IOP Conf. Series: Materials Science and Engineering **245** (2017) 062012 doi:10.1088/1757-899X/245/6/062012

Logistic Regression in the Identification of Hazards in Construction

Wojciech Drozd [1]

[1] Institute of Building and Transport Management, Tadeusz Kościuszko Cracow University of Technology, Warszawska 24 St., 31-155 Kraków, Poland

wdrozd@ztob.pk.edu.pl

Abstract. The construction site and its elements create circumstances that are conducive to the formation of risks to safety during the execution of works. Analysis indicates the critical importance of these factors in the set of characteristics that describe the causes of accidents in the construction industry. This article attempts to analyse the characteristics related to the construction site, in order to indicate their importance in defining the circumstances of accidents at work. The study includes sites inspected in 2014 - 2016 by the employees of the District Labour Inspectorate in Krakow (Poland). The analysed set of detailed (disaggregated) data includes both quantitative and qualitative characteristics. The substantive task focused on classification modelling in the identification of hazards in construction and identifying those of the analysed characteristics that are important in an accident. In terms of methodology, resource data analysis using statistical classifiers, in the form of logistic regression, was the method used.

1. Explanation of the selection of the topic and the variables for analysis

A large number of accidents in the construction industry and the resulting fatalities causes that the issue of safety on a construction site is of particular importance, is very current [1, 2, 3, 4] and requires constant attention. This article analyses data relating to construction sites and construction accidents. These are actual data, recorded in official reports of district labour inspectorates. Based on inspection sheets (quantity: 339), drawn up by the inspectors of the Regional Labour Inspectorate in Krakow (Poland), patterns of accidents have been identified and profiled and classification modelling carried out to identify threats on construction sites. The time interval used for the study covered the years 2014-2016. This choice was determined by the timeliness of data on work safety on construction sites. The acquired data were stored in the form of the following sheet:

- *"Inspections 2014 - 16"* (Table 1.1a to 1.1c - limited to showing 2 of 339 observations),
-

Table 1a. Inspections 2014 – 16.

LP	Lpr	Lzat	PDP										ZTB											
			PDPis		PDPbl		PDPor		PDPoi		PDPom		ZTBob		ZTBmn		ZTBsm		ZTBze		ZTBzpe		ZTBhig	
			b	u	b	u	b	u	b	u	b	u	b	u	b	u	b	u	b	u	b	u	b	u
1	8	8	4	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	2	0	0	1
2	32	32	2	2	1	0	1	0	4	3	1	0	0	1	1	0	0	0	1	0	0	0	0	0

Table 1b. Inspections 2014 – 16 (cont.).

DB						**SP**				**RZ**						**PW**							
DBbioz		DBibwr		DBnad		SPln		SPsp		RZzś		RZzw		RZsu		PWoz		PWzp		PWzo		PWoi	
b	u	b	u	b	u	b	u	b	u	b	u	b	u	b	u	b	u	b	u	b	u	b	u
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0
0	0	2	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0

Table 1c. Inspections 2014 – 16 (cont.).

RU												**MU**					
RUod		RUpos		RUbal		RUkom		RUpwm		RUkot		MUos		MUins		MUudt	
b	u	b	u	b	u	b	u	b	u	b	u	b	u	b	u	b	u
0	0	2	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0
1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0

where:

b – Missing,
u – Shortfall,
Lzat – Number of employees,
Lpr – Number of people working,
PDP – Preparation for work,
PDPis – Job training,
PDPbl – Medical examination,
PDPor – Work clothing and footwear,
PDPoi – Personal protection equipment,
PDPom – Licence to operate construction machinery and equipment subject to approval by inspection authority,
PDP – Preparation for work,
PDPis – Job training,
PDPbl – Medical examination,
PDPor Work clothing and footwear,
PDPoi – Personal protection equipment,
PDPom – Licence to operate construction machinery and equipment subject to approval by inspection authority,
ZTB – Siteworks,
ZTBob – Securing the site and the work zone against unauthorised access (fencing, gates, etc.)
ZTBmn – Securing and marking hazardous zone and place,
ZTBsm – Stacking and/or storage of materials,
ZTBze – Anti-shock protection for electrical systems and devices (including protection against mechanical damage),
ZTBzpe – Measurement of the effectiveness of anti-shock protection against direct and indirect contact,
ZTBhig – Toilets,
DB – Safety documentation,
DBbioz – Health and safety plan,
DBibwr – Safe working instructions,
DBnad – Supervision of particularly hazardous work,
SP – Work areas and work processes,
SPln – Carrying out works in the vicinity of active overhead power lines,

