Journal of Crop and Weed, 19(1): 210-215 (2023)
ISSN- O : 2349 9400 ; P : 0974 6315
http://cwssbckv.org
www.cropandweed.com

The development of a statistical model for forewarning *Helicoverpa armigera* infestation using beta regression

B. GURUNG, [1]S. DUTTA, K.N. SINGH, [*]A. LAMA, [2]S. VENNILA AND [3]B. GURUNG

ICAR-Indian Agricultural Statistics Research Institute, New Delhi, [1]Bidhan Chandra Krishi Viswavidyalaya, Nadia, [2]ICAR-National Research Centre for Integrated Pest Management, New Delhi
[3]School of Agriculture, ITM University, Gwalior

Received : 15.09.2022 ; Revised : 30.12.2022 ; Accepted : 04.01.2023

DOI: https://doi.org/10.22271/09746315.2023.v19.i1.1681

ABSTRACT

Regression analysis is one of the most commonly employed statistical tool for analyzing the "cause and effect" relationship. The regression coefficients are tested to find if they are affecting the dependent variable and the sign and the values of the coefficients help us to know how and by much they are affecting the dependent variables. However, this technique may not be appropriate for statistical analysis where the dependent variable is in proportional scale as they are restricted to the interval (0, 1). As such, the assumptions of error terms being normal and homoscedastic are violated.A reasonable alternative is to use transformations, but this again may lead to bias in estimation. Taking this into consideration, researchers have developed techniques to capture the changing proportional data such that interpretation is easier and also it becomes more flexible than transformation. In our research, we developed a statistical model for forewarning Helicoverpa armigera infestation by employing beta regression. By maximizing the likelihood function by employing the "optim" function through the analytical gradients available in R package, the unknown coefficients are estimated.Moreover, Fisher scoring iteration through the use of expected information and analytical gradients were used to obtain better estimates through the use of "optim". Residual diagnostics have also been carried out. The developed model could also be employed for forewarning pest infestation in other crops.

Keywords **:** Beta regression, Count data, Forewarning, *Helicoverpa armigera,* Homoscedasticity, Modeling

Regression models are very efficient, flexible, and appropriate for studying the relationships between a set of regresors and the dependent variable, which is also called the "cause and effect" relationship. In many practical day to day circumstances, proportional data are analyzed, wherein the dependent variable is expressed in fractions, proportions or percentages. In such scenarios, the usual regression technique is inappropriate as proportions are only bound between zero and one. So, as such, the data may not be Gaussian, and their variance is generally heteroscedastic across the different values the regresors take. A promising way out is to convert the percentage data such that it yields values on the real line, R. The average of the altered response can now be fitted using linear regresors built on a set of predictor variables. However, this approach has shortcomings, one of which is that it will make the estimation of coefficient biassed and, therefore, the prediction based on such estimation procedure will be unrealistic and unstable. Further, the mechanistic interpretation of the estimated coefficients is also tricky in such a situation. Another limitation is that proportion data characteristically display skewness, which in turn will lead to illusory and misleading interpretation if the understanding is based on the normality.

To this end, beta regression may be employed where it is assumed that the dependent variable has beta distribution, which allows us to put up skewed data because of the flexibility provided by the beta distribution. Further, this model has an advantage over the usual logistics regression as the logistic regression will be able to capture the variability in the data, which shows an S-shaped curve only. In contrast, the beta regression can capture any probability-curve depending upon the two parameters of the beta model. A general form of beta regression called the Dirichlet regression may also be employed in some general cases (Hijazi and Jernigan,2009; Boukal, 2015).

In the project funded by 'National Innovations in Climate Resilient Agriculture (NICRA)' of the Indian Council of Agricultural Research, the Government of India, we have data on disease and pest incidences, and their corresponding weather parameters. Regression may not be appropriate for such a scenario as the pest and disease incidence are in percentages. It is to be kept in mind that a considerable amount of data on proportions are collected in agriculture too, so it becomes imperative to study the modeling of such data.

