MATEMATIKA, 2013, Volume 29, Number 1b, 11-16

Logistic Regression Model of Hazardous Event on Floating Offshore Units

[1]Aminatul Hawa Yahaya, [2]Noraini Abdullah and [3]Zainodin Haji Jubok
[1,2,3]Mathematics with Economics Programme, School of Science & Technology, Universiti Malaysia Sabah
88400 Kota Kinabalu, Sabah, Malaysia
[1]Malaysian Institute of Marine Engineering and Technology, Universiti Kuala Lumpur
33200 Lumut, Perak, Malaysia
e-mail: [1]aminatulhawa@yahoo.com, [2]norainiabdullah.ums@gmail.com, [3]zainodin@gmail.com

Abstract Logistic Regression (LR) is used to analyze the relationship between non-metric dependent variable and metric or dichotomous independent variables. The overall test of relationship among the independent variables and groups defined by the dependent is based on the reduction in the likelihood values for a model which does not contain any independent variables and the model that contains the independent variables. This difference in likelihood follows a chi-square distribution, and is referred to as the model chi-square. The significance test for the final model chi-square (after the independent variables have been added) is our statistical evidence of the presence of a relationship between the dependent variable and the combination of the independent variables. In this study, a hazardous event such as accident has been analyzed using binary logistic regression. The presence of a relationship between the dependent variable and combination of independent variables is based on the statistical significance of the final chi-square model. The result of the best model shows that number of Injuries/Fatalities, number of Chains and operation mode gives the significant contribution in predicting hazardous event on floating offshore units.

Keywords Binary Logistic Regression (BLR); Binary Dependent Variable; Chi-Square Distribution; Floating Units; Oil and Gas.

2010 Mathematics Subject Classification 62J05

1 Introduction

The Det Norske Veritas (DNV), on behalf of the UK Health & Safety Executive (HSE) is given the responsibility to obtain the accident statistics for offshore fixed and floating units on the UK Continental Shelf (UKCS) since 1999.The main objective of the project is to obtain complete statistics for accidents and incidents having occurred on offshore fixed and floating units engaged in oil and gas exploration and exploitation on the UKCS in the period 1990-2007, including numbers of accidents and incidents with corresponding frequencies per type of installation/rig. The most recent project related to fixed and floating units, *Accident Statistics for Offshore Units on the UKCS 1990 – 2006* was completed in March 2008. As mention in [1], offshore oil and gas exploitation is a hazardous activity. Many safety and security incidents involving offshore installations around the world occurred lately. The Piper Alpha accident, open the eyes on the implementation of risk assessment model focussing on early hazard detection to minimize the chain of reaction toward fatalities, [2].

Based on historical database, prediction on the hazardous event can be made, [3]. From this database, much can be learned for future risk management, on other offshore platforms as well as in other industrial sectors. A better assessment of the risks involved before other accidents occur and should point to a variety of technical and organizational risk management measures. In [4], hazardous event is defined as a situation with a potential for causing harm to human safety, the environment, property or business. It may be a physical situation (e.g. a shuttle tanker is a hazard because it may collide with the production installation), an activity (e.g. crane operations are a hazard because the load might drop) or a material (e.g. fuel oil is a hazard because it might catch fire). In practice, the term “hazard” is often used for the combination of a physical situation with particular circumstances

that might lead to harm, e.g. a shuttle tanker collision, a dropped load or a fuel oil fire. The essence of a hazard is that it has a potential for causing harm, regardless of how likely or unlikely such an occurrence might be. Floating units in this project are defined to comprise drilling, accommodation, and floating production and storage units. In addition, FPSO's, FSU's, and TLP's are classified as "floating units" although they are classified as "fixed installations" by the HSE under the Safety Case Regulations. As accident is the most severe hazardous event, so this study will predict the accident occurrence besides the others events such as incident, near-missed and insignificant by using the logistic regression model.

Logistic regression (LR) allows one to predict a discrete outcome, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. The model is also known as polytomous or polychotomous logistic regression in the health sciences and as the discrete choice model in econometrics, [5]. Generally, the dependent or response variable (*DV*) is dichotomous, such as presence/absence or success/failure. Discriminant analysis is also used to predict group membership with only two groups. However, discriminant analysis can only be used with continuous independent variables. Thus, in instances where the independent variables (*IV*)'s are a categorical, or a mix of continuous and categorical, logistic regression is preferred. The maximum likelihood estimation (MLE) is the most widely-used general method of estimation procedures and is treated as a standard approach to parameter estimation and inference in statistics, [6].

