 in the above equation.

Evaluating Statistical Model Performance

Two model performances were adopted for the research; Mean Prediction Error (MPE) and Mean Absolute Prediction Error (MAPE). The lower the values the better the model in estimating healthcare cost using either population mean or the covariates effects. The functions/equations have been ascertained in the equations below.

Cite this article as:
Todoko, C.A.K., Ijomah, M.A., Biu, E.O., Abude, F.M., &. Abledu, G.K. (2024). Developing a robust statistical model for predicting mean healthcare costs across different sample size distributions in the Volta region of Ghana. *FNAS Journal of Mathematical and Statistical Computing*, *2*(1), 26-38.

$$MPE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (22)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (23)$$

Akaike Information Criteria (AIC)

Akaike (1974) also developed another procedure known as the Akaike Information Criteria. Below is the form of the statistic.

$$\text{AIC}=\left(\frac{ESS}{T}\right)e\left(\frac{2k}{T}\right) \quad (24)$$

The value of AIC decreases when some variable is dropped (Ramanathan, 1995).

Results

Descriptive Statistics of the Covariate

Table 1: Sample size of (n) = 2444

Variables	Min.	Max.	Mean	Std. Dev.	Variance	Skewness
Hospital Visits	1	8	2.3200	1.1605	1.3471	1.9140
Age (years)	3	71	29.4301	15.6431	244.7010	0.3064
Length of treatment (days)	2	29	4.0700	3.5151	12.3531	4.1331
Sex	1	2	1.5000	0.5000	0.2503	0.0071

Source: Researchers' Computation, 2022

Data in Table 1 Shows the results of the sample size of n = 2444. The corresponding hospital visits had a mean (M=2.3200) day of approximately 2 days. SD =1.1605, variance = 1.340 and skewness of 1.9140. The age representation of the patients had a mean (M=29.43), indicating the mean or average age of participants was 29years, SD =15.641, and skewness of 0.306The length of treatment had a mean (M = 4.0700) thus they spent about 4days for treatment of their ailments, whiles the SD = 3.515, and skewness = 4.1331. The sex of patients had a mean (M =1.5000) with SD = 0.5000, variance = 0.2500, and skewness = 0.007

Table 2: Sample size of (n) = 1100

Variables	Min.	Max.	Mean	Std. Dev.	Variance	Skewness
Hospital Visits	1.0	8.0	2.3071	1.1745	1.3800	2.0460
Age (years)	3	71	29.3312	15.4801	239.6310	0.3061
Length of treatment (days)	2	29	3.9412	3.3251	11.0531	4.3031
Sex	1	2	1.5001	0.5001	0.2502	-0.0110

Source: Researchers' computation, 2022

Data in Table 2 shows the results of the sample size of n = 1100. The hospital visits had a mean (M =2.3071) which indicated on average patients visited the hospital 2 days with their ailment or for treatment the standard deviation (SD = 1.1745), variance = 1.3806 and skewness was established to be equal to 2.046. The age of the patients had a mean (M=29.3312) indicating an average of 29 years from the sample, the SD=15.480, variance = 239.631, and skewness of 0.3061) of the sample. This data is followed by the length of treatment in days, the mean (M =3.94) indicating that it takes an average of 4 days, the standard deviation (SD =3.325, the variance = 11.035, and skewness of 4.303.

Cite this article as:
Todoko, C.A.K., Ijomah, M.A., Biu, E.O., Abude, F.M., &. Abledu, G.K. (2024). Developing a robust statistical model for predicting mean healthcare costs across different sample size distributions in the Volta region of Ghana. *FNAS Journal of Mathematical and Statistical Computing*, *2*(1), 26-38.

