International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.04, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Application of Generalized Linear Models in Agriculture

[1]Yasmeena Ismail*, [2]Nageena Nazir*

**Division of agricultural statistics,*

Sher-e-Kashmir University of Agricultural Sciences and Technology-Kashmir (INDIA)

ABSTRACT

A generalized linear model is an extension of classical linear models with the response from an exponential family of distributions. The Akiake Information Criterion is used as a comprehensive tool for selecting the adequate model. In this paper, we have fitted both classical as well as generalized linear models on a dataset pertaining to yield and yield attributes of tulip (Tulipa Lilioideae). On the basis of AIC criterion it was observed that gamma generalized linear model comes out to be the best fit for the above dataset. This study suggests that the parameter estimates in these models can be greatly influenced by the assumption about the error structures in the estimation and that gamma distributions are appropriate for the production model in assessing the yield of tulip. We recommend that generalized linear models should be used to identify the appropriate error structure in modeling production of tulips.

Keywords: Akaike Information Criteria, Generalized Linear Models, Gamma, Statistical Models, Tulips.

I INTRODUCTION

Tulips occupy a prominent position among the top 10 flowers of the world. In the international flower market, tulips command good demand on account of their elegant flowers of different hues and shapes. Tulip is a temperate crop and the bulbs require a cold temperature regime for flower initiation. Tulips are less influenced by light. However, under KASHMIR conditions, partial reduction in sunlight results in healthier plants with longer flower stalks. Tulips prefer light soil with a low salt content and pH of 6-7. Well decomposed FYM should be mixed @ 5-10 kg per m2, depending on the soil condition, to enrich the soil. Plants can be grown outdoor and under green house conditions. To get a good crop of flower, bulbs of 10-12cm size or more should be planted 15cm x 30cm apart i,e, 10,000 bulbs per kannal area. The tunics should be removed before planting the bulbs. The flowering plants will produce 4-6 leaves. Smaller (less than 6-8 cm) bulbs will produce plants with single, leaf and will not Page 2 of 9 produce any flower until the bulbs reach proper size. Recommended cultivation practices are followed to get quality flowers [1]. Using the ordinary multiple-linear regression in addressing and analyzing the production of tulip might be restrictive, as it assumes that the response variable follows the normal distribution only, which is not convenient in

practice. The generalized linear models [2] assume a more general class of distributions to the response variable, which makes modeling production data more feasible.

The generalization in the generalized linear models over the ordinary multiple linear regression is in two ways. The variable of our main interest i.e., y (dependent variable) is allowed to follow any distribution that belongs to the exponential family of distributions, and not just the normal distribution [3]. The mean of the variable y does not need to be the linear on the explanatory variable $x(x = (x_1, x_2, ..., x_p)^T)$, as long as it is linear on another scale. In order to establish the notation for the components of the generalized linear models we define the model to have the following systems of equations

$$y = \mu + \varepsilon$$
$$g(\mu) = x^T \beta$$

Where y is the response variable to some covariate $(x = (x_1, x_2, ..., x_p)^T)$.

With the subsequent development and spread of GLM computer software, the importance of these models in practical data analysis has greatly expanded [4].

II MATERIAL AND METHODS

In the generalized linear models the distribution of *y* is assumed to be a member of the exponential family. A distribution is said to be a member of the exponential family if its probability mass/density function can be represented in the form:

$$f(y \mid \theta, \phi) = \exp\left[\frac{y(\theta) - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$

Where a, b and c are given functions. Moreover, θ and ϕ are parameters. In particular, θ is called the natural parameter, and it is the main parameter of interest. ϕ is called the dispersion parameter [5]. When $y \sim Gamma(\nu, \lambda)$ then the density of the gamma distribution is usually given by

$$f(y) = \frac{1}{\Gamma(\nu)} \lambda^{\nu} y^{\nu-1} e^{-\lambda y}; y > 0$$

where ν describes the shape and λ describes the scale of the distribution. However, for the purposes of a GLM, it is convenient to reparameterize by putting $\lambda = \nu / \mu$ to get:

$$f(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} e^{-\left(\frac{y\nu}{\mu}\right)}; y > 0$$

Now $E(y)=\mu$ and $y=\mu^2/\nu=(E(y))^2/\nu$. The dispersion parameter is $\phi=\nu^{-1}$. The canonical parameter is $-1/\mu$, so the canonical link is $\eta=1/\mu$. [6].