1.1 The Model

The *DV* in logistic regression is usually dichotomous, that is, the DV can take the value 1 with a probability of success $\pi(W)$, or the value 0 with probability of failure 1- $\pi(W)$. This type of variable is called a Bernoulli (or binary) variable. Although not as common and not discussed in this treatment, applications of logistic regression have also been extended to cases where the *DV* is of more than two cases, known as multinomial or polytomous ([7] use the term polychotomous).

The *IV*'s in logistic regression can take any form. That is, logistic regression makes no assumption about the distribution of the *IV*'s. They do not have to be normally distributed, linearly related or of equal variance within each group. The relationship between the *IV*'s and *DV* is not a linear function in logistic regression; instead, the logistic regression function is used, which is the logit transformation of $\pi(W)$:

From the general form of Multiple Linear Regression model:

$$Y = \Omega_0 + \Omega_1 W_1 + \Omega_2 W_2 + \cdots + \Omega_k W_k + u \tag{1}$$

Hence, the logistic model has the following general format:

$$\pi(W) = \frac{e^{(\Omega_0 + \Omega_1 W_1 + \cdots + \Omega_k W_k)}}{1 + e^{(\Omega_0 + \Omega_1 W_1 + \cdots + \Omega_k W_k)}} \tag{2}$$

where, W_j = *j-th* independent variable (single/dummy/interaction/generated/transformed),
Ω_0 = constant term, Ω_j = *j-th* coefficient of *j-th* independent variable W_j , k = number of independent variables, (k+1) = number of parameters in the model, Y = dependent variable and *u* = error term, for j=1, 2, 3, …, k. An alternative form of the logistic regression equation is:

$$\frac{\pi(W)}{1 - \pi(W)} = e^{(\Omega_0 + \Omega_1 W_1 + \cdots + \Omega_k W_k)} \tag{3}$$

$$\text{logit}\left[\pi(W)\right] = \log\left(\frac{\pi(W)}{1 - \pi(W)}\right) = \Omega_0 + \Omega_1 W_1 + \cdots + \Omega_k W_k \tag{4}$$

2 Materials and Methods

2.1 Data Collection

A total of 4205 samples were collected for accidents having occurred on floating offshore units engaged in the oil and gas activities on the UKCS in the period 1980-2007. Floating units in this project were defined as comprising semi-submersibles, jackups, ships and tension-leg platforms engaged in drilling, accommodation, production and storage. Det Norske Veritas AS was contracted to undertake the work. This published dataset, together with the complete report, have been downloaded from the websites of Oil & Gas UK and HSE, http://www.oilandgasuk.co.uk/ and http://www.hse.gov.uk/research/rrhtm/index.htm
respectively.

2.2 Model Building Procedure

STEP 1: All Possible Models
The number of all possible models, N can be calculated by using the formula:

$$N = \sum_{j=1}^{q} \left(C_j^q \right) \tag{5}$$

Where N is the number of possible models generated and q is the number of independent variables excluded the dummy variable and j = 1,2,…, q.

STEP 2: Selected Models
Multicollinearity is the intercorrelation of IV. The higher correlation coefficient will increase the standard error of the beta coefficients and produce assessment of the unique role of each independent resulting in difficult or impossible output. Multicollinearity exist if |Correlation Coefficient| > 0.95. Zainodin-Noraini multicollinearity remedial procedures had been applied and details are explained in [8] and [9].

Next, the coefficient test should be carried out as an elimination procedure of insignificant variable. To justify the removal of the insignificant variable, Wald Test ([10]) should be applied to the possible models upon the completion of all the elimination procedure of insignificant variables.

STEP 3: Best Models
Identification of the best model should be based on Modified Eight Selection Criteria (M8SC). The objective is to determine a model with the lowest value of a criterion statistic. Voglevag [11] suggested that instead of minimizing the value of SSE, modification on 8SC should be made by maximizing likelihood by replacing it with the Deviance statistic value. The calculation of the criterion statistics will be based on the deviance statistics value, number of estimated parameters including constant term (k+1) and the sample size (n). The deviance statistics value can be calculated as follow:

$$G^2 = \text{deviance statistics} = -2\sum_{i=1}^{n} \left[Y_i \ln\left(\hat{p}_i\right) + \left(1 - Y_i\right) \ln\left(1 - \hat{p}_i\right) \right] \tag{6}$$

