RESEARCH REVIEW JOURNALS
Volume-03
Issue-07
July-2018

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Study of Logit Regression - A statistical analysis of GSRTC Data

[1]Bhavika D Shah & [2]Pravender

[1]Research Scholar, Department of Statistics, Gujarat University, Ahmadabad (India)
[2]Research Guide, Department of Statistics, Gujarat University, Ahmadabad (India)

INTRODUCTION TO LOGIT REGRESSION

The back-slide models for dichotomous data, including vital backslide and probabilistic model. These models are appropriate when the response takes one of only two possible characteristics addressing progress and dissatisfaction, or all the generally the proximity or nonattendance of a property of premium.

LOGISTIC REGRESSION

It starts by familiarizing with outline the assessment of twofold data. By then it is discussed the stochastic structure of the data to the extent the Bernoulli and binomial movements, and the conscious structure similarly as the logit change. The result is a summarized straight model with binomial response and association logit.

THE BINOMIAL DISTRIBUTION

It considers first the circumstance where the response Yi is matched, tolerating only two characteristics that for convenience we code as one or zero. For example, we could describe.

$$Y = \begin{cases} 1 \text{ if the value is higher than the median} \\ 0 \text{ Otherwise} \end{cases}$$

It is seen that Yi as an understanding of an irregular variable Yi that can yield the measures 1 and 0 with probabilities θi and 1 - θi, individually. The dispersal of Yi is known as a Bernoulli circulation with factor θi, and can be recorded in thick structure as:$P(Yi) = \theta i^{yi}(1 - \theta i)^{1-yi}$

For Y_i = 0 or 1, if Y_i = 1 will produce θ_i, and if Y_i = 0 than it will produce 1- θi. It is impartially easy to prove by through intention that the estimated value and alteration of Yi are

$$E(Yi) = \pi i = \theta i, \text{and}$$
$$\text{Var}(Yi) = \epsilon_i^2 = \theta i(1 — \theta i).$$

It is likewise seen that the mean and distinction depend upon the essential probability θi. Any impact that impacts the probability will adjust the mean just as the difference in the discernment. This suggests an immediate model that empowers the pointers to impact the mean anyway expect that the vacillation is consistent won't be palatable for the assessment of combined data.

Expect since the units under assessment can be orchestrated by the components of excitement into k bundles so all individuals in a social event have undefined estimations of all abrades.

It is watched Yi as an affirmation of a sporadic variable Yi that takes the characteristics 0, 1... θi. If the θi observations in each social event are free, and they all have a comparable probability ni of having the property of interest, by then the dispersal of Yi is binomial with parameters θi and ni, which can be make:

$$Y_i \sim Bi(\theta_i n_i)$$

The probability density function of dependent variable can be given as:

$$P(Yi) = NC_Y . \theta i^{yi}(1 - \theta i)^{1-yi}$$

Here, the main stage demonstrates higher the estimation of perception than middle and second stage is showing bring down the incentive than middle.

The least complex technique to get this result is according to the accompanying. Let Y_j be a marker variable that takes the characteristics one or zero if the j-th unit in social event iis a triumph or a error, independently. Note that Yj is a Bernoulli self-assertive variable with mean and change as given in Equation 3.2. We can make the amount out of triumphs Yi in social affair I in general of the individual marker factors, so Yi = XjYij. The mean of Yi is then the aggregate of the individual strategies, and by opportunity, its distinction is the entire of the individual changes, inciting the result in induced condition. Note again that the mean and contrast depend upon the concealed probability m. Any factor that impacts this probability will impact both the mean and the change of the recognitions.

From a logical point of view the amassed data definition given here is the most wide one; it joins solitary data as the outstanding circumstance where we have n social events of size one, so k = n and ni = 1 for all I. It furthermore joins as a remarkable case the different over the top where the essential probability is the equal for all individuals and we have a single social event, with k = 1 and n_i = n. Thusly, all need to consider with respect to estimation and testing is the binomial transport.

