Ship Damage Rates Analysis Based on Poisson Regression

Gao Shang*
School of Computer Science and Engineering
Jiangsu University of Science and Technology
Zhenjiang, China
e-mail: gao_shang@just.edu.cn

Qian qiang
School of Computer Science and Engineering
Jiangsu University of Science and Technology
Zhenjiang, China
e-mail: qianqiang_just@163.com

Abstract—Ship damage rates analysis is a important in shipbuilding. Poisson regression is a regression model for analyzing the dependent variable of count data. Ship damage rates forecast model based on Poisson regression is proposed. Using SPSS Clementine data mining tool, the ship data is analysis by Poisson regression. Some interpretations are made based on the parameter estimates.

Keywords- Ship damage rate;, Poisson regression; SPSS; parameter estimates

I. INTRODUCTION

Ship damage stability is a long-lasting puzzle in shipbuilding field, which involves many complex technical issues as rolling, flooding and capsizing of damaged ship in random waves[1-5].There are many phenomena where the dependent variable is the count type (which can take on nonnegative inter values: {0,1,2,…}), such as the number of restatement of financial statements during a period, or the number of patents received by a firm per year. The underlying variable in each case, the outcome variable, is discrete. Note that these numbers are actual counts, which are different form the ordinal numbers. Sometimes count data can also refer to `rare`, or `infrequent`, occurrences such as failing *n* tests. Just as the Bernoulli distribution was chosen to model the yes/no decision, the probability model distribution that is specifically suited for count data is the Poisson probability distribution. A generalized linear model can be used to fit a Poisson regression for the analysis of count data[6-10]. For example, a dataset presented and analyzed elsewhere concerns damage to cargo ships caused by waves. The incident counts can be modeled as occurring at a Poisson rate given the values of the predictors, and the resulting model can help you determine which ship types are most prone to damage.

II. THE POISSON DISTRIBUTION AND POISSON REGRESSION

In statistics, Poisson regression is a form of regression analysis used to model count data and contingency tables[1]. The basic distribution for describing counts is the Poisson distribution, which arises in connection with the Poisson process.

A Poisson distributed random variable Y with parameter $\lambda > 0$ is defined for all nonnegative integer numbers 0, 1, 2, ...

The density of a Poisson distribution is

$$P\{Y = y\} = \frac{\lambda^y}{y!} e^{-\lambda} \tag{1}$$

Mean and variance are

$$E(Y) = \lambda \,,\ Var(y) = \lambda$$

Hence, for a Poisson distributed variable the mean and variance are equal.

Notice an interesting feature of the Poisson distribution: Its variance is the same as its mean value. The parameter λ on the Poisson regression model may be written as a log-linear model (Lawless J. E. ,1984; Hilbe 2007):

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + \cdots + \beta_k x_{ki}$$

It equivalents to (because exp() is always positive))

$$\lambda_i = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + \cdots + \beta_k x_{ki})$$

where the $X_i = (x_{1i}, x_{2i}, \cdots, x_{ki})^T$ are some of the variables that might affect the mean value. Then, the probability that *y* equals the value *h*, conditional on X_i, is:

$$P(y = h \mid X_i)$$

$$= \frac{1}{h!} \cdot \exp[-\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + \cdots + \beta_k x_{ki})]$$

$$[\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + \cdots + \beta_k x_{ki})]^h$$

The parameters $\beta = (\beta_0, \beta_1, \cdots, \beta_k)^T$ can be estimated by the maximum likelihood estimated method:

$$L = \prod_{i=1}^{n} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \tag{2}$$

Thus,

$$\log L = \sum_{i=1}^{n} [-\lambda_i + y_i \log(\lambda_i) - \log(y_i!)]$$

$$= -\sum_{i=1}^{n} \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + \cdots + \beta_k x_{ki}) +$$

$$\sum_{i=1}^{n} y_i (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + \cdots + \beta_k x_{ki}) - \sum_{i=1}^{n} \log(y_i!)$$

Notice that the parameters $\beta = (\beta_0, \beta_1, \cdots, \beta_k)^T$ only appear in the first two terms of each term in the summation. Therefore, given that we are only interested in finding the best value for $\beta = (\beta_0, \beta_1, \cdots, \beta_k)^T$ we may drop the $\sum_{i=1}^{n} \log(y_i!)$ and simply write

$$H = -\sum_{i=1}^{n} \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + \cdots + \beta_k x_{ki})$$

$$+ \sum_{i=1}^{n} y_i (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + \cdots + \beta_k x_{ki})$$

This equation has no closed-form solution. However, The negative log-likelihood, $-H$, is a convex function, and so standard convex optimization or gradient descent techniques can be applied to find the optimal value of $\beta = (\beta_0, \beta_1, \cdots, \beta_k)^T$.