In this paper, to study the statistical model of tulip we have taken the data of tulip pertaining to the yield and yield attributes of the same. The data consists of three varieties on which three treatments are applied each treatment is replicated three times. We check the normality of the data by Shapiro-Wilk test [7] and we find that the response fails to follow the normal distribution. Thus, we fit four generalized linear models on each of the responses one is the emergence and other bud variables of the tulip, where these response variables are allowed to follow any of the distributions from an exponential family of distributions. To select the appropriate model (model selection) Akaike Information Criteria (AIC) [8] is used and the model with minimum AIC is selected to be the best for the model.

To establish notation for the proposed generalized linear model we let y_{ijk} be the observations in the cell *ijk* where *i*, *j* and *k* represents the treatment level, replication level and variety consecutively. Additionally, *i=1,2,3; j=1,2,3; k=1,2,3.* In general, when $y_{ijk} \sim Gamma(\nu_{ijk}, \lambda_{ijk})$, the first model we will fit is

$$y_{ijk}=\mu+\alpha_i+\beta_j+\gamma_k+\varepsilon \qquad \text{(model I)}$$

Where, y_{ijk}=Emergence of tulip flower, μ is the intercept α_i, β_j and γ_k are the parameters representing treatment, replication and variety consecutively.

III RESULTS AND DISCUSSION

On the basis of the AIC criteria it is observed that the gamma distribution has the minimum AIC value and we fit the generalized linear model based on the gamma distribution shown in the table 1.

Table 1: AIC value of different distributions

Distribution	**Emergence**	
	Link	AIC
Gamma	Log	106.56
Inverse Gaussian	Identity	108.6
Weibull	log	109.62
Multinomial	Cumulative Logit	1368

Table 2: Estimates of the parameters of the model I

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	4.715e+00	2.981e-03	1581.769	< 2e-16 ***
VARIETYV2	4.950e-02	3.481e-03	14.219	2.47e-14 ***
VARIETYV3	8.245e-03	3.481e-03	2.368	0.02500 *
TREATT2	-1.055e-02	4.020e-03	-2.625	0.01388 *
TREATT3	-1.655e-03	4.020e-03	-0.412	0.03366 *
TREATT4	2.930e-03	4.020e-03	0.729	0.04203 *
TREATT5	-6.241e-03	4.020e-03	-1.553	0.03175 *
TREATT6	3.139e-03	4.020e-03	0.781	0.04141 *
TREATT7	-4.655e-03	4.020e-03	-1.158	0.04665 *
TREATT8	-1.892e-03	4.020e-03	-0.471	0.04150 *
REPLICATR2	-1.309e-03	3.481e-03	-0.376	0.70969
REPLICATR3	7.461e-03	3.481e-03	2.143	0.04091 *
VARIETYV2:TREATT2	7.828e-03	4.403e-03	1.778	0.08631 .
VARIETYV3:TREATT2	5.493e-03	4.403e-03	1.247	0.22257
VARIETYV2:TREATT3	8.519e-03	4.403e-03	1.935	0.06320 .
VARIETYV3:TREATT3	2.761e-03	4.403e-03	0.627	0.53573
VARIETYV2:TREATT4	-4.759e-03	4.403e-03	-1.081	0.28903
VARIETYV3:TREATT4	-1.036e-03	4.403e-03	-0.235	0.81573
VARIETYV2:TREATT5	1.953e-03	4.403e-03	0.444	0.66079
VARIETYV3:TREATT5	6.824e-03	4.403e-03	1.550	0.13246
VARIETYV2:TREATT6	1.518e-03	4.403e-03	0.345	0.73289
VARIETYV3:TREATT6	-3.618e-03	4.403e-03	-0.822	0.41823
VARIETYV2:TREATT7	9.332e-03	4.403e-03	2.119	0.04306 *
VARIETYV3:TREATT7	1.270e-02	4.403e-03	2.883	0.00749 **
VARIETYV2:TREATT8	-4.536e-04	4.403e-03	-0.103	0.91868
VARIETYV3:TREATT8	5.838e-03	4.403e-03	1.326	0.19564
VARIETYV2:REPLICATR2	7.536e-03	2.697e-03	2.795	0.00927 **
VARIETYV3:REPLICATR2	1.277e-03	2.697e-03	0.474	0.63936
VARIETYV2:REPLICATR3	6.912e-03	2.697e-03	2.563	0.01603 *
VARIETYV3:REPLICATR3	-2.286e-03	2.697e-03	-0.848	0.40365