From a sensible point of view it is basic to observe that if the indicators are discrete factors and the outcomes are self-sufficient, we can use the Bernoulli allotment for the individual zero-one data or the binomial scattering for collected data involving incorporates of triumphs in each social event. The two systems are proportionate, as in they lead to a similar likelihood work and therefore comparable evaluations and standard errors. Working with collected data when it is possible has the extra favored outlook that, dependent upon the degree of the social occasions; it winds up possible to test the conventionality of assault of the model. To the extent our model we can work with 16 social affairs of women (or less when we dismiss a part of the pointers) and get the extremely same measures as we would if we worked with the 1607 individuals.

THE LOGIT TRANSFORMATION

The accompanying stage in describing a model for data concerns the productive structure. It should have the

probabilities ni depend upon a vector of viewed abrades x'. The most clear idea is allowed ni to be an immediate limit of the abrades, state

$$\theta_i = X_i . b$$

Here, b is a vector of backslide coefficients. The described model is every so often called the immediate probability appear. This model is much of the time surveyed from individual data using standard least squares.

One issue with this model is that the probability n_i on the left-hand-side must be some place in the scope of zero and one, yet the straight marker X'b on the right-hand-side can take any real regard, so there is no affirmation that the foreseen characteristics will be in the correct range aside from if complex impediments are constrained on the coefficients.

An essential reaction for this issue is to change the probability to remove the range restrictions, and model the change as an immediate limit of rubs. It does this in two phases.

The principal likelihood can be resolved as odd to the worth higher than the middle:$\rho = \text{logit}\,(\theta_i) = \log\frac{\theta_i}{1-\theta_i}$.It has the effect of clearing the floor impediment. To see this point note that as the probability goes down to 0 the odds approach 0 and the logit methodologies negative wearisome characteristics. At the other exceptional, as the probability approaches one the odds approach in this way does the logit. As such, logits map probabilities from the range (0, 1) to the entire real line. Note that if the probability is 0.5 the odds are even and the logit is 0. Negative logits address probabilities underneath one half and positive logits identify with probabilities more than one half.

The backslide coefficients b can be deciphered correspondingly as in direct models, recalling that the left side is a logit instead of a mean. Therefore, b_j addresses the alteration in the logit of the probability related with a unit change in the j^{th} marker holding each and every other pointer relentless. While imparting results in the logit scale will be new from the outset, it has the great position that the model is genuinely clear in this particular scale. Exponentiation Equation it is find that the odds for the i^{th} unit are given by exponential type of condition.

This verbalization portrays a multiplicative model for the odds. For example on the off chance that we by one way or another happened to change the jth pointer by one unit while holding each and every other variable unfaltering, we would copy the odds by foreseen estimation of b. To see this point surmise the immediate pointer is Xi. b and we increase Xj by one, to get X'i. b + bj. Exponentiation gets exponential Xi. b times foreseen b. Consequently, the exponentiation coefficient b addresses an odds extent. Making an understanding of the results into multiplicative ramifications for the odds, or chances extents, is much of the time helpful, in light of the fact that we can deal with a continuously unmistakable scale while holding a decently fundamental model.

Making do with the probability θi in the logit show gives the more confounded model

$$\theta_i = \frac{\text{Exp}\,(X_i^{'}.b)}{1 + \text{Exp}\,(X_i^{'}.b)}$$

While the left side is in the characteristic probability scale, the correct side is a non-direct limit of the pointers, and there is no essential technique to express the effect on the probability of growing a marker by one unit while holding substitute factors consistent. We can obtain an inaccurate reaction by accepting backups with respect to Xj, which clearly looks good only for steady pointers. Using the rest of we get

$$\frac{\partial\theta_i}{\partial X_{ij}} = b_i\theta_i(1-\theta_i).$$

As such, the effect of the jth pointer on the probability θi depends upon the coefficient bj and the estimation of the probability. Analysts every so often survey this thing setting θi to the model mean. The result approximates the effect of the covariate near the mean of the response.

APPLICATION OF LOGIT REGRESSION

This model is determined by taking the binary panel structure of data. The binary data are recorded based on the actual data. The bifurcation is given by taking median of the actual data. Higher the values than median is coded as 1 and lower the value than median is coded as 0. In order to define structure the binary panel data has been compiled.

