World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:18, No:4, 2024

Identifying Factors Contributing to the Spread of Lyme Disease: A Regression Analysis of Virginia's Data

Fatemeh Valizadeh Gamchi, Edward L. Boone

Open Science Index, Mathematical and Computational Sciences Vol:18, No:4, 2024 publications.waset.org/10013586.pdf

Abstract—This research focuses on Lyme disease, a widespread infectious condition in the United States caused by the bacterium *Borrelia burgdorferi* sensu stricto. It is critical to identify environmental and economic elements that are contributing to the spread of the disease. This study examined data from Virginia to identify a subset of explanatory variables significant for Lyme disease case numbers. To identify relevant variables and avoid overfitting, linear poisson, and regularization regression methods such as ridge, lasso, and elastic net penalty were employed. Cross-validation was performed to acquire tuning parameters. The methods proposed can automatically identify relevant disease count covariates. The efficacy of the techniques was assessed using four criteria on three simulated datasets. Finally, using the Virginia Department of Health's Lyme disease dataset, the study successfully identified key factors, and the results were consistent with previous studies.

Keywords—Lyme disease, Poisson generalized linear model, Ridge regression, Lasso Regression, elastic net regression.

I. Introduction

LYME disease is one of the most often reported vector-borne diseases in the US. Since it was first identified in 1975 in the town of Old Lyme, Connecticut, the illness has been referred to as Lyme disease. Lyme disease, which is carried by the blacklegged tick (*Ixodes scapularis*) in the eastern United States and is caused by the bacterium *Borrelia burgdorferi*, is the most common vector-borne illness in North America [1]. Incidence of Lyme disease grew nationally between 1992 and 2002, although the total number of confirmed cases has subsequently remained mostly constant [1]. Recently, Lyme disease has become more common in some areas; in Virginia, the number of verified cases nearly doubled between 2006 and 2007 [1]. A 1990 study on Lyme disease cases in Virginia found that while the illness was not common in the early 1980s, it seemed to have increased in incidence and geographic distribution by the late 1980s, which led researchers to conclude that the condition was spreading south [2]. Virginia is therefore a great location to study the illness's mechanism and discover the essential factors that contribute to the formation of Lyme disease. The bacteria is spread to humans through the bites of Ixodid species ticks. Many people do not realize they have been bitten since blacklegged tick nymphs are tiny, difficult to notice, and do not cause any itching or irritation. Early signs of Lyme disease include a rash, fever, headaches, and fatigue. In those who are not treated when the disease is still in its early stages, severe and persistent symptoms may appear. Memory issues, shooting pains, numbness in the hands or feet, and arthritis in the main joints are some of the enduring symptoms of this disease. According to Maes et al. [3], the disease's anticipated treatment costs impose a major burden on the public's health. Therefore, research on the origin of Lyme disease is important for public health in general. Tick abundance, host species populations and infection rates, human population patterns, awareness, and behavior, habitat, climate, and other variables all have an impact on the spread of Lyme disease. Understanding the long-term effects of environmental and social factors that may play a role in the development of lyme disease is possible via research on the disease.

The major goal of this study is to determine a subset of explanatory variables that are important for Lyme disease case numbers in northern and western Virginia. Bivariate analysis and other straightforward techniques have been utilized in earlier studies to find important factors [4]–[6].

We opted to focus on Virginia's Northern Piedmont, Blue Ridge, Ridge and Valley, and Central Appalachian regions for our study. These regions are crucial for the research of Lyme disease because they are home to a diversity of tick species, habitats, deer, and potentially at risk individuals. In order to identify which environmental factors in northern and western Virginia were most strongly associated with Lyme disease, we employed Poisson generalized linear model, Ridge regression, lasso regression, and Elastic net penalty.

II. Methods

There are experiments having counts as their potential outcomes in many real-world issues in several fields, including engineering, medicine, biology, economics, and the sciences. The frequency of an event occurring over a certain period of time or location might be the target variable in any of these areas. Examples are number of failures, number of errors, etc. Here, we are looking to determine the relationship between a count response variable and the regressors. Poisson model is the best linear model to model count or rate data [7]. In this model, the response variable y is assumed to have a Poisson distribution.

Poisson generalized linear models (GLMs) are a form of regression model for modeling count data. They are

Fatemeh Valizadeh Gamchi is PhD candidate with Department of Statistical Sciences and Operations Research, Virginia commonwealth University, Richmond, VA, 23284 USA (e-mail: valizadehf@vcu.edu).

Edward L Boone is Professor with Department of Statistical Sciences and Operations Research, Virginia commonwealth University, Richmond, VA, 23284 USA.

Open Science Index, Mathematical and Computational Sciences Vol:18, No:4, 2024 publications.waset.org/10013586.pdf

a broader form of Poisson regression, which implies the response variable has a Poisson distribution. Poisson GLMs may simulate a broader range of count data, such as under- or over-dispersed counts, zero-inflated or zero-truncated counts, and correlated counts. Other factors, or covariates, that may impact the count outcome can also be included [7]–[11].