Table 1 Modified Eight Selection Criteria (M8SC) for best model identification

AIC: $\left(\frac{G^2}{n}\right)e^{\frac{2(k+1)}{n}}$	RICE: $\left(\frac{G^2}{n}\right)\left(1-\frac{2(k+1)}{n}\right)^{-1}$	FPE: $\left(\frac{G^2}{n}\right)\frac{n+k+1}{n-(k+1)}$	SCHWARZ: $\left(\frac{G^2}{n}\right)(n)^{\frac{2(k+1)}{n}}$
GCV: $\left(\frac{G^2}{n}\right)\left(1-\frac{k+1}{n}\right)^{-2}$	SGMASQ: $\left(\frac{G^2}{n}\right)\left(1-\frac{k+1}{n}\right)^{-1}$	HQ: $\left(\frac{G^2}{n}\right)(\ln n)^{\frac{2(k+1)}{n}}$	SHIBATA: $\left(\frac{G^2}{n}\right)\frac{n+2(k+1)}{n}$

The Akaike Information Criterion (AIC), [12] and Finite Prediction Error (FPE), [13] are developed by Akaike. The Generalised Cross Validation (GCV) is developed by [14] while the HQ criterion is suggested by [15]. The RICE criterion is discussed by [16] and the SCHWARZ criterion is discussed by [17]. The SGMASQ is developed by [10] and the SHIBATA criterion is suggested by [18].

STEP 4: Model's Goodness of Fits

The following phase is to check the validation of the best model. There are two tests on the model's goodness of fits. The two tests that have been suggested by [19] are Pearson and Deviance Chi-Square Goodness of Fits tests.

3 Statistical Analysis

3.1 Models Generated

In the development of the LR models for this datasets, Event Category would be the dependent variable (DV) noted by *Y*, whereas Injury/Fatality (X_1) and No. of Chain (X_2) would be the independent variables (IV). As shown in Table 2, Unit Type (D_1) and Operation Mode (D_2) were included as independent dummy variables included in the models. Dummy variables were executed during the calculation of the possible models but included after in the models before next model building procedure was carried out. All possible models in this study when q = 2 (excluded the 4 dummies) is $N = \left(C_1^2\right) + \left(C_2^2\right) = 3$, as shown in Table 3.

Table 2 Data and variable summary

VARIABLE TYPE	VARIABLE NOTATION	VARIABLE INFORMATION	DATA TYPE
Dependent Variable	Y	Event Category	Binary 1 = Accident, 0 = Others
Independent Variables	X1	No. of Injuries/Fatalities	Discrete
	X2	No. of Chains	Discrete
	D1	Unit Type	Binary 1 = Semi-Submersible, 0 = Others
	D2	Operation Mode	Binary 1 = Drilling, 0 = Others

Table 3 Summary of all possible models

NO. OF VARIABLE	SINGLE	MODEL'S NAME	VARIABLES IN THE MODEL
1	2	M1	Y,X1,D1,D2
		M2	Y,X2, D1,D2
2	1	M3	Y,X1,X2,D1,D2

Pearson Correlation analysis verifies that Y has a positive correlation with all four IV's. There is no existence of multicollinearity between IV's. Thus, no elimination should be made among the independent variables.

Next, the coefficient test should be carried out as an elimination procedure of insignificant variable by using the backward elimination as shown by [20]. In this paper, model M1 is selected for the illustration purpose. For M2, the final model named as M1.0.3, with zero variable removed due to multicollinearity and four insignificant variables eliminated.

3.2 Modified Eight Criteria of Model Selection (M8SC)

From 3 possible models generated during the stage of this analysis, all three models have been selected with different G^2 value and number of model parameters. The best model was then chosen from the selected models by using the M8SC based on the majority of least values as shown in Table 4. The best model selected is M3.0.1.

Table 4 Values of Eight Selection Criteria (8SC) for selected models

Selected Model	m = k+1	G^2	AIC	RICE	FPE	SCHWARZ	GCV	SGMASQ	HQ	SHIBATA
M1.0.1	3	2511.806	49.926	50.236	49.932	53.913	50.075	47.393	52.071	49.659
M2.0.1	3	2474.558	49.186	49.491	49.191	53.113	49.333	46.689	51.299	48.923
M3.0.1	4	2196.307	45.243	45.756	45.254	50.122	45.486	42.237	47.853	44.823