Email: chllm6@gmail.com

How to cite : Gurung, B., Dutta, S., Singh, K.N., Lama, A., Vennila, S. and Gurung, B. 2023. The development of a statistical model for forewarning *Helicoverpa armigera* infestation using beta regression, *J. Crop and Weed*, 19 (1): 210-215.

In this manuscript, we have used beta regression to model and capture the variability in the data related to pest infestation. The developed technique could be employed for forewarning any other pest as well as disease infestation where weather parameters are used as regressor variables.

MATERIALS AND METHODS

Ferrari and Cribari-Neto (2004) and Kieschnick and McCullough (2003) first hypothesized in the 2000s the beta regression model. It is employed for datasets where the dependent variable is bound between two values, usually 0 to 1, and is centred on the assumption that the response variable is beta distributed (Douma and Weedon, 2019). It can tackle the problem of non-constant variance by assuming that the response variable is beta distributed over a range of linear independent variables (Cribari-Neto and Zeileis, 2010). Since its development, beta regression has been popularly employed in social science research, mostly in human behavioral studies (Larrea and Kawachi, 2005; Blane *et al*., 2008).

Model description

The beta regression falls into the class of generalized linear model, but differs from traditional linear regression model as it fits a regress and variable that follows a beta distribution, and not a normal distributed regress and. It can be parameterized by its mean andvariance alone, similar to normal distribution.A heartening feature of the standard deviation of a beta distribution is that it is a function of its mean and a scalar quantity called the 'precision' parameter. This scalar quantity describes how closely assembled the observations are.

The density of the beta regression can be written as:

$$\pi\left(y; p, q\right)=\frac{\Gamma\left(q+p\right)}{\Gamma\left(q\right)\Gamma\left(p\right)}\, y^{p-1}(1-y)^{q-1},\ 0<y<1 \qquad (1)$$

where p>0, q>0, and Γ(.) denotes the gamma function. The arithmetic mean and variance of the regressand can respectively be written as,

$$E\left(y\right)=\frac{p}{\left(p+q\right)}$$

$$Var\left(y\right)=\frac{pq}{\left(p+q\right)^{2}\left(p+q+1\right)}$$

Let

$$u=\frac{p}{(p+q)}$$

$$\varphi = p+q$$

So

$$E\left(y\right)=u$$

$$Var\left(y\right)=\frac{V(u)}{1+\varphi}$$

where $V\left(u\right)=u(1-u)$

u is the arithmetic mean of the dependent variable. The precision parameter is denoted as φ. For fixed μ, the greater the value of φ, the lesser is the variance of y. After the new parameterization, the density of y can now be written as:

$$f\left(y;u,\varphi\right)=\frac{\Gamma\varphi}{\Gamma\left(u\varphi\right)\Gamma\left((1-u)\varphi\right)}\, y^{u\varphi-1}(1-y)^{(1-u)\varphi-1},\ 0<y<1$$

where 0 < u < 1 and $\varphi > 0$

Let $y_1, y_2, y_3,.....y_n$ be n independent random variables, where each y_t follows beta distribution with the density function given in Eq. 1. The random variables have mean u_t, and unknown precision φ. The model is achieved by supposing that the mean can be written as

$$g\left(u_t\right)=\sum_{i=1}^{k}x_{ti}\beta_i=\eta_t$$

where β_i are unknown coefficients and x_{ti} are the data on k independent variables which are supposed to be fixed and also known. Depending on the appropriateness of the data there are numerous likely choices for the link function $g(.)$.

Based on a sample of n independent observations, the likelihood function can be written as :

$$\ell\left(\beta,\varphi\right)=\sum_{t=1}^{n}\ell_t\left(\mu_t,\varphi\right)$$

$$\ell_t\left(\mu_t,\varphi\right)=log\,\Gamma\varphi-log\,\Gamma\left(\mu_t\varphi\right)-log\,\Gamma\left(\left(1-\mu_t\right)\varphi\right)+\left(\mu_t\varphi-1\right)log\,y_t+\left\{\left(1-\mu_t\right)\varphi-1\right\}log\left(1-y_t\right)$$