A. Regularized Regression

The compexity of a model is defined by the number of its parameters and their values. More parameters increases the risk of overfitting. Overfitting occurs when the estimation is a good fit for the particular dataset and may fail to fit another dataset or to predict future observation. Regulization is a way to avoid overfitting by constraining the coefficient estimation to zero. The size of the coefficient and the error term are penalized. Some of the simple approaches to reduce model comlexity and avoid overfitting are to add a Ridge, Lasso or Elastic Net penalty to the model [12], [13].

Ridge regression addresses overfitting by adding a penalty term to the objective function of the regression model, which shrinks the coefficients of the predictors towards zero while still allowing all variables to contribute to the model. The penalty term is proportional to the square of the magnitude of the coefficients, and it can be controlled by a hyperparameter called lambda or alpha [14].

Poisson ridge regression is a type of ridge regression that is specifically designed for count data. It assumes that the response variable follows a Poisson distribution. The goal of Poisson ridge regression is to find a set of coefficients that minimizes the sum of squared errors while controlling for overfitting [15]. Also it tries to solve problems caused by multicollinearity due to a linear association among regressors (explanatory variables) [16], [10]. Multicollinearity may present some issues in the regression model such as having unstable coefficients, large variance, prediction can be very poor, poor power of tests and yet the fit of the regression model may be good. Ridge regression is a biased estimation technique that tries to reduce the effect of the collinearity by obtaining a decrease in variance and an increase in the stability of the regression coefficients [15].

Lasso regression stand for least absolute shrinkage and selection operator introduced by Tibshirani [17] is a type of regression analysis that is used to both prevent overfitting and select variables. Lasso regression imposes the L_1-norm on the regression coefficient which means the sum of the absolute value of the coefficients is restricted. Lasso regression adds a penalty term to the objective function of the regression model that shrinks the coefficients of the variables that are not important, effectively performing variable selection and creating a more parsimonious model. The penalty term is proportional to the absolute value of the coefficients, and it can set some of them exactly to zero. Lasso regression is particularly useful when the number of predictors is large and some of them may be irrelevant or redundant [13], [18]. Poisson lasso regression is a type of Lasso regression analysis and is particularly useful when dealing with count data, which is often more prone to overfitting than continuous data. Poisson lasso regression effectively controls overfitting by selecting only the most significant predictors, which provides a more interpretable model with better predictive performance [18]–[24].

Poisson elastic net regression is a count data modeling approach that combines the advantages of ridge and Lasso regression. The L1-norm and L2-norm penalties are linearly combined in Poisson elastic net regression. This enables it to choose a restricted set of predictors while simultaneously accounting for predictor correlations [25], [12], [10].

III. Simulation

In the simulation study we consider the following model:

$$y_i \sim Poisson(\lambda(x_i)) \tag{1}$$

where $\lambda(x_i)$ is the mean and variance of the response variables based on Poisson distribution given as $exp(\beta_0 + x_i\beta)$. We ran three different settings for our simulation, each of them with 100 observations and 15 explanatory variables. In addition, $\beta_0 = 0.2$ is the intercept and $\beta = (-0.5, -0.5, -0.5, -0.5, -0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0, 0, 0, 0, 0)$ represents the vector of regression coefficients. We note that, some of the coefficients are equal to zero which shows that the covariates corresponding to these coefficients are not active in the model.

In the first scenario, the explanatory variable is a 15-dimension vector following a multivariate normal distribution with mean zero and covariance I_{15}. In the second scenario, the vector of explanatory variables follows a multivariate normal distribution with mean zero and block diagonal covariance matrix where $cov(x)_{ij} = (0.7)^{|i-j|}$ for $i, j = 1, 2, ..., 5$, showing strong colinearity between first 5 covariates, and $cov(x)_{ij} = 1$ for $i = j$ and 0 for the rest of covariates. In the third scenario, the explanatory covariates follow a multivariate normal distribution with mean zero and strong colinearity between all of them with a covariance matrix where $cov(x)_{ij} = (0.7)^{|i-j|}$ for $i, j = 1, 2, ..., 15$. The estimation for the coefficients for the first simulated dataset are shown in the Table II for the Linear regression (LM), GLM, Ridge, Lasso and Elastic net GLM. Also, MSE, Cross-Validation(CV), R-square and AIC criteria are calculated.

The goal of these simulations is to compare the performance of LM, GLM, Ridge GLM, Lasso GLM and Elastic net GLM, together. We consider four different criteria to measure the performance of these models which are: (a) MSE: the mean squared error based on response residuals; (b) CV: the 10-folder cross validation based on deviance residuals; (c) R^2 shows the percentage of the total variance explained by the model and is defined to be 1- (sum of squares of residuals/ total sum of squares); (d) AIC is the Akaike information criterion defined as $-2 * log - likelihood + p * n$. In addition, at the end in Table.VII we consider three measures to variable selection accuracy: (a) avre.size: it indicates the average size of the models; (b) corr.coef: it shows the average number of coefficients set to 0 correctly; (c) mis.coef: it calculates the average number of coefficients set to 0 incorrectly.