Test of Normality on Raw Healthcare Cost

Table 3: Cost of Patient Healthcare (n = 2444)

Parameters	**Values (in Ghana Cedis)**
Mean	313.02
Std. Deviation	712.68
Variance	507915.85
Skewness	5.210
Kurtosis	28.29
Minimum	25.00
Maximum	5210.00

Table 3 shows that the different sample sizes used for the normality tests were 2444, 1100, 400, 100, and 25. About the sample size of n = 2444, Figure 4.1 with the corresponding table 4.8 showed that skewness = 5.213, with mean of health care cost M = 313.016. The maximum value was 5210 GH and the minimum was 25 GH. Figure 3 with the associated data showed a positively skewed healthcare cost data.

Table 4: Cost of Patient Healthcare (n = 1100)

Parameter	Value (in Ghana Cedis)
Mean	296.58
Std. Deviation	676.45
Variance	457586.00
Skewness	5.54
Kurtosis	32.31
Minimum	25.12
Maximum	5210

Table 4 shows the context of the sample size of 1100. The mean value of 296.58 indicates that, on average, patients spend GHS 296.58. Also, the skewness = 5.540 with a kurtosis of 32.317, which points to a highly positively skewed data test of normality on transformed data. In this study, the predicted data came from Weibull regression which was used to represent the Cox proportional hazard model.

Table 5: General Statistics on Predicted Data (n = 2444)

Parameters	Mean	Std. Dev	Coeff. of Skewness	Coeff. of Kurtosis
Log normal $\sigma^2 = 0.5$	0.978	0.7169	0.454	0.197
Log normal $\sigma^2 = 1$	1.0052	0.9978	0.433	-0.352
Log normal $\sigma^2 = 1.5$	0.9767	1.245	0.527	-0.514
Log normal $\sigma^2 = 2$	0.9824	1.4383	0.603	-0.582
Weibull $\alpha = 0.5$	1.07	2.73	1.1	0.07
Weibull $\alpha = 1$	1	1.01	1.85	0.37
Weibull $\alpha = 5$	1.01	0.23	-0.24	-0.2
Gamma $\alpha = 1$	1.02	1.01	2.09	7.11
Gamma $\alpha = 2$	0.99	0.71	1.44	2.85
Gamma $\alpha = 4$	1.18	0.59	1.08	1.68
Poisson $\alpha = 0.5$	0.5	0.71	1.46	2.52
Poisson $\alpha = 1$	1	0.98	0.89	0.51
Poisson $\alpha = 5$	2.53	1.59	0.69	0.6

Table 5 takes into consideration, the log normal, the higher the variance, the greater the skewness value across the sample sizes of 2444 and 1100. About the sample size of 2444, the Gamma skewness kept decreasing while the variance increased from $\alpha = 1$, 2, and 4. This was the same with the Weibull regression model which decreased in skewness while the shape kept increasing. Similarly, the Poisson regression showed that the coefficient of the skewness decreases while the variance increases from 0.5, 1, and 5. Thus, the variance reduced from 1.4600,

Cite this article as:
Todoko, C.A.K., Ijomah, M.A., Biu, E.O., Abude, F.M., &. Abledu, G.K. (2024). Developing a robust statistical model for predicting mean healthcare costs across different sample size distributions in the Volta region of Ghana. *FNAS Journal of Mathematical and Statistical Computing*, *2*(1), 26-38.

0.8900 and 0.6900. However, the Log normal skewness kept increasing with the variance. This meant that except for the log-normal, the predicted data of Weibull, Poisson, and Gamma all kept decreasing in skewness while their variance increased. In all, Weibull produces the highest skewness level.

Table 6: General Statistics on Predicted Data (n = 1100)

Parameters	**Mean**	**Std. Dev**	**Coeff. of Skewness**	**Coeff. of Kurtosis**
Log normal $\sigma^2 = 0.5$	0.978	0.7169	0.454	0.197
Log normal $\sigma^2 = 1$	1.0052	0.9978	0.433	-0.352
Log normal $\sigma^2 = 1.5$	0.9767	1.245	0.527	-0.514
Log normal $\sigma^2 = 2$	0.9824	1.4383	0.603	-0.582
Weibull $\alpha = 0.5$	1.07	2.73	1.1	0.07
Weibull $\alpha = 1$	1	1.01	1.85	0.37
Weibull $\alpha = 5$	1.01	0.23	-0.24	-0.2
Gamma $\alpha = 1$	1.02	1.01	2.09	7.11
Gamma $\alpha = 2$	0.99	0.71	1.44	2.85
Gamma $\alpha = 4$	1.18	0.59	1.08	1.68
Poisson $\alpha = 0.5$	0.5	0.71	1.46 00	2.52
Poisson $\alpha = 1$	1	0.98	0.89	0.51
Poisson $\alpha = 5$	2.53	1.59	0.69	0.6