TREATT2:REPLICATR2	1.929e-03	4.403e-03	0.438	0.66475
TREATT3:REPLICATR2	-3.688e-03	4.403e-03	-0.837	0.40942
TREATT4:REPLICATR2	1.502e-03	4.403e-03	0.341	0.73565
TREATT5:REPLICATR2	2.068e-03	4.403e-03	0.470	0.64230
TREATT6:REPLICATR2	5.855e-03	4.403e-03	1.330	0.19434
TREATT7:REPLICATR2	-1.556e-03	4.403e-03	-0.353	0.72648
TREATT8:REPLICATR2	9.295e-05	4.403e-03	0.021	0.98331
TREATT2:REPLICATR3	4.318e-03	4.403e-03	0.981	0.33514
TREATT3:REPLICATR3	-2.013e-03	4.403e-03	-0.457	0.65115
TREATT4:REPLICATR3	-2.583e-03	4.403e-03	-0.587	0.56223
TREATT5:REPLICATR3	2.612e-03	4.403e-03	0.593	0.55779
TREATT6:REPLICATR3	-3.369e-03	4.403e-03	-0.765	0.45055
TREATT7:REPLICATR3	-4.275e-03	4.403e-03	-0.971	0.33994
TREATT8:REPLICATR3	-4.852e-03	4.403e-03	-1.102	0.27993

Additionally, we get that the null deviance is 0.046 on 71 degrees of freedom. The residual deviance is 0.00040699 on 28 degrees of freedom AIC is 106.58. Significance of the parameters can be looked upon by their *pr(>|z|)* values in table 2. We can see that the parameters for varieties and treatments are significant and for the third replication at 5% significance level, because their *pr(|z|)* values are less than 0.05.

The second model we will fit is

$$z_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon \qquad \text{(model II)}$$

Where, z_{ijk} =Bud of flower

Table 3:AIC value of different distributions

Distribution	Link	AIC
Gamma	Log	197.25
Inverse Gaussian	Identity	204.49
Weibull	log	199.92
Multinomial	Cumulative logit	648.04

In Table 3 we see that the Gamma distribution has the minimum AIC and we fit the generalized linear model based on the gamma distribution. Fitting the model is shown in table 4 as :

Table 4: Estimates of the parameters of the model II

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.9293307	0.0024399	2020.339	< 2e-16 ***
VARIETYV2	0.0141211	0.0028491	4.956	3.13e-05 ***
VARIETYV3	0.0033143	0.0028491	1.163	0.2545
TREATT2	0.0066579	0.0032899	2.024	0.0526 .
TREATT3	0.0011518	0.0032899	0.350	0.7289
TREATT4	0.0082034	0.0032899	2.494	0.0188 *
TREATT5	0.0054653	0.0032899	1.661	0.0178 *
TREATT6	0.0069501	0.0032899	2.113	0.0437 *
TREATT7	0.0024608	0.0032899	0.748	0.0407 *
TREATT8	0.0010036	0.0032899	0.305	0.0626
REPLICATR2	0.0023436	0.0028491	0.823	0.0177 *
REPLICATR3	0.0032930	0.0028491	1.156	0.0575.
VARIETYV2:TREATT2	-0.0051016	0.0036039	-1.416	0.1679
VARIETYV3:TREATT2	-0.0046529	0.0036039	-1.291	0.2072
VARIETYV2:TREATT3	0.0020259	0.0036039	0.562	0.5785
VARIETYV3:TREATT3	0.0004088	0.0036039	0.113	0.9105
VARIETYV2:TREATT4	-0.0030358	0.0036039	-0.842	0.4067
VARIETYV3:TREATT4	-0.0017022	0.0036039	-0.472	0.6404
VARIETYV2:TREATT5	-0.0062869	0.0036039	-1.744	0.0920 .
VARIETYV3:TREATT5	-0.0022315	0.0036039	-0.619	0.5408
VARIETYV2:TREATT6	-0.0049317	0.0036039	-1.368	0.1821
VARIETYV3:TREATT6	-0.0004930	0.0036039	-0.137	0.8922
VARIETYV2:TREATT7	0.0017049	0.0036039	0.473	0.6398
VARIETYV3:TREATT7	0.0010513	0.0036039	0.292	0.7727
VARIETYV2:TREATT8	0.0015277	0.0036039	0.424	0.6749
VARIETYV3:TREATT8	0.0020728	0.0036039	0.575	0.5698
VARIETYV2:REPLICATR2	-0.0037735	0.0022069	-1.710	0.0984 .
VARIETYV3:REPLICATR2	-0.0015457	0.0022069	-0.700	0.4895
VARIETYV2:REPLICATR3	-0.0002473	0.0022069	-0.112	0.9116
VARIETYV3:REPLICATR3	0.0006013	0.0022069	0.272	0.7873