Open Science Index, Mathematical and Computational Sciences Vol:18, No:4, 2024 publications.waset.org/10013586.pdf

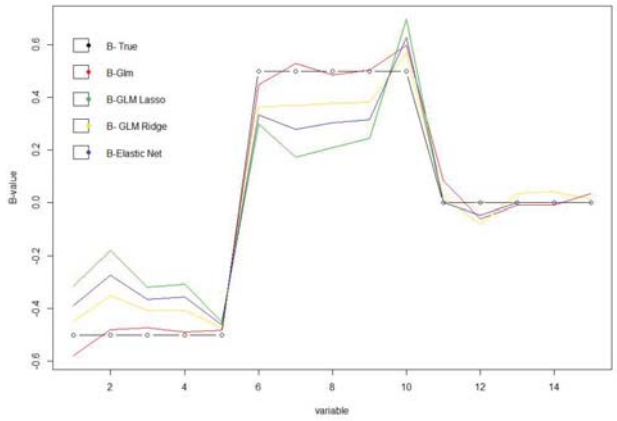

Fig. 1 Comparison of estimated value for each β using GLM, Ridge, Lasso and Elastic net regression based on first simulated dataset

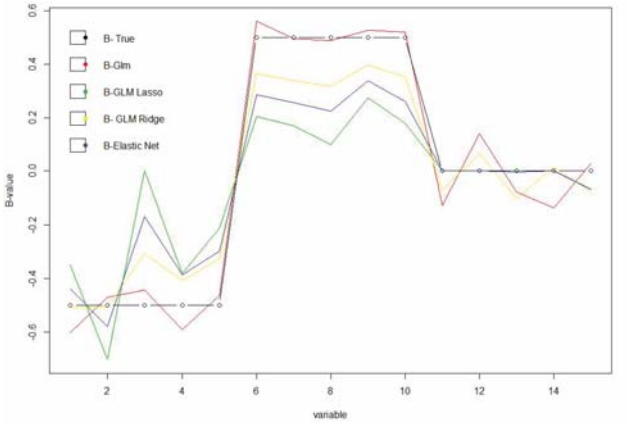

Fig. 2 Comparison of estimated values for each β using GLMs, Ridge, Lasso and Elastic net regression based on second the simulated dataset

We used the glmnet package [26] in R to fit Ridge, Lasso and Elastic GLM net. the Mass [27] for generating correlated predictors. In GLM, goodness-of-fit statistics only use for summarizing how well models could fit data. We utilized the cross validation (CV) technique to find the optimal tuning parameter for each simulation in Table I. We calculated optimal value for the tuning parameter through CV by considering deviance type of measure for Poisson regression which result is in Table I.

By having considered the optimal values in Table I, we were able to fit the discussed models to the dataset and represent, in Table II, MSE, R^2, CV and AIC of each model.

Table II presents the results of the first simulation with independent covariates. The GLM method outperformed the other methods, producing the highest R-squared value of 0.987, the lowest MSE and AIC, indicating a better fit. Lasso and Elastic net regression methods provided better results than Ridge regression in terms of cross-validation. Furthermore, Lasso and Elastic net were able to select relevant predictors by setting some coefficients to zero. The results suggest that GLM provides accurate coefficient estimation. Fig. 1 visually compares the estimates of the different models, and shows that the estimation under GLM is the closest to the true value of β.

Now, we repeat the same steps for the the second simulated dataset where there is high colinearity between the first five explanatory variables. Fig. 2 provides the plots for the second simulated dataset corresponding to those in Table III that visually show the optimum value for the tuning parameter. Also, the optimal tuning parameter for the second simulated data considering CV with deviance measurement are represented in Table IV.

In Table IV it is clear that the GLM model has the lowest MSE, highest R-squared value, and lowest AIC value among all the models, indicating it is the best-performing model. However, its CV score is higher than some of the other models, indicates that the model's predictions are less accurate. The LM model is the worst-performing, while the Ridge, Lasso, and Elastic-net models perform reasonably well. Also the estimation value for the coefficients are close to the real values

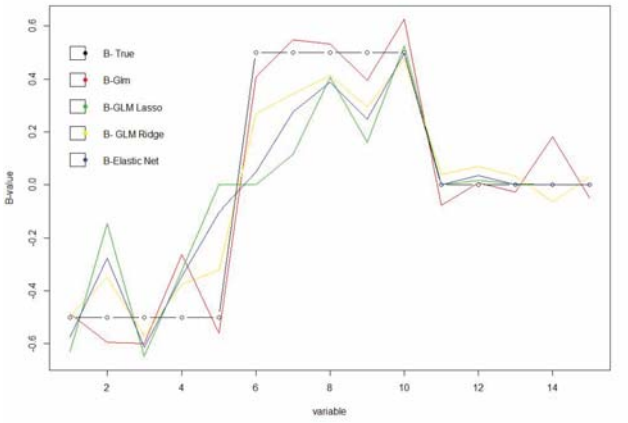

Fig. 3 Comparison of estimated values for each β using GLM, Ridge, Lasso and Elastic net regression based on the third simulated dataset

as shown Fig. 2.

The third simulated dataset has a strong colinearity between all the covariates. The optimal tuning parameters following CV method with deviance measurement are in Table V.