Table 6 shows the sample size n = 1100, there were some similarities in terms of the skewness in both the Weibull, Gamma, and Poisson distributions of predicted data. Thus, at $\alpha = 0.5, 1, 5$ *of* the Weibull, the skewness coefficient decreased over time from 4.2600, 1.9800, and -0.2200 respectively, this was in line with that of the Gamma with $\alpha = 1, 2, 4$ showing a decreasing function of the skewness from 2.4, 1.37 and 1.0600, the Poisson distribution of predicted data also showed the same trend of decreasing skewness as against increasing variance or shapes. Except for the log-normal distribution which increased in value with the coefficient of skewness concerning the variance.

Table 7: Performance of estimators on a sample size of 2444

Model	**MPE**	**MAPE**	**MSE(β)**
Weibull			
Shape $\alpha = 0.5$	4.0071	0.7892	16.0611
Shape $\alpha = 1$	4.0782	0.803	16.6275
Shape $\alpha = 5$	4.068	0.801	16.5461
Gamma			
Shape $\alpha = 1$	4.0881	0.8991	16.464
Shape $\alpha = 2$	4.0581	0.805	16.7092
Shape $\alpha = 4$	3.8902	0.7676	15.1919
Log OLS	**MPE**	**MAPE**	
Variance $\sigma^2 = 0.5$	4.1001	0.8074	16.8074
Variance $\sigma^2 = 1$	4.0723	0.802	16.0168
Variance $\sigma^2 = 1.5$	4.1011	0.8076	16.815
Variance $\sigma^2 = 2$	4.095	0.8065	16.7714
Poisson			
$\alpha = 0.5$	4.8774	0.9015	15.0364
$\alpha = 1$	4.597	0.803	16.0943
$\alpha = 5$	4.5472	0.5017	16.4907
Cox Prop.			
Shape $\alpha = 0.5$	4.107	0.7892	16.0615
Shape $\alpha = 1$	4.3251	0.803	16.6275
Shape $\alpha = 5$	4.068	0.801	16.5461

Cite this article as:
Todoko, C.A.K., Ijomah, M.A., Biu, E.O., Abude, F.M., &. Abledu, G.K. (2024). Developing a robust statistical model for predicting mean healthcare costs across different sample size distributions in the Volta region of Ghana. *FNAS Journal of Mathematical and Statistical Computing*, *2*(1), 26-38.

Taking into consideration the data in Table 7, the estimators witness lower MPE by declining skewness and increasing sample sizes in relation to the results. Considering the sample size n = 2444, the OLS with 4.0723 exhibited lower MPE and levels compared to the Weibull, Gamma, Cox proportional and poison. Thus, producing a decreasing value of MPE while the shape kept increasing at $\sigma^2 = 1$, Weibull =4.0782, Gamma=4.0581, Cox proportional=4.0680 and Poisson = 4.5970 at $\alpha = 1$ for all respectively. However, there were lower values for MPE for estimators the better model Log OLS preferred to others Weibull, Gamma, Poisson, and Cox proportional hazard models. Under the MAPE which depicts an accuracy measure on each of the estimators, this was illustrated in table 7. The values of the MAPE were quite lower overall for all the estimators across the various sample sizes with minimal value variations. In the context of the sample size n=2444, some of the MAPE values were similar with little differences, thus, $\sigma^2 = 1$, Log OLS=0.802, and Weibull=0. 8030 Gamma=0.8991, Poisson=0.8030, and Cox proportional=0.8030 values respectively. The Log OLS had the least MAPE values across the GLM models compared to the Cox Proportional, Weibull Poisson, and Gamma.