The estimation of parameters of Beta regression

The above beta regression model is not possible to be estimated by the 'Method of least squares' as minimization of the residual sum of squares return ordinary equations which are nonlinear in the parameters. Since it is not possible to solve the nonlinear equations accurately, the next best way is to obtain fairly accurate logical estimates by using procedures that are iterative, namely, linearization method, steepest descent method, and levenberg-marquardt's method. The

popularity of these methods is certainly due to the fact that they are simple, robust and as well as the availability of many softwares nowadays which have such in-built algorithms. Through the use of maximum likelihood via ‘‘optim” using analytical gradients the estimation of the unknown coefficients is carried out. The initial value, by default, is the coefficient estimates obtained by running a linear regression on a transformed dependent variable. Subsequently, the ‘‘optim” result may be improved by an additional expected information and Fisher scoring iteration through analytical gradients (Grun *et al*., 2012).

Data description

Field experiments were conducted during Rabi 2016-20 at experimental plots in research station farm of Bidhan Chandra KrishiViswavidyalaya (BCKV), Kalyani, Nadia (Elevation: 16m, Latitude: 22° 59’ 15.2’’ and Longitude: 88° 27’ 26.6’’) under the national flagship programme on NICRA, funded by the Government of India, to assess the role of weather variables in the infestation of *Helicoverpa armigera*. It i*s* one of the species of Lepidoptera in the Noctuidae family. The larvae feed on a variety of cultivated crops and is one of the major pests in tomatoes. Data on percentage incidence of *Helicoverpa armigera* along with weather variables like minimum temperature, maximum temperature, morning relative humidity, evening relative humidity, wind-speed, rainfall, sunshine-hour, and rainy-day as the independent variables were employed for illustration purpose and were collected from All India Coordinated Research Project (AICRP) on Agro-Meteorology, Bidhan Chandra KrishiViswavidyalaya, Kalyani, West Bengal. The descriptive statistics are given in Table 1.

Many researches have been carried out on the effect of weather variables on pest or disease infestation. Divyasree *et al*. (2021) conducted an experiment on pigeon pea (*Cajanus cajan* (I) Mill sp), wherein they studied the population build-up of gram pod borer, *Helicoverpa armigera* by using pheromone and light traps. They also studied the impact of weather parameters on trap catch. Wind speed, temperature, and evaporation were found to have contributed significantly to the occurrence of the larval pest population. Adult and larval populations were significantly correlated with maximum temperature and evaporation. Srivastava *et al*. (2016) studied the effect of temperature, growing degree day (GDD), and rainfall on the peak population and the larval incidence of *Helicoverpa armigera* on chickpea and its growth in Bundelkhand, Madhya Pradesh. Moreover, the rising and falling phase of the larval population was also examined. Vikram *et al*. (2018) studied how weather parameters were affecting the *Helicoverpa armigera* (Hubner) incidence on tomatoes. Weather parameters such as wind-velocity, sunshine-hours, and maximum and minimum temperature had a significant positive correlation with the larval population, while morning and evening relative humidity had a nonsignificant negative correlation. Further, rainfall had a non-significant positive correlation with the larval population. Huang and Hao (2020) also conducted an experiment to study how climate change and crop planting structure affected the incidence of cotton bollworms. They found out that the principal factors that affected the first, second, and third generations of moths were mean temperature in June. Gautam *et al*. (2018) studied population dynamics and management of *Helicoverpa armigera* (Hubner) on chickpea. They inferred that the mean population of *Helicoverpa armigera* showed a non-significant negative correlation with rainfall (-0.266), minimum (-0.335) and maximum temperature (-0.220) while RH showed non-significant positive correlation (0.394). Singh *et al*. (2015) studied how weather variables such as maximum temperature, minimum temperature, relative humidity, rainfall, and sunshine hours affected the population and the larval parasitization of *Helicoverpa armigera* in the chickpea ecosystem. In all the above research, beta regression was not employed, so the research we have carried out has the novelty over all of them in that regard. Recently, Varma *et al*. (2022) explored the use of the machine learning algorithms for crop yield forecasting using weather variables where the dependent variable, yield, was on a ratio scale.

RESULTS AND DISCUSSION

The density plot of the *Helicoverpa armigera* incidence data was generated. The density plot gives us an idea regarding the shape the distribution takes and whether it is bell-shaped. It was seen that the plot was far from being bell-shaped or normal.

