Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 2197-2202
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.6092

Atan regularized for the high dimensional Poisson regression model

Ali Hameed Yousif[a,*], Ahlam Hanash Gatea[b]

[a] *College of Administration and Economic, Wasit University, Iraq*
[b] *College of Languages, University of Baghdad, Iraq*

(Communicated by Madjid Eshaghi Gordji)

Abstract

Variable selection in Poisson regression with high dimensional data has been widely used in recent years. we proposed in this paper using a penalty function that depends on a function named a penalty. An Atan estimator was compared with Lasso and adaptive lasso. A simulation and application show that an Atan estimator has the advantage in the estimation of coefficient and variables selection.

Keywords: Poisson regression, Lasso, Adaptive Lasso, Atan

1. Introduction

Poisson regression is one of the statistical models which has been widely used in recent years and in different fields, such as medicine, engineering, and the social sciences [4]. Poisson regression is an appropriate technique in analyzing count data and it studies the relationship between the mean of count data and explanatory variables [8]. If there is a relation between the explanatory variables this will lead to multicollinearity Increase in set of the explanatory variables will lead to large variance with difficult interpretation [3]. The classical methods cannot deal with these problems so we had to look for new methods that can deal with them The commonly used method of dealing with dimensional and multicollinearity problems is a ridge regression that was proposed by [7]. Although ridge regression has certain properties, it cannot make variable selection. [12] proposed Lasso regression that can perform estimation and variable selection. [6] mentioned that a good penalty function gives an estimator that has three properties, including Unbiasedness, Sparsity and,

[*] Corresponding author

Email addresses: ahameed@uowasit.edu.iq (Ali Hameed Yousif), ahameed@uowasit.edu.iq (Ahlam Hanash Gatea)

Received: October 2021 *Accepted:* December 2021

Continuity. [10] introduced an algorithm for compute the estimators in generalized linear models. [2] focused on the identification of outliers in the Poisson regression model. [3] did a modification on adaptive Lasso to get an adjusted Lasso. [17] studied the performance of estimators for Poisson regression models with highly correlated variables. [9] proposed using the elastic net method for Poisson regression model in state estimation and variable selection. [15] proposed using LAD-Atan estimator in the regression model with high dimensional data. [16] proposed using a new penalty Atan for the quantile Regression with high dimensional data. [14] proposed robust LAD-Shrink set estimator in regression model with high dimensional data. In this paper, we proposed using a new a penalty functions in the Poisson regression with high dimensional data named Atan. This paper is organized as follows. In the second part includes the maximum likelihood method for poisson regression. The third part includes the regularized poisson regression with Lasso, Adaptive Lasso estimators. The fourth part presents the Atan poisson regression with high dimensional data. The fifth and sixth parts illustrate simulation results and application. Finally, the seventh part is devoted to presenting research conclusions.

2. Maxumim Likelihood Method for Poisson Regression

Let y_i represents a number of events according to the Poisson distribution with means μ_i where the probability function is shown in the following formula:

$$f(y_i|x_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}; \quad y_i = 0, 1, 2, ... \tag{2.1}$$

Let $\mu_i = exp(x_i'\beta)$ represents the conditional mean in Poisson regression method. Let $y_1, y_2, ..., y_n$ are observations according to Poisson distribution with mean μ_i, then the likelihood function for observations are shown in the following formula:

$$L(y_i, \mu_i) = \frac{(\prod_{i=1}^n \mu_i^{y_i})(e^{-\sum_{i=1}^n \mu_i})}{\prod_{i=1}^n y_i} \tag{2.2}$$

when we taking log of the above equation, we get:

$$l(\beta) = \sum_{i=1}^n y_i \log(\mu_i) - \sum_{i=1}^n \mu_i - \log\left(\prod_{i=1}^n y_i!\right) \tag{2.3}$$

but $\mu_i = exp(x_i'\beta)$ then

$$l(\beta) = \sum_{i=1}^n y_i x_i'\beta - \sum_{i=1}^n exp(x_i'\beta) - \log\left\{\prod_{i=1}^n y_i!\right\} \tag{2.4}$$