As indicated in Table 7, with the focus of the objectives also considering the estimates of the covariates β_1 coefficients in estimating the model, the coefficients at 95% Confidence Interval (CI) were taken across sample sizes and estimators or models. The values showed that as classified within different sample sizes, there were greater similarities between the coefficient of the β_1 MSE values across the estimators or models at the sample size of n = 2444, the Poisson regression showed a decreasing MSE β_1 coefficient values at increasing shapes, thus β_1 = 15.0364 to 16.6275, to corresponding to shapes of α = 1, 0.5, and 5. However, the values across the estimators; Weibull, Cox prop, Gamma, and Log OLS have been ascertained to be slightly different with all being the high values of 15.01 – 17.90, which account for the higher values come from the results of the real world data.

Table 8: Performance of Estimators with Sample Size of 1100

Model	MPE	MAPE	MSE(β)
Weibull			
Shape α =0.5	4.1254	0.8160	17.0190
Shape α =1	3.9354	0.7785	15.4874
Shape α =5	3.955	0.7824	15.6452
Gamma			
Shape α =1	4.2454	0.8398	18.0235
Shape α =2	4.0354	0.7982	16.2845
Shape α =4	3.8654	0.7646	14.9414
Log OLS			
Variance σ^2=0.5	4.0497	0.801	16.3997
Variance σ^2=1	4.0810	0.8073	16.6555
Variance σ^2=1.5	4.0717	0.8054	16.5778
Variance σ^2=2	4.0732	0.8057	16.5909
Poisson			
Shape α =0.5	4.055	0.8022	16.4463
Shape α =1	4.575	0.9051	20.9343
Shape α =5	3.005	0.5945	9.0325
Cox Proportional			
Shape α =0.5	4.1541	0.865	16.3392
Shape α =1	4.3105	0.8635	16.4745
Shape α =5	4.4965	0.8824	16.6726

However, considering the Log OLS model, there was a depiction of decreasing MPE values of 4.0110 taking to $\sigma^2 = 1$ and an increase in α =1, of other models with a corresponding value. Taking a critical look at the sample size n = 1100, Weibull = 4.9350, Gamma = 4.2454, Poisson =4.5751 and Cox proportional = 4.9551 displayed decreasing MPE values at increasing shape of α = 1 The Log OLS of y had very close MPE values across the different variances, with only minimal variations. The Poisson and Cox Prop all had decreasing values of MPE at the increasing shapes. About the sample size of 1100, there was not much change in the MAPE values compared to the sample size of n = 2444. However, few variations were spotted for the Weibull = 0.8984 Gamma 0.8397 with Poisson = 0.9050 and Cox proportional = 0.8174 which saw an increase in values respectively the values

Cite this article as:
Todoko, C.A.K., Ijomah, M.A., Biu, E.O., Abude, F.M., &. Abledu, G.K. (2024). Developing a robust statistical model for predicting mean healthcare costs across different sample size distributions in the Volta region of Ghana. *FNAS Journal of Mathematical and Statistical Computing*, *2*(1), 26-38.

were reducing with rising shapes. The Log OLS values showed smaller MAPE compared to the other Poisson models thus the Cox Proportion, Weibull, and Gamma. Taking into consideration the sample size n = 1100, the Cox proportion provided a decreasing value of the coefficient of MSE as the shape kept increasing from α = 0.5, 1, and 5 corresponding to MSE = 17.02, 15.49, and 15.65. While that of the Log OLS was within the range of 16.3997 and 16.591 in an increasing order corresponding to the variance 0.5, 1, 1.5, 2. Comparatively, the Poisson had the lowest values as noted in the earlier sample size, while the Gamma and Weibull regression showed similar decreasing values with rising shapes.