A quantile-quantile plot (Q-Q plot) was also plotted, which helps to find correlation between a dataset and the normal distribution, was also plotted (Fig. 2). A 45-degree line was also plotted for reference. The plot shows the deviation the dataset under consideration has from an average plot.

Visual inspections which were described above is usually not reliable enough. So a statistically sound test may be employed to compare the datasets to normal distribution to establish statistically whether data follow normal distribution or not. For testing normality, the Shapiro-Wilk normality test was carried out. The test statistics was 0.598 with a p-value <0.0001. We can infer that we have enough evidence to reject the null hypothesis of data being normal.So regression analysis is not a valid option for the data under consideration.

Table 1: Descriptive statistics of variables used for illustration

	Max. Temp. (°C)	Min. Temp. (°C)	RH. Morn. (%)	RH. Even (%)	Rain (mm)	Sun (h/day)	Wind (km/h)	Rainy Day	Incidence of *Helicoverpa*
Mean	26.40	12.21	89.45	55.00	1.50	5.88	0.00	0.13	0.06
Standard Error	0.12	0.12	0.23	0.34	0.28	0.08	0.00	0.02	0.00
Median	26.24	11.84	90.57	56.29	0.00	6.00	0.00	0.00	0.04
Mode	24.79	9.80	92.71	59.86	0.00	7.43	0.00	0.00	0.03
Standard Deviation	2.67	2.73	5.19	7.78	6.32	1.78	0.02	0.48	0.07
Sample Variance	7.11	7.45	26.94	60.59	39.94	3.18	0.00	0.23	0.00
Kurtosis	-0.01	1.73	-0.46	0.00	42.90	-0.50	52.54	28.75	11.75
Skewness	0.21	0.88	-0.50	0.24	6.21	-0.41	7.37	4.86	3.17
Range	15.24	18.45	22.00	48.57	58.80	9.43	0.14	4.00	0.50
Minimum	18.79	6.94	76.43	36.29	0.00	0.57	0.00	0.00	0.00
Maximum	34.03	25.39	98.43	84.86	58.80	10.00	0.14	4.00	0.50

Table 2: Estimated coefficients of the beta regression model

Coefficients (Mean model with a logit link)					
Estimate	**Estimate**	**Std. Error**	**Z value**	**Pr(>|z|)**	
(Intercept)	-2.10	1.37	-1.52	0.12	
Max.Temp. (°C)	-0.18	0.02	-6.40	1.56E-10	***
Min.Temp. (°C)	0.18	0.02	8.15	3.37E-16	***
RH.Morn. (%)	0.01	0.01	2.20	0.02	*
RH.Even (%)	-0.01	0.01	-1.20	0.22	
Rainfall (mm)	0.01	0.01	1.91	0.05	.
Sunshine (h/day)	0.14	0.02	5.86	4.49E-09	***
Wind (km/h)	-1.57	1.76	-0.88	0.37	
RainyDay	-0.16	0.10	-1.62	0.10	
(phi)	28.69	1.86	15.37	<2e-16	***

Note: * is significant at 5%, ** is significant at 1% and *** is significant at <.01% level of significance

A better option in the form of Beta regression is employed.

The estimated coefficients of the Beta regression model is obtained and presented in Table 2. We can see that maximum-temperature, minimum-temperature, and sunshine-hour were highly significant, while relative humidity (morning) was significant at a 3% level of significance. Rainfall too can be seen to be significant at a 6% level of significance. The parameter phi was also found to be highly significant.The results are in sync with various such researches in many other parts of the world. (Divyasree *et al*., 2021; Srivastava *et al*., 2016; Vikram *et al*., 2018; Huang and Hao, 2020).

Further, to validate that the Beta regression model is a good fit to the data set, we need to check the behaviour of the residuals. Statistically, we check for the independence of the residuals. To do so, we have employed Box-Pierce test, which has a null hypothesis of residuals being independent. It was found that the residuals were independent, as Box-Pierce test statistic calculated value was 0.20131 with a p-valueof 0.6537. The obtained result statistically proves that the fitted Beta regression model is able to model the *Helicoverpa armigera* incidence data appropriately.