III. Using Poisson Regression to Analyze Ship Damage Rates

The **SHIP** data shown in Table 1 represent damage caused by waves to the forward section of certain cargo-carrying vessels. The purpose of the investigation was to set standards for future hull construction. In order to do so, the investigators needed to know the risk of damage associated with five ship types (**TYPE**), year of construction (**YEAR**), and period of operation (**PERIOD**). These three variables are the classification variables. **MONTHS** is the aggregate number of months in service and is an explanatory variable. **Y** is the response variable and represents the number of damage incidents.

In order to do so, the investigators needed to know the risk of damage associated with

- TYPE: Ship type: A–E,
- YEAR: Year of construction: 1960-64, 1965-69, 1970-74, 1975-79,
- PERIOD: Period of operation: 1960-74, 1975-79,
- MONTHS: The aggregate number of months in service.

Y is the response variable and represents the number of damage incidents. The numeric data of ship is shown in table 2.

TABLE I. Ship Data

	Ship type	Year of construction	Period of operation	Logarithm of aggregate months of service	Number of damage incidents
1	A	1960-64	1960-74	4.844	0
2	A	1960-64	1975-79	4.143	0
3	A	1965-69	1960-74	6.999	3
4	A	1965-69	1975-79	6.999	4
5	A	1970-74	1960-74	7.321	6
6	A	1970-74	1975-79	8.118	18
7	A	1975-79	1960-74		
8	A	1975-79	1975-79	7.716	11
9	B	1960-64	1960-74	10.712	39
10	B	1960-64	1975-79	9.751	29
11	B	1965-69	1960-74	10.261	58
12	B	1965-69	1975-79	9.922	53
13	B	1970-74	1960-74	8.863	12
14	B	1970-74	1975-79	9.480	44
15	B	1975-79	1960-74		
16	B	1975-79	1975-79	8.870	18
17	C	1960-64	1960-74	7.072	1
18	C	1960-64	1975-79	6.314	1
19	C	1965-69	1960-74	6.661	0
20	C	1965-69	1975-79	6.516	1
21	C	1970-74	1960-74	6.663	6
22	C	1970-74	1975-79	7.575	2
23	C	1975-79	1960-74		
24	C	1975-79	1975-79	5.613	1
25	D	1960-64	1960-74	5.525	0
26	D	1960-64	1975-79	4.654	0

27	D	1965-69	1960-74	5.663	0
28	D	1965-69	1975-79	5.257	0
29	D	1970-74	1960-74	5.855	2
30	D	1970-74	1975-79	7.097	11
31	D	1975-79	1960-74		
32	D	1975-79	1975-79	7.626	4
33	E	1960-64	1960-74	3.807	0
34	E	1960-64	1975-79		
35	E	1965-69	1960-74	6.671	7
36	E	1965-69	1975-79	6.080	7
37	E	1970-74	1960-74	7.054	5
38	E	1970-74	1975-79	7.678	12
39	E	1975-79	1960-74		
40	E	1975-79	1975-79	6.295	1

TABLE II. SHIP NUMERIC DATA

	Type	Construction	Operation	Log_months_service	Damage_incidents
1	1	60	60	4.844	0
2	1	60	75	4.143	0
3	1	65	60	6.999	3
4	1	65	75	6.999	4
5	1	70	60	7.321	6
6	1	70	75	8.118	18
7	1	75	60		
8	1	75	75	7.716	11
9	2	60	60	10.712	39
10	2	60	75	9.751	29
11	2	65	60	10.261	58
12	2	65	75	9.922	53
13	2	70	60	8.863	12
14	2	70	75	9.480	44
15	2	75	60		
16	2	75	75	8.870	18
17	3	60	60	7.072	1
18	3	60	75	6.314	1
19	3	65	60	6.661	0
20	3	65	75	6.516	1
21	3	70	60	6.663	6
22	3	70	75	7.575	2
23	3	75	60		
24	3	75	75	5.613	1
25	4	60	60	5.525	0
26	4	60	75	4.654	0
27	4	65	60	5.663	0
28	4	65	75	5.257	0
29	4	70	60	5.855	2
30	4	70	75	7.097	11
31	4	75	60		
32	4	75	75	7.626	4
33	5	60	60	3.807	0
34	5	60	75		
35	5	65	60	6.671	7
36	5	65	75	6.080	7
37	5	70	60	7.054	5
38	5	70	75	7.678	12
39	5	75	60		
40	5	75	75	6.295	1

The data provides information on the number and exposure for ship damage incidents, where the exposure was expressed in terms of aggregate number of month service. The risk of ship damage incidents was associated with three rating factors: ship type, year of construction and period of operation. The fitting procedure only involves thirty-four data points because six of the rating classes have zero exposures.