Comparing Goodness of Fit

Table 9: Test of Goodness of Fit

Data	Estimator	Akaike Information Criterion (AIC)
Sample size n=2444	OLS for Ln(Y)	6944.6
	Poisson	6945.31
	Weibull	6944.8
	Gamma	14076.94
	Cox. Prop	7440.08
Sample size n= 1100	OLS for Ln(Y)	2264.87
	Poisson	2285.72
	Weibull	45240.1
	Gamma	2700.91
	Cox. Prop	5113.7

The results in Table 9 above show that at a sample size of n = 2444, the AIC values for all estimators were not widely different, however, the OLS for log(y) produced a robust goodness of fit with the minimum AIC = 6944.5950. At sample size n = 1100, the values were slightly wider in between than the previous sample size, the Log OLS had the best estimator with AIC = 2264.87. Conclusively, while, there are no best models across the various sample sizes, the OLS for log (y), was adjudicated as providing the best model for predicting the robustness and goodness of fit for healthcare data.

Discussion

Statistical modelling has become paramount in the estimation of health costs, A lot of strategies were adopted to control health costs by using Monte Carlo simulations was used to estimate the mean health cost of different models (Malehi et al., 2015). However, we used the real cost of patient data collected from health hospitals. The research of Malehi identified the Gamma distribution as an appropriate estimator of the mean, but we identified that as the sample size increases the OLS for Ln(y) improves significantly. This agrees with the statement by many researchers that "selecting the optimal models depends on the research objectives". We again identified that Cox proportional hazard models, despite their theoretical merits, may not come out with the best results most importantly in non-ideal data conditions. We found many researchers who stated that there are no universally best model across all the data conditions using both simulation and real patient cost data. Our findings suggest that healthcare cost estimation sometimes involves the transfer from one health system to another involves an assumption about production process and efficiency since the data sources vary. It is also consistent with research work which suggested that in modeling healthcare cost sensitivity to assumption should be consider (Gregori, 2011).

Depending on the sample size and estimator models, the effects of the covariates used produce higher precision in estimating healthcare costs. With different sample sizes, the use of the mean of the healthcare cost data provided robust results as compared to the covariates and simulation. Also, they were easy to read and interpret without much ambiguity. Taking into consideration the various results shown in the above tables MPE with sample size Gamma performed better than all the distributions this goes contrary to research by (Gregori et al., 2011). However, the Gamma regression model had the smallest biases across all data-generating processes. From Table 7, poison distribution performs better than all the distributions since it provides less value. The mean population estimates across the various sample sizes and variance where necessary and the estimators in comparison with the various results with the unique variance or shape (shape α =1) or Variance=1, the OLS performed better than all other distributions. The MPE for the poison distribution was better than all other distributions also MAPE poison

Cite this article as:
Todoko, C.A.K., Ijomah, M.A., Biu, E.O., Abude, F.M., &. Abledu, G.K. (2024). Developing a robust statistical model for predicting mean healthcare costs across different sample size distributions in the Volta region of Ghana. *FNAS Journal of Mathematical and Statistical Computing*, *2*(1), 26-38.

again performed better than ordinary least square (OLS), log(y), Poisson, Cox proportional hazard, Weibull, and Gamma distributions.

Comparatively from the results under the unique variance or shape (shape α=1 or variance=1) of the distributions, the OLS, regression model provided more accurate estimates of inpatient cost than is the mean. However, the OLS produces higher robustness at higher, and the Poisson differs slightly in all the shapes and the variances. This means that the OLS performed sufficiently better on real-world data and we disagree with a research work with stated that, The Weibull estimator produces much higher robust precision with lower MSE of the covariates and that of the population mean results. Thus, the GLM models were robust compared to the Cox proportional hazard model and OLS for log (y). One cannot choose an estimator over another across all the sample sizes; however, the OLS (Log normal) produces a much better prediction, this is confirmed in the Table as OLS produced less value compared to the Gamma, Poisson, and Cox proportional hazard models. However, in this study, the predicted data came from Weibull regression which was used to represent the Cox proportional hazard model.