By performing the partial differentiation of equation (2.4) with respect to β and using the weighted iterative least squares algorithm, the maximum likelihood estimator for β is as follows [4]:

$$\hat{\beta}_{ML} = (\acute{X}\hat{W}X)^{-1}(\acute{X}\hat{W}\hat{z}) \tag{2.5}$$

Where the $\hat{W}$ represent a matrix diagonal with diag (μ_i) and $\hat{z}$ is a vector equal to

$$\hat{z} = \log(\mu_\imath) + \frac{y_i - \hat{\mu_\imath}}{\hat{\mu_\imath}}$$

3. Regularized Poisson Regression

The classical methods for estimating parameters of the Poisson regression model is not suitable deal with high dimensional data. The appropriate method to deal with high dimensional data is called the regularized method which has been widely used to overcome high dimensional problems in the Poisson regression model and to improve the accuracy of prediction and we can get the estimator by minimizing the following [3, 11].

$$\hat{\beta} = \arg\min_{\beta}\{-l(\beta) + P_{\lambda}(\beta)\} \tag{3.1}$$

Where $P_{\lambda}(.)$ represents the penalty function, where there are many penalty functions used in Poisson regression. One poplar penalty called Lasso was proposed by [12], the idea of Lasso put restricted to coefficients is equal to the sum absolute of coefficients. The lasso estimator becomes one of the estimators that is widely used in the analysis of regression models because it makes estimation and selection of variable. The term of penalty for the lasso is as follows.

$$P_{\lambda}(|\beta|) = \sum_{j=1}^{p} |\beta_j| \tag{3.2}$$

Where λ is a penalty parameter, then lasso for poisson regression model

$$\hat{\beta}_{\text{Lasso}} = \arg\min_{\beta}\left\{-l(\beta) + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{3.3}$$

[18] proposed the unbiased penalty called Adaptive Lasso, that uses different weights for the coefficients. The Adaptive Lasso leads to reduce bias in the process of selecting variables and thus determines the appropriate model. The Adaptive Penalty Function LASSO came after [12] who suggested a penalty function called LASSO, in which Zou proved that LASSO Penalty Function is a biased function as a result of its dependence on a single weight for all coefficients. The Adaptive LASSO penalty is shown in the following formula

$$P_{\lambda}(|\beta|) = \sum_{j=1}^{p} w_j\,|\beta_j| \tag{3.4}$$

Where w_j represents a vector with px1 and compute by the following formula.

$$w_j = [abs\,(\beta_j)]^{\gamma} \tag{3.5}$$

And γ is positive constant equal and $\gamma > 0$

[3] used adaptive lasso for poisson regression model

$$\hat{\beta}_{\textbf{A}\text{Lasso}} = \arg\min_{\beta}\left\{-l(\beta) + \lambda\sum_{j=1}^{p}w_j\,|\beta_j|\right\} \tag{3.6}$$

4. Atan Poisson Regression Method

In this section, we proposed using Atan penalty in Poisson Regression to get estimation and variable selection with the same time. Atan penalty has been proposed by [13] to include the following formula:

$$P_{\lambda,\alpha}(|\beta|) = \lambda \left(\alpha + \frac{2}{\pi}\right) \arctan\left(\frac{|\beta|}{\alpha}\right); \quad \alpha = 0.005 \tag{4.1}$$

Then the estimator of Poisson Regression with using Atan penalty can be computed as follows :