The basic logit regression is than run to examine the parametric impact on total loss. It is presented in table 1.1 as follows:

Table 1.1: Logit, using 192 observations Dependent variable: Loss

	Coeffi.	SE	z	p-value	
Const.	−1.39	0.740	−1.88	0.0591	*
Effective KM	0.444	0.575	0.77	0.4399	
No. of Passengers	1.869	0.565	3.308	0.0009	***
Total EPKM	−0.116	0.798	−0.145	0.8843	
Total CPKM	1.683	0.563	2.991	0.0028	***
Load Factor	0.547	0.498	1.09	0.2720	
Vehicle Utilized per Day	−1.046	0.525	−1.99	0.0464	**
Fleet Utilization	−0.327	0.418	−0.782	0.4342	
Crew Utilization	−0.688	0.641	−1.073	0.2831	
Diesel KMPL	−0.239	0.583	−0.409	0.6820	
Engine Oil KMPL	0.124	0.525	0.236	0.8130	

Break Down	−0.130	0.571	−0.228	0.8194	
Accidents	0.136	0.645	0.211	0.8328	

Mean dependent var	0.473958	S.D. dependent var	0.500627
McFadden R-squared	0.276111	Adjusted R-squared	0.178237
Log-likelihood	−96.14966	Akaike criterion	218.2993
Schwarz criterion	260.6468	Hannan-Quinn	235.4503

The mathematical formation is given for the derived logit regression model as:

$$\text{Loss} = -1.39 + 0.444\ \text{Effective KM} + 1.869\ \text{No. of Passangers} - 0.116\ \text{total EPKM} + 1.683\ \text{total CPKM} + 0.547\ \text{load factor} - 1.046\ \text{Vehical Utilized per day} - 0.327\ \text{Fleet Utilization} - 0.628\ \text{Crew utilization} - 0.239\ \text{diesel KMPL} + 0.124\ \text{Engine oil KMPL} - 0.130\ \text{Breakdown} + 0.136\ \text{Accidents}$$

It is observed that the base of the model is defined negative. Thus, negative signed parameters have increased the loss and positive singed parameters can reduce the loss to GSRTC. It is observed that effective kilometers, total number of passengers, total cost per kilometers, load factors, engine oil consumption kilometer per liter and number of accidents are positively associated with dependent variable loss. These parameters sound in favor to GSRTC. Rest of the variables such as; total earning per kilometer, vehicle utilization per day, fleet utilization, crew utilization, use of diesel kilometer per liter and break downs of vehicles are the major responsible parameters having effect to increase the loss to GSRTC. This model is clearly indicating that total number of passenger and effective kilometers are playing very important role to reduce the total loss. GSRTC has to examine their resource to make the decided variable in their favor.

The basic statistics are also very important to compare the model to understand the best fit results. For binary statistics the co-efficient of correlation is computed by taking McFadden r. The R –square value is computed very poor. It shows that 27.61% results are due to inclusive impact of GSRTC functioning. Other 72.39% variation is not explained between the variables. Other factors are playing very important role to caused GSRTC in loss. The model testing parameters - Log-likelihood (−96.14), Schwarz criterion (BIC = 260.6), Akaike criterion (AIC = 218.3) and Hannan-Quinn (HQ = 235.45) are computed lower. These can support to understand the nature of different models while comparison of two or more models.

References:

1. Agresti, A. 2002. Categorical Data Analysis, 2nd ed. New York: John Wiley and Sons.
2. Aldrich, J. H., and F. D. Nelson. 1994. Linear Probability, Logit and Probit Models. Thousand Oaks, Calif.: Sage Publications, Inc.
3. Finn, J. D. 1974. A general model for multivariate analysis. New York: Holt, Rinehart and Winston.
4. Fox, J. 1984. Linear statistical models and related methods: With applications to social research. New York: John Wiley and Sons.
5. Hosmer, D. W., and S. Lemeshow. 2000. Applied Logistic Regression, 2nd ed. New York: John Wiley and Sons.
6. Kirk, R. E. 1982. Experimental design, 2nd ed. Monterey, California: Brooks/Cole.
7. McCullagh, P., and J. A. Nelder. 1989. Generalized Linear Models, 2nd ed. London: Chapman & Hall.
8. Hosmer DW Jr, Lemeshow S, Sturdivant RX (2013) Applied Logistic Regression. Third Edition. New Jersey: John Wiley & Sons.
9. Long JS (1997) Regression Models for categorical and limited dependent variables. Thousand Oaks, CA: Sage Publications.