SPSP – Securing permanent work areas against falling objects and weather conditions,
RZ – **Earthworks and excavations,**
RZzś – Securing excavation walls,
RZzw – Entry to the excavation,
RZsu – Storage of excavated material and materials in the vicinity of the excavation,
PW – **Work at heights,**
PWoz – Use in the work areas of collective protection measures against falls from height (e.g. open edges of floors, roofs),
PWzp – Securing passageways to work areas and staircases against falls from height,
PWzo – Securing holes in ceilings, exterior walls, lift shafts, etc. from the possibility of persons falling into them,
PWoi – Securing a worker against falling from height using personal protective equipment,
RU – **Scaffolding,**
RUod – Documented inspection of scaffolding by authorised persons,
RUpos – Scaffolding foundation,
RUbal – Protective railings on scaffolding's working platforms,
RUkom Vertical circulation cores on the scaffolding,
RUpwm – Working surfaces of scaffolding filled with catwalks,
RUkot – Anchoring the scaffolding to fixed elements of a building in accordance with documentation,
MU – **Machinery and equipment,**
MUos – Guards and elements securing dangerous machinery and equipment,
MUins – Safety instructions for machinery and equipment,
MUudt – Decisions of the Office of Technical Inspection (UDT) permitting the operation of equipment,

2. Logistic regression

2.1. Introduction

Logistic regression is a method using which researchers can model a two-state (binary) dependent variable [5]. During the construction of the model, one of the states of the dependent variable is coded as 0, and the other as 1. Usually, as the value of 1 encodes the state, which is more interesting to us or desired by us. A logistic model is based on the logistic function in the following form:

$$f(z) = \frac{e^z}{1+e^z}, \tag{1}$$

where:

$Z \in (-\infty,+\infty)$

The course of the logistic function is shown in the graph below:

It can be seen that the shape of the logistic function resembles a stretched letter S and its values are within the range of 0 to 1. Initially, changes in the function are minimal and oscillate close to 0, and when the threshold is reached rapidly increase to 1. Using this method, it can be modelled events characterized by a change of the rate of occurrence after reaching a certain threshold value.

More specifically, the logistic model is defined as follows:

$$P(x) = \frac{\exp(b_0+\sum_{i=1}^{n} b_i x_i)}{1+\exp(b_0+\sum_{i=1}^{n} b_i x_i)}, \tag{2}$$

where:

$P(x)$ Means the probability that the predicted variable will have a value of 1,

$a_1, \ldots, a_n$ Are regression coefficients,
$x_1, \ldots, x_n$ Are independent variables (can be both quantitative and qualitative).

As in any regression model, also here we are trying to estimate the regression coefficients and fit the best model, based on the values of a group of data.

The probability of occurrence of the modelled event for an object described using attributes $x_1, \ldots, x_n$, can be calculated using a logistic function from a linear combination (appropriately weighted sum) of the value of attributes.

