

Apply Logistic Regression to Analyze   
Singapore Workplace Injury Data

EBS5101 Foundation of Business Analytics – Assignment 1

Submitted by:

Ma Min(A0163305N), Muni Ranjan(A0163382E),

Pradeep Kumar (A0163453H), Zheng Weiyu (A0163412R)

[Objective 3](#_Toc474495833)

[Problem Statement 3](#_Toc474495834)

[Exploratory Analysis 4](#_Toc474495835)

[Determination of Key Factors 5](#_Toc474495836)

[Iteration 1 5](#_Toc474495837)

[Observations 5](#_Toc474495838)

[Iteration 2 6](#_Toc474495839)

[Observations 7](#_Toc474495840)

[Iteration 3 7](#_Toc474495841)

[Covariance Test 8](#_Toc474495842)

[Observations 8](#_Toc474495843)

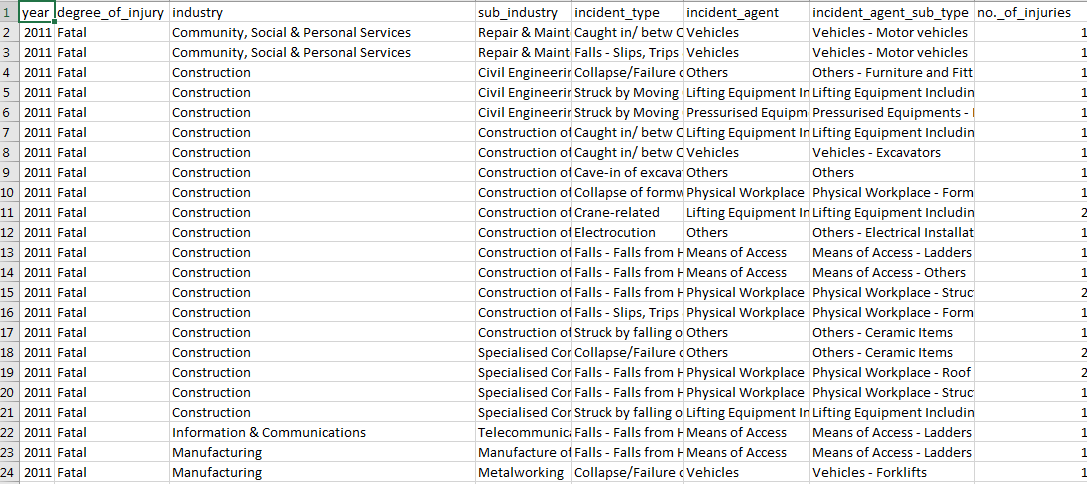
[Conclusion 9](#_Toc474495844)

[References 9](#_Toc474495845)

# Objective

The objective of this report is to explain the data exploration learning technique using Logistic Regression. We have selected the data “Workplace Injury by types” provided by Singapore government. We would like to identify the relationship between different factors provided in the data. We want to find out if there is an independent variable which could be predicted based on one or more dependent variable.

Below is the quick snapshot of data:



Source: [data.gov.sg](https://data.gov.sg/dataset/54a2cbdb-a9b5-46cc-a2de-16ade7212050/resource/109b3957-8826-4d92-b47e-01f58ec22cf3)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

|  |  |
| --- | --- |
|  |  |

# Problem Statement

We found the following information about the data:

* There are total 8 variables provided in this dataset.
* Total number of observations are 16374
* Unique values under the **no.\_of\_injuries** varies from **1 to 261**. This indicates that for a typical accident number of workers injured from 1 to 261
* There are 3 types of degree\_of\_injuries - FATAL, MAJOR, MINOR

We would like to explore the following :

1. Is there any relation between single injury or group injury with other variables?
2. Can we predict the number of Injuries (single vs multiple) based on statisctically signficant variables?

To conduct this analysis we converted the injury\_count to a boolean variable

* 0: Represents 1 people involved in accident
* 1: Represents more than 1 people involved in accident

For all the attributes, an initial exploratory analysis was done. Bar charts were used to find out the relevance of the variables . Since there were no null values, no reduction of data was required.

|  |  |
| --- | --- |
| Exploratory Analysis We first identified the major attributes which could help us create the model for predicting the group injury. For this we compared the unique values in each variable and found out the following:   * Year has no effect on our model. Hence we dropped this variable * We converted different factors under columns “industry”, “incident\_type”, “incident\_agent” to separate variables. * Next we bar plotted the all variables against the “Number of Injury”. Some of them are shown in the images. |  |