Table VI shows that the GLM model has the lowest MSE and highest R-square value, indicating that it has the best fit among the models. The GLM model also has the highest CV value, indicating that it has higher variability in its predictions.The LM model has the highest MSE and lowest R-square value, indicating that it has the worst fit among the models. The Ridg model has a moderate MSE and R-square value, and a low CV value. The Lasso model has a high MSE and low R-square value, but a low CV value. The Elastic-net model has a moderate MSE and R-square value, and a low CV value. It should be noted that both the Lasso and Elastic net models performed variable selection by setting some of the coefficients to zero. Also the estimated values of coefficients are shown in Fig. 3.

In Table VII, three performance measures are evaluated for accuracy of variable selection: average size (avre.size) of the models, correct zero estimate count (corr.coef), and incorrect zero estimate count (mis.coef). In the first model,

TABLE I
OPTIMAL TUNING PARAMETERS FOR FIRST DATASET BASED ON CV

	$Ridge_{glm}$	$Lasso_{glm}$	$Elasticnet_{glm}$
$Tunning-parameter$	0.5584067	0.01320347	0.02640693

TABLE II
MODEL PERFORMANCE BASED ON FIRST SIMULATED DATASET

Model	MSE	CV	R-square	AIC
Lm	103.49	6.8	0.44	781.74
GLm	2.51	16.7	0.98	309.32
$Ridge_{glm}$	5.28	2.20	0.97	
$Lasso_{glm}$	19.22	1.33	0.89	394.42
$Elastic\ net_{glm}$	10.26	1.32	0.94	

the avre.size of coefficients incorrectly set to zero is notably small, and this increases as the correlation of the coefficients in the models rises. This trend is consistent across the other two measures for each model. Elastic Net demonstrates commendable performance in avre.size value and marginally outperforms in minimizing mis.coef, while Lasso regression excels in accurately identifying corr.coef, showcasing its superior function in choosing correct coefficients to be zeroed.

A. Data Analysis

The dataset for this article on Lyme disease which was taken from [28] includes information on cases from 2006 to 2011, as well as statistics on Virginia's population and land use. The Virginia Department of Health gathered statistics on cases of Lyme disease (2006–2011). The 2010 Census provided the demographic information, such as population density, median income, and average age [29]. Data on land cover were gathered for 2006 from the Multi-Resolution Land Cover Consortium [30]. In this paper, we utilized the data for Eco id = 0, which stands for the northern/western subregion, which includes the Northern Piedmont, Blue Ridge, Ridge and Valley, and Central Appalachian. Xie et al. [28] used this dataset to identify important environmental and human factors for the spread of this disease. This dataset is available in the supplement of [28]. As a first step in developing a model, we took into account environmental and demographic variables that may have an impact on the development of the disease. There are 15 variables in this dataset. The description is as follows [28].

- x1: Percentage of developed land in each census tract
- x2: Percentage of forest in each census tract
- x3: Percentage of herbaceous in each census tract
- x5: Sum of area of forested fragments in each census tract divided by the total area
- x6: Sum of forest fragment perimeters in each census tract divided by the total area
- x7: CWED of developed-forest edge
- x8: TECI of developed-forest edge
- x9: CWED of forest-herbaceous edge
- x10: TECI of forest-herbaceous edge
- x11: CWED of herbaceous-developed edge
- x12: TECI of herbaceous-developed edge

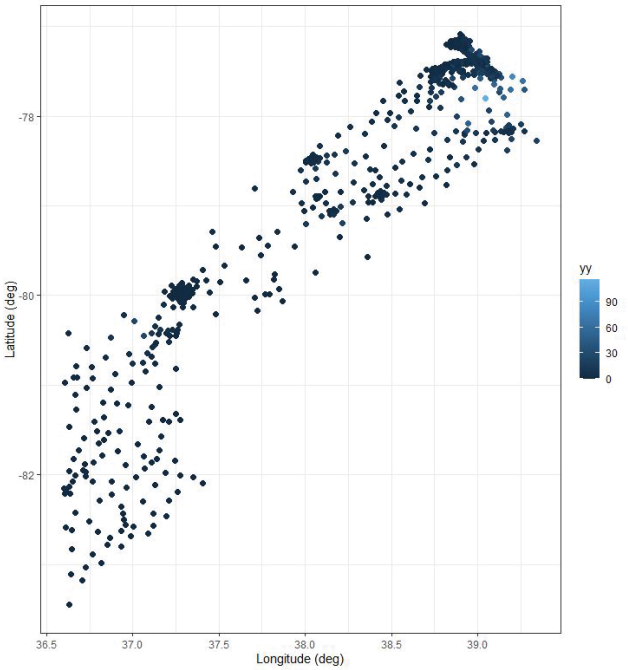

Fig. 4 The response variable y in the Lyme disease dataset

- x13: Tract population density in 2010
- x14: Median_age; Median age at each census tract in 2010
- x15: Mean_income; Mean income (inflation adjusted) at each census tract in 2010

Also, Fig. 4 represents the response variable in spatial locations for region 0 in the Lyme disease data set. Fig. 5 summarizes the distribution of the response variable, the count of Lyme disease in different locations for the study area.