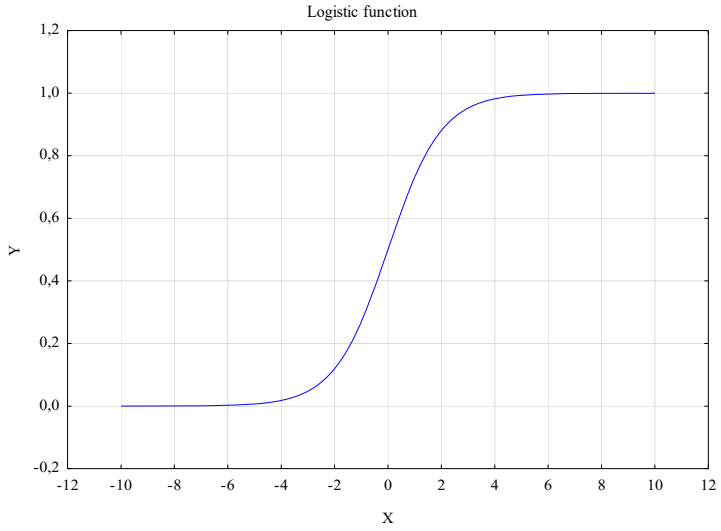

Figure 1. Shape of the logistic function

2.2. Relative risk and odds ratio

Relative risk determines the degree of an increase (decrease) of the probability of occurrence of an event, after changing a factor, and is calculated using the formula:

$$Rw = \frac{p_A}{p_B}, \tag{3}$$

Where p_A and p_B represent the respective probability of occurrence in each group corresponding to the levels of the factor examined.

Odds ratio is calculated based on the classification table 2x2, which shows the observed distribution of cases for a dichotomous random variable:

$$Is = \frac{p_A/(1-p_A)}{p_B/(1-p_B)}, \tag{4}$$

Where p_A and p_B are defined as in the case of relative risk.

The difference between relative risk and odds ratio results from the difference between odds and risk. If out of 100 people involved in accidents 20 dies, the risk of death is 20/100 or 0.2, and the odds that an accident victim dies is 20:80 or 1:4 (0.25). In cases where either result is very rare, relative risk and odds ratio have very similar values. It's also worth knowing that the value of the odds ratio is always more distant from unity than relative risk, thus the causal relationship may seem stronger.

The odds ratio plays an important role in the interpretation of the assessment of logistic regression parameter and is normally taken into account when reporting the results of such models. In particular, if explanatory variables are type 0-1, logistic regression coefficients tell us how the occurrence of a given factor reflects in the odds of the modelled event.

2.3. Model design

For logistic regression, the same data from *Inspections in 2014 - 16* was used. The dependent variable was variable *"Accident"* encoding the occurrence of an accident, and the independent variables was

variable *Lpr* and other variables 0-1 (excluding fixed variables, such as *PDPom u, DBbioz, SPln, SPln u).* Forward selection was used for variable selection. It involves the creation of the best univariate model, and then gradually expanding it by other variables - until we are not able to add a predictor that would significantly improve the match. Assessment of the significance of each variable is based on the Wald test [6]. This method, in addition to free expression, of all the variables selected these: *SPsp u, PWoz, MUos u, PWoi u, RUbal u, DBibwr u, RUbal, ZTBzpe u.* The proposed model is as follows.

Accident ~ SPsp u + PWoz + MUos u + PWoi u + RUbal u + DBibwr u + RUbal + ZTBzpe u,

which means that the probability of an accident depends on:

SPsp u - Securing permanent work areas against falling objects and weather conditions - shortfall,
PWoz - Use in the work areas of collective protection measures against falls from height (e.g. open edges of floors, roofs) - missing,
MUos - Guards and elements securing dangerous machinery and equipment - missing,
PWoi u - Securing a worker against falling from height using personal protective equipment - shortfall,
RUbal u - Protective railings on scaffolding's working platforms,
DBibwr u - Safe working instructions - shortfall,
RUbal - Protective railings on scaffolding's working platforms - missing,
ZTBzpe u - Measurement of the effectiveness of anti-shock protection against direct and indirect contact - shortfall.

The table 2 shows the results of the logistic regression model. Significant results were marked in red.