Conclusion

In this study, we cannot conclude that one estimator is better across various sample sizes and variances and shapes, few of them estimated the mean healthcare cost more appropriately comparatively. Thus, irrespective of the shapes and variances, some produced better estimators using the population mean and covariates. Thus, lower values on MAPE, MSES, and MPE. The OLS log (y) model is robust when using real-world data. In the context of the Akaike Information Criterion (AIC), the best model was adjudicated to be those with lower values, and the OLS for log (y) and Weibull performed well at higher sample sizes. In many situations, the OLS log (y) becomes a dominant estimator without much substantial when dealing with the estimation of mean population cost E(y) and the covariates β_1 of the healthcare cost data.

Recommendations

The following recommendations were made based on the findings of the study:

1. The choice of a statistical model to predict healthcare costs should be determined by the size of data involved in the study. Using the different sample sizes used in this study as a guide will be helpful.
2. Policymakers, particularly in Ghana, could adopt the OLS log (y) to predict healthcare costs irrespective of the sample size involved. This is because it produced robust and accurate estimates across different sample sizes.

References

Almanie, S. A. (2024). *Assessment of Direct Medical Cost of Hospitalized COVID-19 Adult Patients in Kuwait During the First Wave of the Pandemic*. *July*, 509–522.

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0217-0

Gold, H. T., McDermott, C., Hoomans, T., & Wagner, T. H. (2022). Cost data in implementation science: categories and approaches to costing. *Implementation Science*, *17*(1), 1–12. https://doi.org/10.1186/s13012-021-01172-6

Gregori, D., Petrinco, M., Bo, S., Desideri, A., Merletti, F., & Pagano, E. (2011). Regression models for analyzing costs and their determinants in health care: An introductory review. *International Journal for Quality in Health Care*, *23*(3), 331–341. https://doi.org/10.1093/intqhc/mzr010

Khosravi, M., Zare, Z., Mojtabaeian, S. M., & Izadi, R. (2024). Artificial Intelligence and Decision-Making in Healthcare: A Thematic Analysis of a Systematic Review of Reviews. *Health Services Research and Managerial Epidemiology*, *11*. https://doi.org/10.1177/23333928241234863

Malehi, A. S., Pourmotahari, F., & Angali, K. A. (2015). Statistical models for the analysis of skewed healthcare cost data : a simulation study. ??? https://doi.org/10.1186/s13561-015-0045-7

Ravangard, R., Hatam, N., Teimourizad, A., & Jafari, A. (2014). Factors affecting the technical efficiency of health systems: A case study of economic cooperation organization (ECO) countries (2004–10). *International Journal of Health Policy and Management*, *3*(2), 63–69. https://doi.org/10.15171/ijhpm.2014.60

Russel, S. (2004). The economic burden of illness for households in developing countries: A review of studies focusing on malaria, tuberculosis, and human immunodeficiency virus/acquired immunodeficiency syndrome. *American Journal of Tropical Medicine and Hygiene*, *71*(2 SUPPL.), 147–155. https://doi.org/10.4269/ajtmh.2004.71.147

Sendi, P., Matter-Walstra, K., & Schwenkglenks, M. (2021). Handling uncertainty in cost-effectiveness analysis: Budget impact and risk aversion. *Healthcare (Switzerland)*, *9*(11). https://doi.org/10.3390/healthcare9111419

Stanmore, E. K., Mavroeidi, A., De Jong, L. D., Skelton, D. A., Sutton, C. J., Benedetto, V., Munford, L. A., Meekes, W., Bell, V., & Todd, C. (2019). The effectiveness and cost-effectiveness of strength and balance Exergames to reduce falls risk for people aged 55 years and older in UK assisted living facilities: A multi-centre, cluster randomised controlled trial. *BMC Medicine*, *17*(1), 1–14. https://doi.org/10.1186/s12916-019-1278-9

Sturmberg, J. P., & Bircher, J. (2019). Better and fulfilling healthcare at lower costs: The need to manage health systems as complex adaptive systems. *F1000Research*, *8*, 1–13. https://doi.org/10.12688/F1000RESEARCH.19414.1