Clementine is the SPSS enterprise-strength data mining workbench [11]. Clementine helps organizations to improve customer and citizen relationships through an in-depth understanding of data. Organizations use the insight gained from Clementine to retain profitable customers, identify cross-selling opportunities, attract new customers, detect fraud, reduce risk, and improve government service delivery.

Clementine's visual interface invites users to apply their specific business expertise, which leads to more powerful predictive models and shortens time-to-solution. Clementine offers many modeling techniques, such as prediction, classification, segmentation, and association detection algorithms. Once models are created, Clementine Solution Publisher enables their delivery enterprise-wide to decision makers or to a database.

The poisson regression modeling is shown on Figure 1. The results of parameter estimates are shown in Table 3.

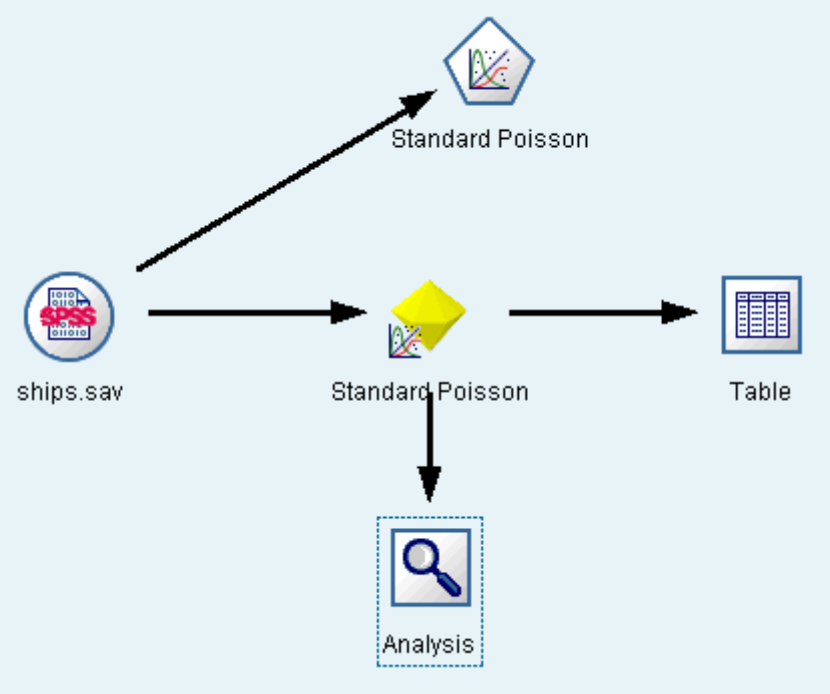

Figure 1. Sample stream to analyze damage rates

TABLE III. PARAMETER ESTIMATES

Parameter	B	Std. Error	95% Wald Confidence Interval	Wald Chi-Square	df	Sig.
intercept	-6.406	0.2828	[-6.960,-5.852]	513.238	1	0.000
Type=5	0.326	0.3067	[-0.276,0.927]	1.127	1	0.288
Type=4	-0.076	0.3779	[-0.817,0.665]	0.040	1	0.841
Type=3	-0.687	0.4279	[-1.526,0.151]	2.581	1	0.108
Type=2	-0.543	0.2309	[-0.996,-0.091]	5.536	1	0.019
Type=1	0					
Construction=75	0.453	0.3032	[-0.141,1.048]	2.236	1	0.135
Construction=70	0.818	0.2208	[0.386,1.251]	13.743	1	0.000
Construction=65	0.697	0.1946	[0.316,1.079]	12.835	1	0.000
Construction=60	0					
Operation=75	0.384	0.1538	[0.083,0.686]	6.249	1	0.12
Operation=60	0					
Scale	1.691					

The parameter estimates table summarizes the effect of each predictor. While interpretation of the coefficients in this model is difficult because of the nature of the link function, the signs of the coefficients for covariates and relative values of the coefficients for factor levels can give important insights into the effects of the predictors in the model.