Table 2. Results of the logistic regression model

	Evaluation	Standard Error	Wald Stat	GU upper 95%	GU lower 95%	p	Odds ratio	Confidence OR -95%	Confidence OR +95%
Free expression	-7.207	1.023	49.590	-9.213	-5.201	0.000	0.001	0.000	0.006
SPsp u	1.955	0.642	9.264	0.696	3.214	0.002	7.065	2.006	24.881
PWoz	2.925	0.617	22.462	1.715	4.134	0.000	18.630	5.558	62.449
MUos u	2.458	0.668	13.536	1.149	3.767	0.000	11.680	3.153	43.261
PWoi u	1.702	0.609	7.811	0.508	2.895	0.005	5.483	1.662	18.082
RUbal u	1.582	0.607	6.783	0.391	2.772	0.009	4.864	1.479	15.997
DBibwr u	-3.961	1.463	7.327	-6.829	-1.093	0.007	0.019	0.001	0.335
RUbal	2.452	0.717	11.709	1.048	3.857	0.001	11.616	2.851	47.327
ZTBzpe u	1.270	0.632	4.041	0.032	2.508	0.044	3.561	1.032	12.286

The table consists of the following columns:

- ***Evaluation*** - defines the regression coefficient for each variable,
- ***Standard Error*** - returns the standard error for the specified coefficients from the Evaluation column.
- ***Wald Stat*** - the value of the Wald statistic allowing to test the hypothesis whether the true value of the regression parameter evaluation is different from 0,

- ***GU upper 95%* and *GU lower 95%*** - sets the upper and lower limit of the confidence interval for the designated parameter. If the range does not cover zero, we assume that the actual value of the regression parameter evaluation is different from zero,
- ***p*** - test probability. If p<0.05, then we assume that the actual value of the regression parameter evaluation is different from zero and this variable in the model is significant,
- ***Odds ratio*** - the ratio of chance occurrence of the event in one group to the chance of its occurrence in a different group,
- ***Confidence OR-95%* and *Confidence+95%*** - the upper and lower confidence interval for the odds ratio.

The logistic function takes the form:

$$P(\boldsymbol{Accident} = \text{YES}) = \exp(-7.207 + 1.955 * \boldsymbol{SPsp\ u} + 2.925 * \boldsymbol{PWoz} + 2.458 * \boldsymbol{MUos\ u} + 1.702 * \boldsymbol{PWoi\ u} + 1.582 * \boldsymbol{RUbal\ u} - 3.961 * \boldsymbol{DBibwr\ u} + 2.452 * \boldsymbol{RUbal} + 1.270 * \boldsymbol{ZTBzpe\ u}) / (1 + \exp(-7.207 + 1.955 * \boldsymbol{SPsp\ u} + 2.925 * \boldsymbol{PWoz} + 2.458 * \boldsymbol{MUos\ u} + 1.702 * \boldsymbol{PWoi\ u} + 1.582 * \boldsymbol{RUbal\ u} - 3.961 * \boldsymbol{DBibwr\ u} + 2.452 * \boldsymbol{RUbal} + 1.270 * \boldsymbol{ZTBzpe\ u}))$$

Therefore, if the variable *SPSP* has the value of 1 (i.e. the inspection found shortfalls in the given area), the linear part of the model increases by 1,955 and, consequently, the odds ratio of an accident increases by a factor of 7.065 (we can read this from the Odds Ratio column in Table 2.).

2.4. Assessment of fitness

One of the results of the logistic regression model is the probability of membership of particular cases to the modelled class (an accident occurred or not). The values of this probability can be used to support the decision-making process. The best decision rule should guarantee the best results - the minimum number of errors. To precisely define the relevant criterion, we introduce rule quality measures: **specificity** and **sensitivity**. Sensitivity tells us what percentage of objects actually belonging to the indicated condition (sites where accidents took place) was classified correctly by the model. Specificity tells us what fraction of the objects belonging to the non-indicated condition was correctly classified by the model. The change of the values of sensitivity and specificity, under the impact of changing the cut-off point, can be observed in the graph below (Figure 2.).