As we can see in Fig. 5 plots, there are many locations with zero cases which suggests that using zero-inflated Poisson regression in future might be a good idea to model this dataset. Table VIII shows summary statistics for the response variable. It is worth noting that, the mean of the data is 6.346, and the standard deviation is 10.54087. The Poisson log linear regression model for the expected rate of the disease cases at location i is $log\frac{\mu_i}{m_i} = \beta_0 + \beta^T x_i$ or $log\mu_i = \beta_0 + \beta^T x_i + log(m_i)$ where m_i is the population for location i and $-log(m_i)$ is the offset term. To continue, we analysed this dataset by LM, GLM, Ridge, GLM, Lasso GLM and Elastic net GLM. To calculate the estimation of coefficients, we need to find the optimal tuning parameter for Ridge, Lasso and Elastic net penalty. Fig. 6 is the plot of coefficient vs log of tuning parameter to find the minimum value for tunning parameter by having MSE as a type measure.

Table IX indicates these optimal tuning parameter by CV

Open Science Index, Mathematical and Computational Sciences Vol:18, No:4, 2024 publications.waset.org/10013586.pdf

TABLE III
OPTIMAL TUNING PARAMETERS FOR SECOND SIMULATED DATASET BASED ON CV

	$Ridge_{glm}$	$Lasso_{glm}$	$Elasticnet_{glm}$
$Tunning - parameter$	1.913204	0.02358688	0.035685223

TABLE IV
MODEL PERFORMANCE BASED ON SECOND SIMULATED DATASET

Model	MSE	CV	R-square	AIC
Lm	726.824	14.6	0.40	976.65
GLm	2.088506	26.9	0.99	307.43
$Ridge_{glm}$	36.32	5.63	0.97	
$Lasso_{glm}$	169.42	0.98	0.86	698.68
$Elastic\ net_{glm}$	90.34	1.28	0.92	

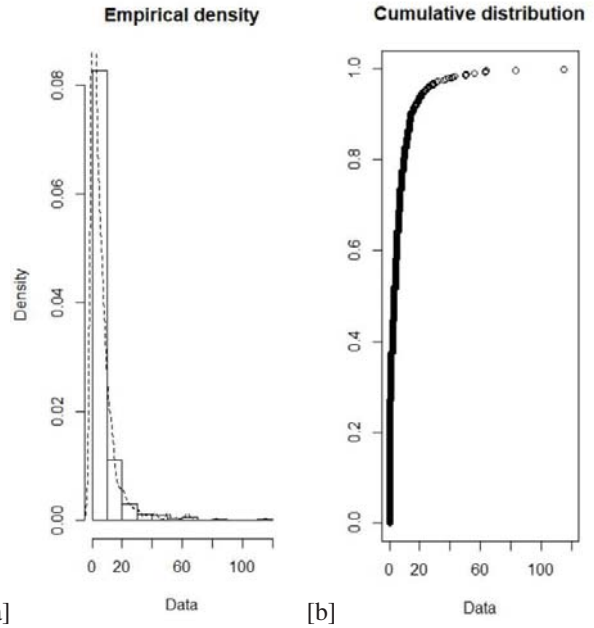

Fig. 5 (a) Probability density function (pdf) y for the Poisson distribution with mean μ and plot (b) Cumulative distribution function (cdf) of y

method based on deviance measurement in Lyme disease dataset.

Result for the estimation of parameters are represented in Table X. The asterisks in this table indicate the level of significance of the estimated coefficients, with more asterisks indicating higher significance. Based on the evaluation metrics, GLM appears to have the best performance in estimating unknown parameters, with the lowest MSE value and the highest R-square value. However, it has a high CV score, indicating that it may be over fitting the data. LM appears to have the worst model performance among the evaluated models. Lasso, Ridge, and Elastic net appear to be reasonable models with relatively low MSE values. Among these models, elastic net has the best performance, with a relatively low MSE and CV score and a high R-squared value. Additionally, elastic net performed variable selection by assigning zero coefficients to some of the predictors, which can be beneficial for model interpretability and reducing over fitting.

Based on the analysis, the results indicate that several variables have a significant role in the occurrence of Lyme disease cases. Specifically, the variables that have the most significant role are the Percentage of forest in each census tract, Percentage of herbaceous in each census tract, Sum of forest fragment, Sum of forest fragment perimeters in each census tract divided by the total area, perimeters in each census tract divided by the total area, CWED of forest-herbaceous edge, TECI of herbaceous-developed edge, Median age at each census tract in 2010, and Mean income at each census tract in 2010. These variables had a higher coefficient estimate compared to other variables, indicating that they have a stronger relationship with the occurrence of Lyme disease cases. Most of the finding are consistent with the literature for this subregion [28]. Herbaceous environments can offer deer and mice a suitable place to live. White- footed mice or deer have been demonstrated to be highly significant tick hosts in earlier investigations [28], [6]. For certain host species, the presence of both forest and herbaceous regions is attractive. For instance, deer always stay close to forest edge. Consequently, there may be a relationship between the occurrence of Lyme disease and the interspersion of forest and herbaceous land. The outcome is in line with [28]. The percentage of forest cover is significant. It was also discovered to be a crucial factor in literature [28], [6]. In line with the research on Lyme disease, it was also found that the mean income was an active variable. People with lower incomes may be less likely to have health insurance, live in areas with more green space, participate in outdoor activities, or know how to prevent Lyme disease. Furthermore, higher income may be correlated with better health status and access to healthcare, which could result in earlier diagnosis and treatment of Lyme disease. [31], [28], [6]. The median age was an active variable in our results on Lyme disease because older individuals may have a weakened immune system or be more likely to be exposed to ticks. It also found in [32]. Fig. 7 shows the visualization of these estimations.