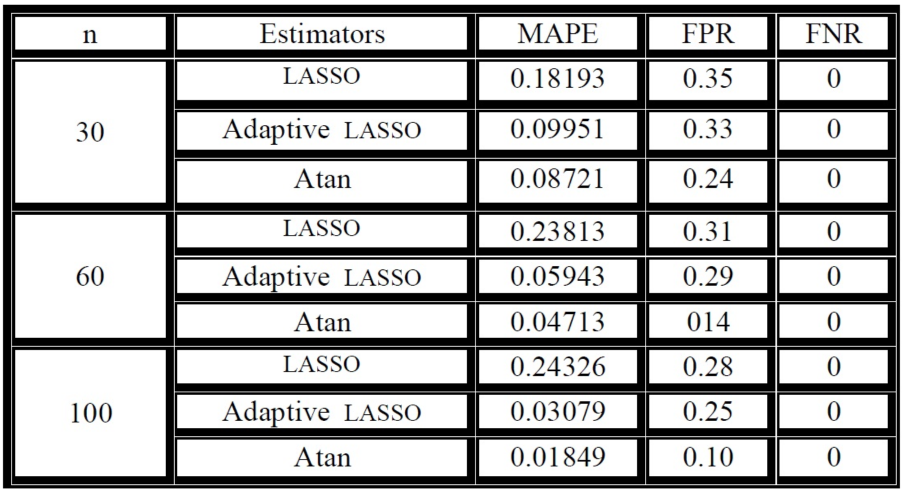
$$\hat{\beta}_{\text{Atan}} = \arg\min_{\beta}\{-l(\beta) + \lambda P_{\lambda,\alpha}(|\beta|)\} \tag{4.2}$$

5. Simulation

In order to compare the behavior of the regularized estimators of the poisson regression with high-dimensional data, simulations were used by the Monte Carlo method. In simulation experiments, the response variable is generated through a Poisson distribution with mean $\mu_i = exp(x_i'\beta)$. The explanatory variables are generated by $x \sim N_p(0, \sum)$, where $\sum_{ij} = \rho^{|i-j|}$, where $\rho = 0.5$. The simulation experiments were repeated 200 times. The sample sizes are 30, 60, and 100. The default values for the parameters are as follows:
First experiment: $\beta = (3, 1.5, 0, 0, 2, 0, ..., 0),\ p = 10$.
Second experiment: $\beta = (1.5, 0.5, 0, 1, 0, 0, 1.5, 0, 0, 0, 1, 0, ..., 0),\ p = 15$.

In the simulation study, we relied on a mean square error (MSE) as criteria for comparison among the estimators who also used two other two criteria for measuring variable selection: (FNR) and (FPR) were proposed by [1].

Table 1: simulation results of the first experiment for all estimators



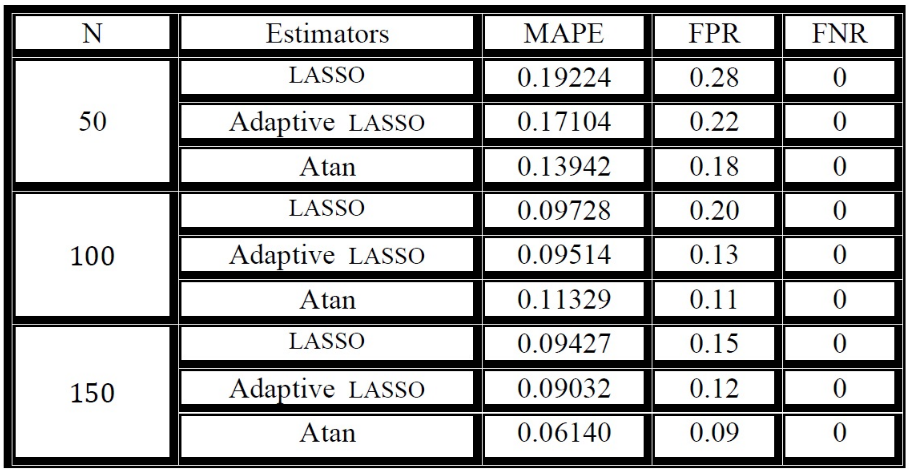
n	Estimators	MAPE	FPR	FNR
30	LASSO	0.18193	0.35	0
	Adaptive LASSO	0.09951	0.33	0
	Atan	0.08721	0.24	0
60	LASSO	0.23813	0.31	0
	Adaptive LASSO	0.05943	0.29	0
	Atan	0.04713	014	0
100	LASSO	0.24326	0.28	0
	Adaptive LASSO	0.03079	0.25	0
	Atan	0.01849	0.10	0