CONCLUSION

In the present study we have developed a model for forewarning *Helicoverpa armigera* incidence through the use of Beta regression instead of multiple regression as the dependent variable takes value between zero and one only, and as such, the usual regression is not appropriate. The pest incidence is assumed to depend on weather variables which is tested using appropriate statistical tools. The developed model in advance will predict the severity of pest incidence depending upon the pattern of the weather variables. This will allow the subject matter specialist to act accordingly and issue

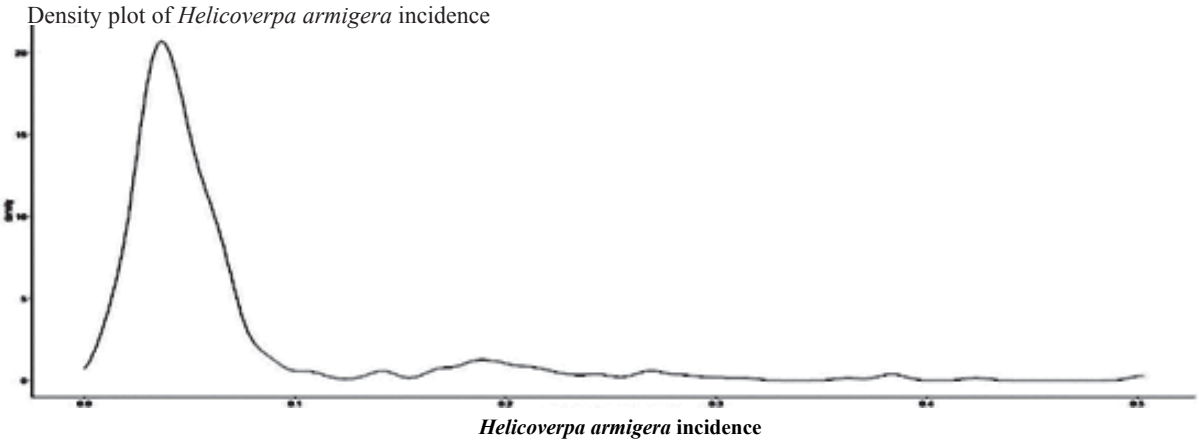

Fig. 1: The density plot of the *Helicoverpa armigera* incidence data

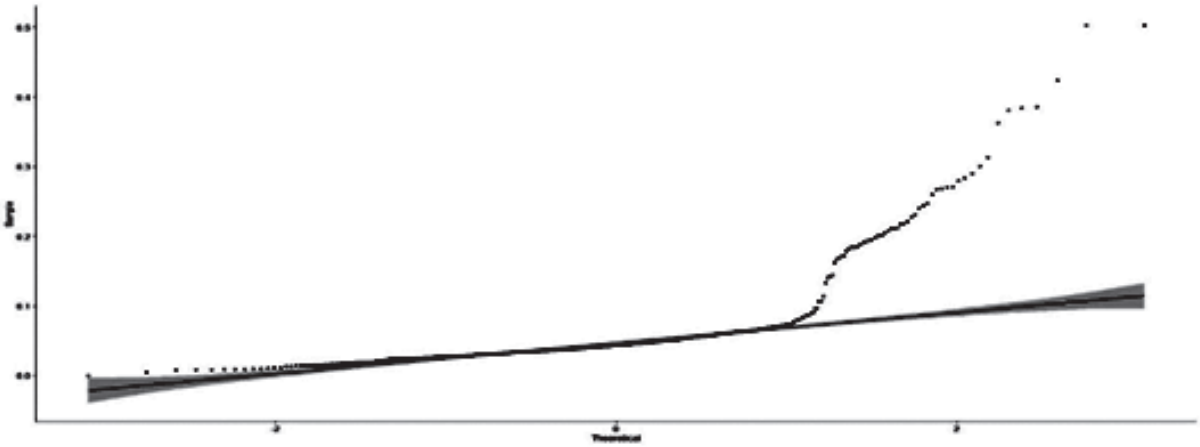

Fig. 2: The Q-Q plot of the *Helicoverpa armigera* incidence data

advisory to the concerned stakeholders for appropriate actions. This in our opinion will help in minimizing crop loss and enhance the income of the farmers. Our study can be extended for pest incidence data on other crops as well. As future scope of work, efforts can be directed towards possibility of variable selection which will make the forewarning model efficient.