IV. CONCLUSION

The study focused on identifying the environmental and economic factors contributing to the spread of Lyme disease in Virginia. Linear Poisson and regularization regression methods were used to identify relevant variables and avoid over fitting.

We simulated three different datasets using multivariate normal distribution. The first dataset served as a baseline, while in the second dataset, we introduced strong colinearity between the first five covariates. In the third scenario, there was strong colinearity between all of the covariates. We applied the mentioned methods to these simulated datasets and evaluated their performance using metrics such as MSE, CV, R-square, and AIC. We then applied these approaches to the Virginia Department of Health's Lyme disease dataset to

Open Science Index, Mathematical and Computational Sciences Vol:18, No:4, 2024 publications.waset.org/10013586.pdf

TABLE V
OPTIMAL TUNNING PARAMETERS FOR THIRD SIMULATED DATASET

	$Ridge_{glm}$	$Lasso_{glm}$	$Elasticnet_{glm}$
$Tunning-parameter$	2.484119	0.03688838	0.05580942

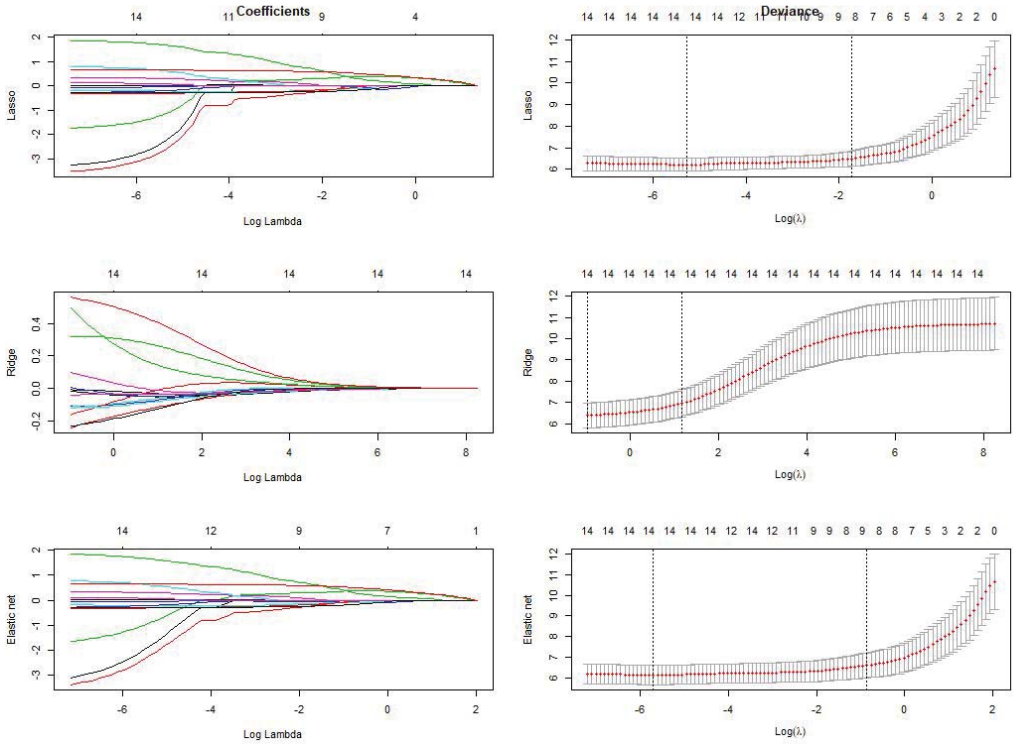

Fig. 6 , Lasso and Elastic net Poisson regression using cross-validation based on Lyme disease dataset

TABLE VI
MODEL PERFORMANCE BASED ON THIRD SIMULATED DATASET

Model	**MSE**	**CV**	**R-square**	**AIC**
Lm	2924.98	24.5	0.32	1115.89
GLm	4.82	90.7	0.99	366.33
$Ridge_{glm}$	52.79	12.02	0.98	AIC
$Lasso_{glm}$	225.05	2.92	0.94	706.51
$Elastic\ net_{glm}$	130.20	1.86	0.96	AIC

TABLE VII
COEFFICIENT SELECTION ACCURACY ACROSS MODELS, HIGHLIGHTING DIFFERENCES BETWEEN ELASTIC NET AND LASSO REGRESSION

	$Lasso_{glm}$	$Elasticnet_{glm}$	True Value
$Model1$			
$aver.size$	13.6	13.9	15
$mis.coef$	0.26	0.25	10
$corr.coef$	1.18	0.836	5
$Model2$			
$aver.size$	12.3	12.7	15
$mis.coef$	1	0.99	10
$corr.coef$	1.71	1.34	5
$Model3$			
$aver.size$	11.9	12.3	15
$mis.coef$	1.03	1.02	10
$corr.coef$	2.04	1.67	5

identify the environmental factors responsible for the spread of Lyme disease in the northern/western area of Virginia. The study found that the percentage of forest cover, percentage of herbaceous land, forest fragment and perimeter, median age, and mean income are the most significant factors in the occurrence of Lyme disease. These findings are consistent with previous studies, which have shown that Lyme disease is more common in areas with more forest cover, more herbaceous land, and older populations. The study's findings can be used to develop strategies for preventing the spread of Lyme disease.