For covariates, positive (negative) coefficients indicate positive (inverse) relationships between predictors and outcome. An increasing value of a covariate with a positive coefficient corresponds to an increasing rate of damage incidents.

For factors, a factor level with a greater coefficient indicates greater incidence of damage. The sign of a coefficient for a factor level is dependent upon that factor level's effect relative to the reference category.

we can make the following interpretations based on the parameter estimates:

(1) Ship type B [type=2] has a statistically significantly (p value of 0.019) lower damage rate (estimated coefficient of –0.543) than type A [type=1], the reference category. Type C [type=3] actually has an estimated parameter lower than B, but the variability in C's estimate clouds the effect. See the estimated marginal means for all relations between factor levels.

(2) Ships constructed between 1965–69 [construction=65] and 1970–74 [construction=70] have statistically significantly (p values <0.001) higher damage rates (estimated coefficients of 0.697 and 0.818, respectively) than those built between 1960–64 [construction=60], the reference category. See the estimated marginal means for all relations between factor levels.

(3) Ships in operation between 1975–79 [operation=75] have statistically significantly (p value of 0.012) higher damage rates (estimated coefficient of 0.384) than those in operation between 1960–1974 [operation=60].

IV. CONCLUSIONS

In actuarial literature, researchers suggested various statistical procedures to estimate the parameters in claim count or frequency model. In particular, the Poisson regression model, which is also known as the Generalized Linear Model (GLM) with Poisson error structure, has been widely used in the recent years. This paper suggests the Poisson regression models as alternatives for handling ship data. Modeling the raw cell counts can be misleading in this situation because the *Aggregate months of service* varies by ship type. Variables like this that measure the amount of "exposure" to risk are handled within the generalized linear model as offset variables. Moreover, a Poisson regression assumes that the log of the dependent variable is linear in the predictors. Thus, to use generalized linear models to fit a Poisson regression to the accident rates.

ACKNOWLEDGMENT

This work was supported by the Open Project Program of Key Laboratory of Intelligent Computing & Information Processing (Xiangtan University), Ministry of Education (No. 2011ICIP05), Artificial Intelligence of Key Laboratory of Sichuan Province, Jiangsu 333 Project , Qing Lan Project. and the National Natural Science Foundation of China under Grant 51008143.

REFRENCES

[1] A. C. Cameron and K. T. Pravin . Regression Analysis of Count Data. Cambridge, Cambridge University Press, New York , 1998.

[2] J. M. Hilbe, Negative Binomial Regression, Cambridge University Press, Cambridge, England , 2007.

[3] H. Hormann, "Damage stability and safety of Ro-Ro passenger ships state of the art review", Proceedings of the6th International Conference on Stability of Ships and O-cean Vehicles(STAB 97). Varna, Bulgaria, pp. 249-252 ,1997.

[4] J. E. Lawless, "Negative Binomial and Mixed Poisson Regression", The Canadian Journal of Statistics, vol.15, pp.209–225, 1984.

[5] P. McCullagh and J. A. Nelder. Generalized Linear Models, Second Edition, London: Chapman and Hal, 1989.

[6] L. J. Ma, Q. Feng and N. Zhang, "Overseas Research Progress in Theoretical Analysis for Ship Damage Stability", Chinese Journal of Ship Research, vol.7,no.2, pp.9-13, 2012. (in Chinese)

[7] D. Vassalos, O. Turan and M. Pawlowski, "Dynamic stability assessment of damaged passenger/Ro-Ro ships andproposal of rational survival criteria", Marine Technology, vol.34, no.4, pp.241-266, 1997.

[8] W. H. Zhao, R. Q. Zhang, J. C. Liu and Y. Z. Lv, "Semi varying coefficient zero-inflated generalized poisson regression mode", Communications in Statistics - Theory and Methods, vol.44, no.1, pp.171-185, January 2, 2015.

[9] G. Lee, Y. Jeong and S. Kim, "The effect of the built environment on pedestrian volume in microscopic space - Focusing on the comparison between OLS (Ordinary Least Square) and poisson regression", Journal of Asian Architecture and Building Engineering, vol.14, no.2, pp.395-402, 2015.

[10] B. M. Kibria, K. Maring;nsson, and G. Shukur, "A simulation study of some biasing parameters for the ridge type estimation of poisson regression", Communications in Statistics: Simulation and Computation, vol.44, no. 4, pp.943-957, April 5, 2015.

[11] W. Xue, Clementine data mining methods and applications. Publishing House of Electronics Industry, September , Beijing, China. 2010. (in Chinese)