From Table 1 and 2, We note that the simulation results were as follows. For the first experiment $p = 10$ and $n = 50, 100, 150$ respectively and the second experiment $p = 15$ and $n = 50, 100, 150$ respectively, the Atan estimator has excellent performance in estimation and variable selection because it has the lower values of MSE, FNR and FPR respectively. While the Adaptive Lasso estimator came in the second rank in estimation and variable selection., le the Lasso estimator came in the last rank in estimation and variable selection.

6. Application

In this section, we used real data to represent a study of prostate cancer for patients based on a study in [5].The data set consist of eight predictors: (lcavol), (lweight), age, (lbph), (svi), (lcp),

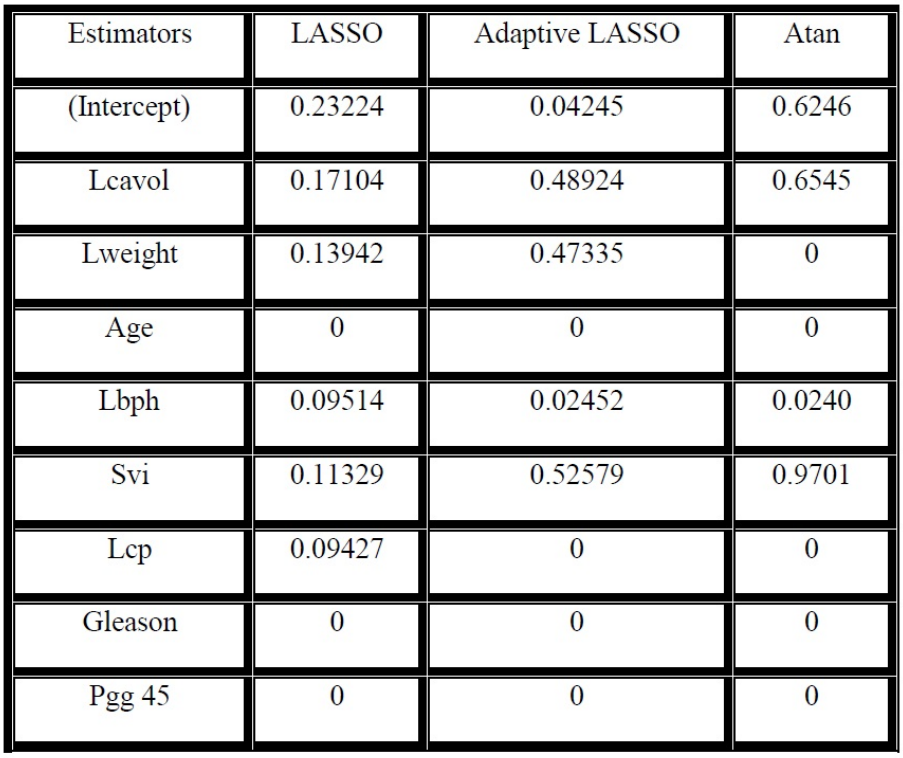
Table 2: simulation results of the second experiment for all estimators



N	Estimators	MAPE	FPR	FNR
50	LASSO	0.19224	0.28	0
	Adaptive LASSO	0.17104	0.22	0
	Atan	0.13942	0.18	0
100	LASSO	0.09728	0.20	0
	Adaptive LASSO	0.09514	0.13	0
	Atan	0.11329	0.11	0
150	LASSO	0.09427	0.15	0
	Adaptive LASSO	0.09032	0.12	0
	Atan	0.06140	0.09	0

(gleason) and (pgg 45), while the response variable is (lpsa) with 97 patients. The results of the application section were displayed in the table 3.