TREATT2:REPLICATR2	-0.0013261	0.0036039	-0.368	0.7157
TREATT3:REPLICATR2	0.0018189	0.0036039	0.505	0.6177
TREATT4:REPLICATR2	-0.0026507	0.0036039	-0.736	0.4682
TREATT5:REPLICATR2	0.0008202	0.0036039	0.228	0.8216
TREATT6:REPLICATR2	-0.0075599	0.0036039	-2.098	0.0451 *
TREATT7:REPLICATR2	0.0013975	0.0036039	0.388	0.7011
TREATT8:REPLICATR2	-0.0011762	0.0036039	-0.326	0.7466
TREATT2:REPLICATR3	-0.0040384	0.0036039	-1.121	0.2720
TREATT3:REPLICATR3	0.0005491	0.0036039	0.152	0.8800
TREATT4:REPLICATR3	-0.0077214	0.0036039	-2.143	0.0410 *
TREATT5:REPLICATR3	-0.0028994	0.0036039	-0.805	0.4279
TREATT6:REPLICATR3	-0.0074839	0.0036039	-2.077	0.0471 *
TREATT7:REPLICATR3	-0.0009387	0.0036039	-0.260	0.7964
TREATT8:REPLICATR3	-0.0068311	0.0036039	-1.895	0.0684 .

The null deviance is 0.0024 on 71 degrees of freedom. The residual deviance is 0.00027 on 28 degrees of freedom AIC is 197.25. Significance of the parameters can be looked upon by their *pr(>|z|)* values in table 4. We observe that the parameters for varieties and maximum number of treatments are significant and for the second replication at 5% significance level, because their *pr(|z|)* values are less than 0.05.

The above study indicates that this type of data sometimes may have the non-normal response and should be evaluated by the generalized linear model methods. Such data need to be checked for normality assumptions. If the response comes out to be non-normal instead of correcting the data for normality the generalized linear model based on the gamma distribution should be applied. Correcting the data for normality leads to the loss of information. Generalized linear models are superior models in situations where the response is non-normal and follows any distribution from a family of exponential distributions.

REFERENCES

[1] Department of Floriculture Kashmir 2012.

[2] Nelder J.A. & Wedderburn, R.W.M. (1972). *Generalized Linear Models*, Journal of the Royal Statistical Society, A, 135, 370–384.

[3] Siddig, Murwan HMA. "Application of the Generalized Linear Models in Actuarial Framework." *arXiv preprint arXiv:1611.02556* (2016).

[4] McCullagh, P. and J.Nelder (1989). *Generalized Linear Models* (2 ed.). London: Chapman & Hall.

[5] Yuan, J. (2014). *Lecture notes for Generalised Linear Models, Goodness-of-fit*. The University of Manchester.

[6] Faraway, Julian J. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Vol. 124. CRC press, 2016.

[7] Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591-611.

[8] Akaike, Hirotugu. "Likelihood of a model and information criteria." *Journal of econometrics* 16.1 (1981): 3-14.