ACKNOWLEDGEMENT

We want to express our respect and gratitude to ICAR- National Centre for Integrated Pest Management and funding under the program of 'National Innovations in Climate Resilient Agriculture' funded by the Indian Council of Agricultural Research, the Government of India for providing us the dataset.

REFERENCES

Blane, D.G., Netuveli, and Montgomery, S.M. 2008. Quality of life, health and physiological status and change at older ages. *Soc. Sci. Med.*, **66**:1579-1587.

Boukal, D.S., Ditrich, T., Kutcherov, D., Sroka, P., Dudová, P. and Papáèek, M. 2015. Analysis of developmental rate isomorphy in ectotherms: Introducing the Dirichlet regression. *PLoS ONE*, **10**(6):e0129341.doi.org/10.1371/journal.pone.012934.

Cribari Neto, F. and Zeileis, A. 2010. Beta regression in R. *J. Stat. Soft.*, **34**(2): 1-24.

Divyasree, C., Sreekanth, M., Chiranjeevi, C.H. and Adinarayana, M. 2021. Monitoring of *Helicoverpa armigera* through pheromone and light traps on pigeon pea and impact of weather parameters on trap catch. *Int. J. Chem. Stud.*, **9**(3): 170-173

Douma, J.C. and Weedon, J.D. 2019. Analyzing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Met.Ecol.Evol.*, **10:** 1412-1430.

Ferrari, S. and Cribari Neto, F. 2004. Beta regression for modelling rates and proportions. *J. App. Stat.*, **31**(7): 799-815.

Gautam, M.P., Chandra, U., Yadav, S.K., Jaiswal, R., Giri, S.K. and Singh, S.N. 2018. Studies on population dynamics of gram pod borer *Helicoverpa armigera* (Hubner) on chickpea (*Cicer arietinum* L.). *J. Entomol. Zool. Stud.*, **6**(1): 904-906.

Grun, B.,Kosmidis, I. and Zeileis, A. 2012. Extended Beta regression in R: shaken, stirred, mixed, and partitioned. *J. Stat. Soft.*, **48**(11): 1-25. URL http://www. jstatsoft.org/v48/i11/.

Hijazi, R. and Jernigan, R. 2009. Modeling compositional data using Dirichlet regression models. *J. App. Prob. Stat.*, **4**(1): 77–91.

Huang, J. and Hao, H. 2020. Effects of climate change and crop planting structure on the abundance of cotton bollworm, *Helicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae). *Ecol. Evol.,***10**: 1324–1338.

Kieschnick, R. and McCullough, B. D. 2003. Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions. *Stat. Model.*, **3**(3): 193–213.

Larrea, C. and Kawachi, I. 2005. Does economic inequality affect child malnutrition? The case of Ecuador. *Soc. Sci. Med.*, **60:**165-178.

Singh, D., Singh, S.K. and Vennila, S. 2015. Weather parameters influence population and larval parasitization of *Helicoverpa armigera* (Hübner) in chickpea ecosystem. *Leg. Res.*, **38**(3): 1-5.

Srivastava, A.K., Nayak, M.K., Yogranjan, Tomar, D.S. and Gurjar, K. 2016. Weather-based prediction of chickpea *Helicoverpa armigera* population in Bundelkhand agroclimatic zone of Madhya Pradesh. *Mausam*, **67**(2): 377-388.

Varma, M., Singh, K.N., and Lama, A. 2022. Exploring the suitability of machine learning algorithms for crop yield forecasting using weather variables. *J. Crop and Weed*, **18**(1): 210-214.

Vikram, A.K., Keshav, M., and Choudhary, R. 2018. Effect of weather parameters on incidence of key pest, *Helicoverpa armigera* (Hubner) on tomato. *J. Entomol. Zool. Stud.*, **6**(1): 97-99.