For future work, as the dataset has more zeros, it may be beneficial to utilize zero-inflated Poisson regression to model these additional zeros independently. A zero-inflated model was fitted and the Vuong test was conducted to compare it with the saturated Poisson regression model. The test statistic in this dataset and a significant p-value indicated that the zero-inflated model would be a good option.

Open Science Index, Mathematical and Computational Sciences Vol:18, No:4, 2024 publications.waset.org/10013586.pdf

TABLE VIII
SUMMARY STATISTICS OF THE RESPONSE VARIABLE y FROM THE LYME DISEASE DATASET

Min	$1stQuantile(25\%)$	Median (50 %)	$Mean$	$3rdQuantile(75\%)$	Max
0	0	3	6.346	8	115

TABLE IX
OPTIMAL TUNING PARAMETERS FOR LYME DIEASE DATASET BASED ON CV

	$Ridge_{glm}$	$Lasso_{glm}$	$Elasticnet_{glm}$
$Tunning-parameter$	0.3819814	0.00991251	0.01245069

TABLE X
PARAMETER ESTIMATION FOR POISSON REGRESSION MODEL BASED ON THE LYME DISEASE DATASET BY ADDING OFFSET

Method	LM	GLM	$Ridge_{glm}$	$Lasso_{glm}$	$Elasticnet_{glm}$
β_0	28.44^{***}	3.035^{***}	3.09	3.15	3.12
β_1	-51.66	-2.20^{***}	-0.04	.	.
β_2	-67.86	-2.43^{***}	-0.24	.	-0.04
β_3	-24.62	-1.05^{***}	0.29	0.39	0.37
β_5	7.10	0.16	-0.05	.	-0.01
β_6	-12.72	-0.39^{*}	-0.09	-0.11	-0.09
β_7	-0.31	0.31^{***}	0.11	.	.
β_8	2.82	$-0.14^{\cdot}$	-0.07	.	.
β_9	-11.32^{***}	-0.35^{***}	-0.22	-0.04	-0.15
β_{10}	29.24^{***}	0.77^{***}	0.30	.	0.12
β_{11}	0.90	0.08	-0.00	.	.
β_{12}	-8.68	-0.26^{*}	-0.29	-0.22	-0.25
β_{13}	-4.96	-0.10^{*}	-0.07	.	.
β_{14}	-5.38^{*}	-0.21^{***}	-0.16	-0.03	-0.09
β_{15}	25.94^{***}	0.63^{***}	0.56	0.458	0.51
MSE	1584.14	58.75	62.61	70.53	66.52
CV	8.3	81.7	5.50	5.49	5.50
$R-square$	0.33	0.47	0.43	0.36	0.40
AIC	5981.90	4542.84	AIC	20321.67	AIC

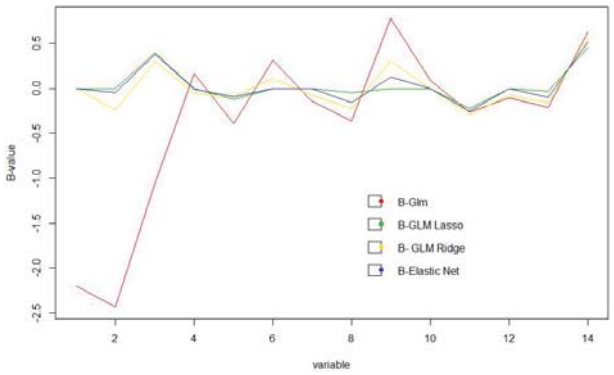

Fig. 7 Comparison of estimated values for each β using GLM, Ridge, Lasso and Elastic net regression based on the Lyme Diease dataset