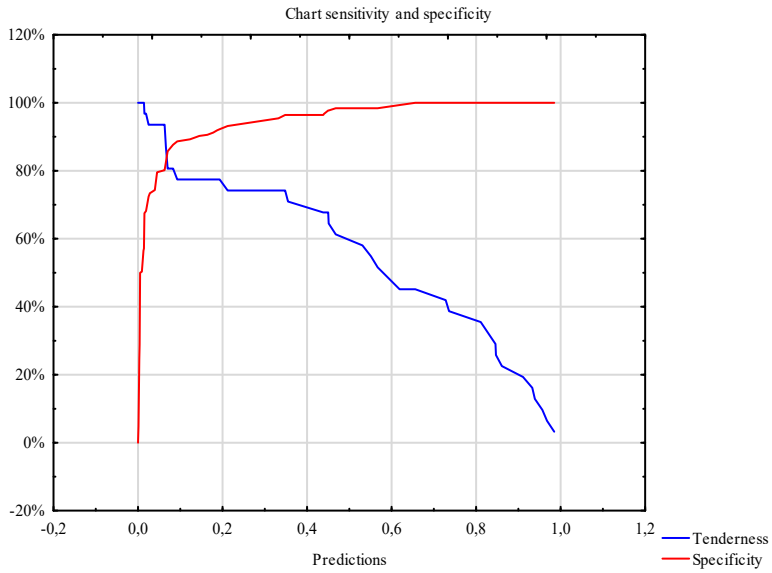

Figure 2. Graph of sensitivity and specificity for the fitted model

As the cut-off point we understand the point above which we assume the occurrence of the modelled class and below we assume the opposite event. A good decision is one that maximizes both values. These graphs are used to construct ROC curves, which illustrate the relationship between sensitivity and specificity (Figure 3.).

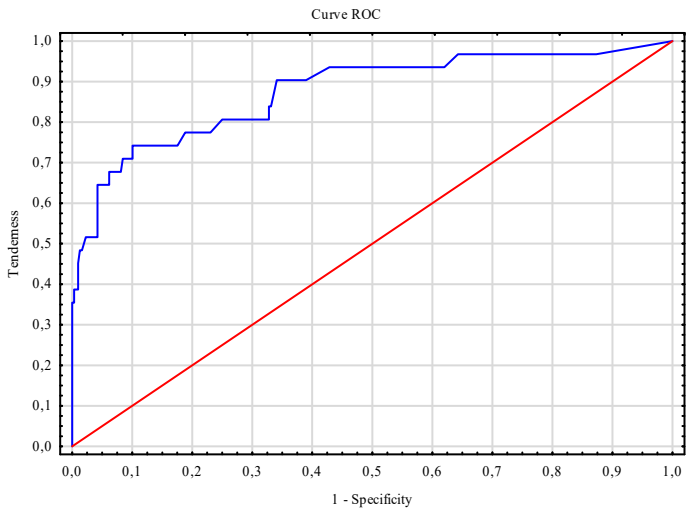

Figure 3. ROC curve for the selected model

ROC curves are often used as a tool to compare models with each other. A very popular approach is calculating the area under the ROC curve graph, denoted as AUC. The value of the AUC index takes values from the interval [0,1]. The bigger it is, the better the model, whereby the AUC of less than 0.5 means that the model is worse than a random prediction of an event.

Calculation of the prediction error on the basis of a learning error does not give a reliable picture of the predictive ability of the model. A prediction error calculated on this set underestimates the actual value of the prediction error, which we might expect when using a model to predict new data. We are able to assess actual ability only on the basis of a set, which was not involved in the process of estimating model parameters. A good method is a v-fold cross-validation, which consists of the division of the set of data into v number of subsets, from which all sets except one are used to estimate the model, for example, first the training set (on the basis of which the model is created) is 1,2,3,4, and the test set (used to evaluate the prediction) is set 5, then the training set is 1,2,3,5, and 4 is the test set, etc. The global evaluation of the prediction error in this case means averaging errors from individual sets.

For the proposed model, the AUC value was 0.944 (±0.0181) in the learning sample, and in a 5-fold cross-validation, we received 0.875 (±0.0391). On this basis we concluded that the model is not excessively fitted to the data and maintains high predictive power for new data.

To determine the goodness of fit of a logistic regression model, we use the Hosmer-Lemeshow test. The null hypothesis and the alternative hypothesis are:

H_0: The observed and expected frequency of events are the same,

H_a: The observed and expected frequency of events are different.