3.3 Best Model Verification

To evaluate the adjustment of the best model, Deviance Goodness of fits tests have been carried out in this phase. The test is carried out based on the residual obtained in the best model M3.0.1. The sum of square of deviance statistics = G^2 is 2196.307 and $\chi^2_{critical}$ is $\chi^2_{0.95;4201}$ = 4051.374. Since the G^2 value is less than $\chi^2_{critical}$, the decision is to accept null hypothesis where the best model M3.0.1 is an appropriate model. From here, the conclusion of best model can be made. Thus, best model M3.0.1 is

$$\hat{Y} = -5.262 + 1.858X_1 + 1.759X_2 - 0.559D_2 \quad (7)$$

where X_1 is the number of Injuries/Fatalities, X_2 is the number of Chains and D_2 is the operation mode. The result for number of Injuries/Fatalities and number of Chains are consistent with the occurrence of accident. The coefficient of D_2 gives the negative values. As the D_2 is the dummy variable, the result shows that the most of the accident happens not because of the drilling activities.

4 Conclusion

The Malaysia oil and gas industries expanded tremendously since its early days of the 1900s with capabilities in the exploration and production of oil. There are about 200 platforms at present operated by various operators in Malaysia. It is crucial to create the awareness among the decision makers, managers, technical professionals in Malaysia oil and gas industries about the requirement to have a complete database on the risk assessment on offshore platforms. As corporations have become more familiar with risk assessment database, this database can be used by applying statistical analysis to improve their decision-making processes. The process of model building involves a search for the best way to specify a relationship between a dependent variable and a set of independent variables. The dummy variable technique enables multiple logistic regressions to handle categorical independent variables. For the comparison of this analysis, future work on multiple logistic regressions with variable interaction should be carried out.

References

[1] Holand, P. *Offshore Blowouts Causes and Control*. Houston: Gulf Publishing Company. 1997.
[2] Lord Cullen (The Hon). *The Public Enquiry into the Piper Alpha Disaster*. London: HMSO. 1990.
[3] Andersen, L.B. Stochastic modeling for the analysis of blowout risk in exploration drilling. *Reliability Engineering and System Safety*. 1998. 61: 53-63.
[4] Vinnem, J.E. *Offshore Risk Assessment, 2nd Edition*. London: Springer-Verlag. 2007.
[5] Hosmer, D.W. and Lemeshow, S. *Applied Logistic Regression*. New York: Wiley. 1989.
[6] Agresti, A. *Categorical Data Analysis (2nd edition)*. New Jersey, USA :Wiley. 2002.
[7] *Tabachnick, B.G. and Fidell, L.S. Using Multivariate Statistics. 3rd Edn*. New York: Harper Collins College Publishers. 1996.
[8] Noraini, A., Zainodin, H.J. & Ahmed, A. Improved Stem Volume Estimation using P- Value Approach in Polynomial Regression Models. *Research Journal of Forestry*. 2011. 5(2): 50-65.
[9] Zainodin, H.J, Noraini Abdullah & Yap, S.J. An Alternative Multicollinearity Approach in Solving Multiple Regression Problems. *Trends in Applied Sciences Research.* 2011. 6(1):1241-1255.
[10] Ramanathan, R. *Introductory Econometrics with Applications*, 5th Edition. South- Western Ohio: Thomson Learning. 2002.
[11] Voglerag, B. *Econometrics: Theory and Applications with EViews*. New York: Addison-Wesley. 2005.
[12] Akaike, H. A New Look at Statistical Model Identification. *IEEE Trans. Auto Control* 1974. 19: 716-723.
[13] Akaike, H. Statistical Predictor Identification. *Annals Instit. Stat. Math*. 1970. 22: 203-217.
[14] Golub, G. H., Heath, M. and Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*. 1979. 21: 215–223.
[15] Hannan, E.J. and Quinn, B. The Determination of the Order of an Autoregression. *J. Royal Stat. Society,* 1979. 41(B): 190-195.
[16] Rice, J. Bandwidth Choice for Nonparametric Kernel Regression. *Annals of Stat.* 1984. 12:1215-1230.
[17] Schwarz, G. Estimating the Dimension of a Model. *Annals of Stat.* 1978. 6: 461-464.
[18] Shibata, R. An Optimal Selection of Regression Variables. *Biometrika.* 1981. 68: 45-54.
[19] Kutner, M.H., Neter, J. and Wasserman, W. *Applied Linear Regression Models.* (4th Edition). Singapore: McGraw-Hill,Inc. 2008.
[20] Noraini, A., Zainodin, H.J. and Nigel Jonney J.B. Multiple Regression Models of the Volumetric Biomass. *Wseas Transaction on Mathematics.* 2008. **7**(7): 492-502.