REFERENCES

[1] R. Murphree Bacon, K. J. Kugeler, and P. S. Mead, "Surveillance for lyme disease–united states, 1992-2006," 2008.
[2] R. J. Brinkerhoff, W. F. Gilliam, and D. Gaines, "Lyme disease, virginia, usa, 2000–2011," *Emerging infectious diseases*, vol. 20, no. 10, p. 1661, 2014.
[3] E. Maes, P. Lecomte, and N. Ray, "A cost-of-illness study of lyme disease in the united states," *Clinical therapeutics*, vol. 20, no. 5, pp. 993–1008, 1998.
[4] B. F. Allan, F. Keesing, and R. S. Ostfeld, "Effect of forest fragmentation on lyme disease risk," *Conservation Biology*, vol. 17, no. 1, pp. 267–272, 2003.
[5] F. Valizadeh Gamchi, Ö. Gürünlü Alma, and R. Arabi Belaghi, "Classical and bayesian inference for burr type-iii distribution based on progressive type-ii hybrid censored data," *Mathematical Sciences*, vol. 13, pp. 79–95, 2019.
[6] L. E. Jackson, E. D. Hilborn, and J. C. Thomas, "Towards landscape design guidelines for reducing lyme disease risk," *International journal of epidemiology*, vol. 35, no. 2, pp. 315–322, 2006.
[7] P. Consul and F. Famoye, "Generalized poisson regression model," *Communications in Statistics-Theory and Methods*, vol. 21, no. 1, pp. 89–109, 1992.
[8] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.
[9] A. F. Zuur, E. N. Ieno, N. Walker, A. A. Saveliev, G. M. Smith, A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith, "Zero-truncated and zero-inflated models for count data," *Mixed effects models and extensions in ecology with R*, pp. 261–293, 2009.
[10] A. Agresti, "An introduction to categorical data analysis," 1996.
[11] G. Rodrıguez, "Models for count data with overdispersion," *Addendum to the WWS*, vol. 509, 2013.
[12] J. Mwikali, S. Mwalili, and A. Wanjoya, "Penalized poisson regression model using elastic net and least absolute shrinkage and selection operator (lasso) penality," *Int. J. Data Sci. Anal*, vol. 5, no. 5, pp. 99–103, 2019.
[13] C. Flexeder, "Generalized lasso regularization for regression models," Ph.D. dissertation, Institut für Statistik, 2010.
[14] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 42, no. 1, pp. 80–86, 2000.
[15] K. Månsson and G. Shukur, "A poisson ridge regression estimator," *Economic Modelling*, vol. 28, no. 4, pp. 1475–1481, 2011.
[16] R. H. Myers and R. H. Myers, *Classical and modern regression with applications*. Duxbury press Belmont, CA, 1990, vol. 2.
[17] R. Tibshirani, "Regression shrinkage and selection via the lasso,"

Open Science Index, Mathematical and Computational Sciences Vol:18, No:4, 2024 publications.waset.org/10013586.pdf

Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996.
[18] M. Y. Park and T. Hastie, "L1-regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 4, pp. 659–677, 2007.
[19] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization," *Mathematical Programming*, vol. 117, pp. 387–423, 2009.
[20] Z. Qin, K. Scheinberg, and D. Goldfarb, "Efficient block-coordinate descent algorithms for the group lasso," *Mathematical Programming Computation*, vol. 5, no. 2, pp. 143–169, 2013.
[21] R. Arabi Belaghi, F. Valizadeh Gamchi, and H. Bevrani, "Likelihood based inference on progressive type-ii hybrid-censored data for burr type iii distribution," *Reliability Theory and its Applications*, p. 194, 2016.
[22] T. T. Wu and K. Lange, "Coordinate descent algorithms for lasso penalized regression," 2008.
[23] L. Kantorovitch, "The method of successive approximation for functional equations," *Acta Mathematica*, vol. 71, no. 1, pp. 63–97, 1939.
[24] S. Hossain and E. Ahmed, "Shrinkage and penalty estimators of a poisson regression model," *Australian & New Zealand Journal of Statistics*, vol. 54, no. 3, pp. 359–373, 2012.
[25] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.
[26] T. Hastie and J. Qian, "Glmnet vignette," *Retrieved June*, vol. 9, no. 2016, pp. 1–30, 2014.
[27] B. Ripley, B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, D. Firth, and M. B. Ripley, "Package 'mass'," *Cran r*, vol. 538, pp. 113–120, 2013.
[28] Y. Xie, L. Xu, J. Li, X. Deng, Y. Hong, K. Kolivras, and D. N. Gaines, "Spatial variable selection and an application to virginia lyme disease emergence," *Journal of the American statistical association*, vol. 114, no. 528, pp. 1466–1480, 2019.
[29] Z. W. Almquist, "Us census spatial and demographic data in r: the uscensus2000 suite of packages," *Journal of Statistical Software*, vol. 37, pp. 1–31, 2010.
[30] J. A. Fry, G. Xian, S. Jin, J. A. Dewitz, C. G. Homer, L. Yang, C. A. Barnes, N. D. Herold, J. D. Wickham *et al.*, "Completion of the 2006 national land cover database for the conterminous united states." *PE&RS, Photogrammetric Engineering & Remote Sensing*, vol. 77, no. 9, pp. 858–864, 2011.
[31] S. E. Seukep, K. N. Kolivras, Y. Hong, J. Li, S. P. Prisley, J. B. Campbell, D. N. Gaines, and R. L. Dymond, "An examination of the demographic and environmental variables correlated with lyme disease emergence in virginia," *Ecohealth*, vol. 12, pp. 634–644, 2015.
[32] H. J. Kilpatrick and A. M. LaBonte, *Managing urban deer in Connecticut: a guide for residents and communities*. Connecticut Department of Environmental Protection, Bureau of Natural . . . , 2007.

Open Science Index, Mathematical and Computational Sciences Vol:18, No:4, 2024 publications.waset.org/10013586.pdf