If the predicted and observed values are close enough, it can be concluded that the model is well fitted, because in this case we expect the lack of significance of the test, and namely a situation in which $p > 0.05$ (where $\alpha=0.05$ is the fixed significance level). For the proposed, fitted model, the value of $p=0.219$, therefore there is no basis to reject the null hypothesis.

The most important practical results of the logistic regression model created (Table 2.) is the list of predictors included in the model and evaluation of their parameters (and their corresponding OR) as well as AUC statistics measuring the goodness of fit.

3. Summary and conclusions

Accidents at work, as events taking place on the site, are random occurrences, difficult or impossible to predict. Therefore, their study and identification of relationships between traits characterising them is not easy. This article attempts a scientific analysis of the multi-dimensional set of data on the construction site, which is characterised by accidents during the execution of works. The subject of the analysis were construction sites located in the Lesser Poland Province, where accidents at work took place in the years 2014 - 16, or were the subject of a routine inspection conducted by the Regional Labour Inspectorate in Krakow. The main aim of this article was to explore the impact of selected characteristics of the construction site on the safety risk in the execution of works, expressed by the behaviour of the employee and the type and status of conditions for the occurrence of an accident.

The following conclusions have been formulated:

- Logical modelling is a good tool for classifying threats to safety in the construction industry, generated by the characteristics of the construction site. Logistic models allow us to discover analytical dependencies, expressing the state of the depicted reality.
- Logistic regression is a method by which we can model the two-state (binary) dependent variable *"Accident"*. During the construction of the model, one of the states of the dependent variable is coded as 0, and the other as 1. Usually, as the value of 1 encodes the state, which is more interesting to us or desired by us.
- In logistic regression, the dependent variable was variable *"Accident"* and independent variables: *SPsp u, PWoz, MUos u, PWoi u, RUbal u, DBibwr u, RUbal and ZTBzpe u*.
- The most important results in practice of the logistic regression model created are:
  - *SPsp u* – Securing permanent work areas against falling objects and weather conditions - shortfall,
  - *PWoz* – Use in the work areas of collective protection measures against falls from height (e.g. open edges of floors, roofs) - missing,
  - *MUos* – Guards and elements securing dangerous machinery and equipment - missing,
  - *PWoi u* – Securing a worker against falling from height using personal protective equipment - shortfall,
  - *RUbal u* – Protective railings on scaffolding's working platforms,
  - *DBibwr u* – Safe working instructions - shortfall,
  - *RUbal* – Protective railings on scaffolding's working platforms - missing,
  - *ZTBzpe u* – Measurement of the effectiveness of anti-shock protection against direct and indirect contact - shortfall.

The tests carried out and the test results presented in the article make it possible to evaluate the relationship between characteristics of the site and work safety during the execution of works. The results can be used by designers and site managers at the stage of preparation of information on safety and the health and safety plan, as well as the stage of designing the site set up and subsequent management of the implementation of work.

In particular, these results may be useful in developing various kinds of activities aimed at improving safety in the construction industry and reducing the number of accidents. The results indicate factors, the involvement of which in generating hazards at work is essential. Analysis confirms the thesis that among the characteristics of the construction site, the material factor is the strongest in the context of creating hazards during the execution of works.

References

[1] B. Hoła, „Modeling qualitative and quantitative accidents in construction," *Oficyna Wydawnicza PWr*, Wrocław 2008.

[2] B. Hoła , "Methodology of hazards identification in construction work course," *Journal of Civil Engineering and Management,* No. 4, 2011, p. 577-585.

[3] E. Błazik – Borowa, „Safety in construction work," *Politechnika Lubelska,* Lublin 2015.
[4] W. Drozd, „Characteristics of the construction site in terms of occupational hazards," *Czasopismo Inżynierii Lądowej, Środowiska i Architektury, Politechnika Rzeszowska,* January – March 2016, pp. 165 – 172.
[5] D. W. Hosmer, S. Lemeshow, *Applied Logistic Regression*, Second Edition, John Wiley &Sons, 2003, p. 16.