Table 3: simulation results of the second experiment for all estimators



Estimators	LASSO	Adaptive LASSO	Atan
(Intercept)	0.23224	0.04245	0.6246
Lcavol	0.17104	0.48924	0.6545
Lweight	0.13942	0.47335	0
Age	0	0	0
Lbph	0.09514	0.02452	0.0240
Svi	0.11329	0.52579	0.9701
Lcp	0.09427	0	0
Gleason	0	0	0
Pgg 45	0	0	0

From Table 3 we note that the lasso estimator has 5 nonzero coefficients, MCP estimator has 4 nonzero coefficients, while Atan estimator has 3 nonzero coefficients,

That means Atan estimator is the best in estimation and variable selection.

7. Conclusion

A study of Atan was proposed by applying the regularized poisson regression. Atan, Lasso and adaptive Lasso were compared based on the simulation and application. The results of the simulation and application the Atan estimator has excellent performance from Lasso and adaptive Lasso estimators in estimation and variable selection.

References

[1] A. Alfons, C. Croux and S. Gelper, *Sparse least trimmed squares regression for analyzing high-dimensional large data sets*, Ann. Appl. Stat. 7(1) (2013) 226–248.

[2] Z.Y. Algamal, *Diagnostic in Poisson regression models*, Electron. J. App. Stat. Anal. 5(2) (2012) 178–186.

[3] Z.Y. Algamal and M.H. Lee, *Adjusted adaptive Lasso in high-dimensional Poisson regression model*, Modern Appl. Sci. 9(4) (2015) 170–177.

[4] C. Choosawat, O. Reangsephet, P. Srisuradetchai and S. Lisawadi, *Performance comparison of penalized regression methods in Poisson regression under high-dimensional sparse data with multicollinearity*, Thai. Statist. 18(3) (2020) 306–318.

[5] G.N. Collins, R.J. Lee, G.B. McKelvie, A.C.N. Rogers and M. Hehir, *Relationship between prostate specific antigen, prostate volume and age in the benign prostate*, Br. J. Urology 71(4) (1993) 445–450.

[6] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc. 96(456) (2001) 1348–1360.

[7] R. Koenker and G. Bassett Jr., *Regression quantiles*, Econometrica: Journal of the Econometric Society, 46(1) (1978) 33–50.

[8] K. Månsson and G. Shukur, *A Poisson ridge regression estimator*, Economic Modelling, 28(4) (2011) 1475–1481.

[9] 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. 5(5) (2019) 99–103.

[10] M.Y. Park and T. Hastie, *L1-regularization path algorithm for generalized linear models*, J. R. Statist. Soc. B 69(4) (2007) 659–775.

[11] F. Shahzad, F. Abid, A.J. Obaid, B.K. Rai, M. Ashraf and A.S. Abdulbaqi, *Forward stepwise logistic regression approach for determinants of hepatitis B & C among Hiv/Aids patients*, Int. J. Nonlinear Anal. Appl. 12(Special Issue) (2021) 1367–1396.

[12] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. R. Statist. Soc. B 58(1) (1996) 267–288.

[13] Y. Wang and L. Zhu, *Variable selection and parameter estimation with the Atan regularization method*, J. Probab. Statist. 2016 (2016) 1–12.

[14] A.H. Yousif, *Proposing robust LAD-shrink set estimator for high dimensional regression model*, Int. J. Agric. Stat. Sci. 17 (2021) 1229–1234.

[15] A.H. Yousif and O.A. Ali, *Proposing robust LAD-Atan penalty of regression model estimation for high dimensional data*, Baghdad Sci. J. 17(2) (2020) 550–555.

[16] A.H. Yousif and W.J. Housain, *Atan regularized in quantile regression for high dimensional data*, J. Phys. Conf. Ser. 1818 (2021) 012098.

[17] C. Zaldivar, *On the Performance of some Poisson Ridge Regression Estimators*, Master of Science Thesis, Florida International University, 2018.

[18] H. Zou, *The adaptive Lasso and its oracle properties*, J. Amer. Statist. Assoc. 101(476) (2006) 1418–1429.