Figure 1: Number of Injuries vs Industrial Machine

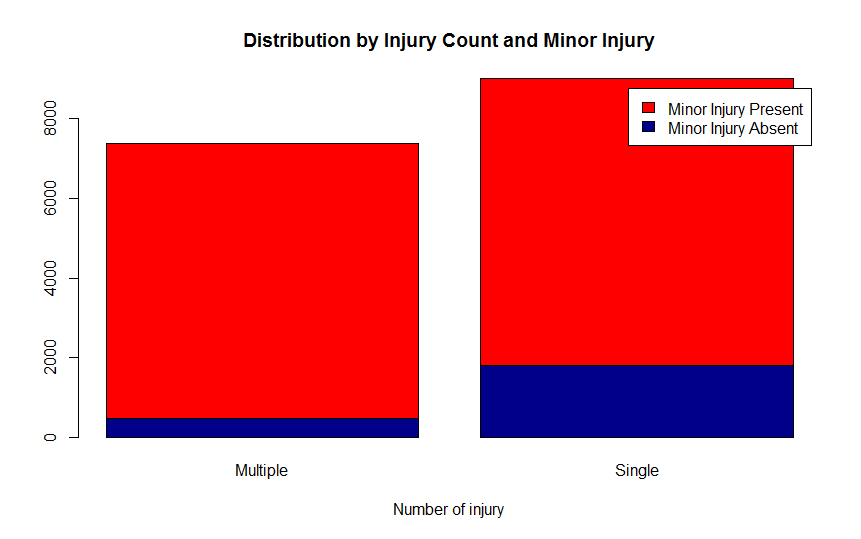
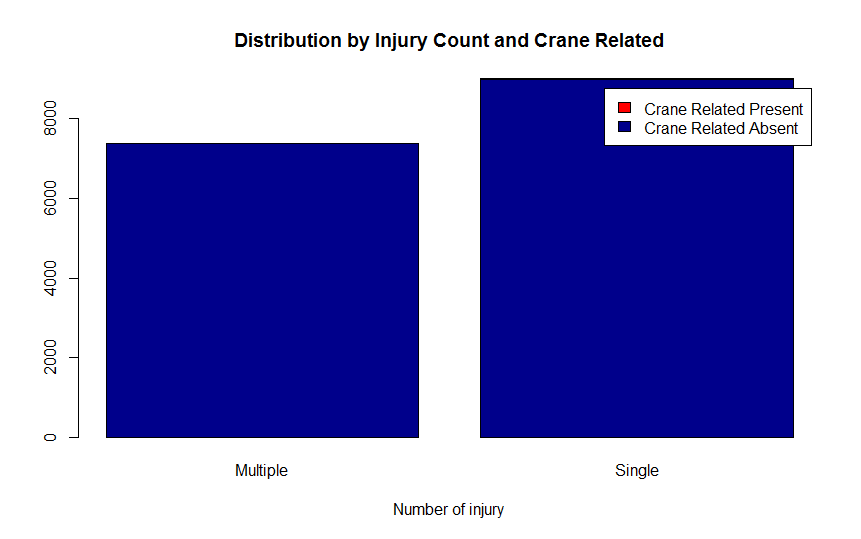


Figure 2: Number of Injuries vs Minor Injury Figure 3: Number of Injuries vs Crane Related injuries

Figure (2) and (3) shares the example of variables which are not useful for preparing the model as they are either not present at all in case of Single and Multiple Injury or they are equally available in both kinds of injuries. Hence they are not considered appropriate in logistic regression.

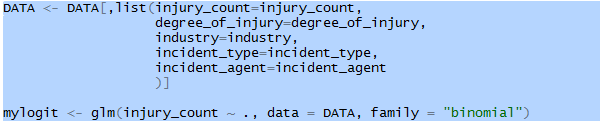
To further confirm our understanding let’s run our first model which takes all the parameters

# Determination of Key Factors

## Iteration 1

In the first run, we considered the most of the variables in degree\_of\_injury, industry, incident\_type, incident\_agent.

The sample code written in “R” is as shown below:



\*\*please note ***injury\_count ~ .*** above :- “.” Represents all other variables in datatable

Below is the output of our model run:

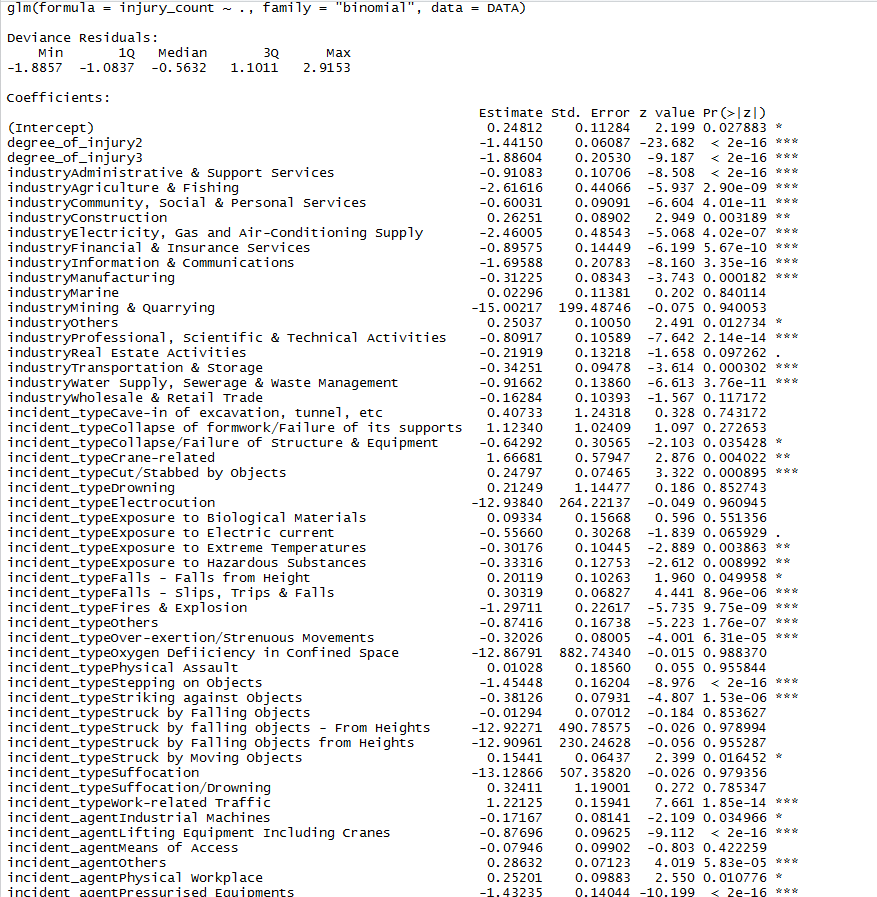


Figure 4: First Iteration of our model

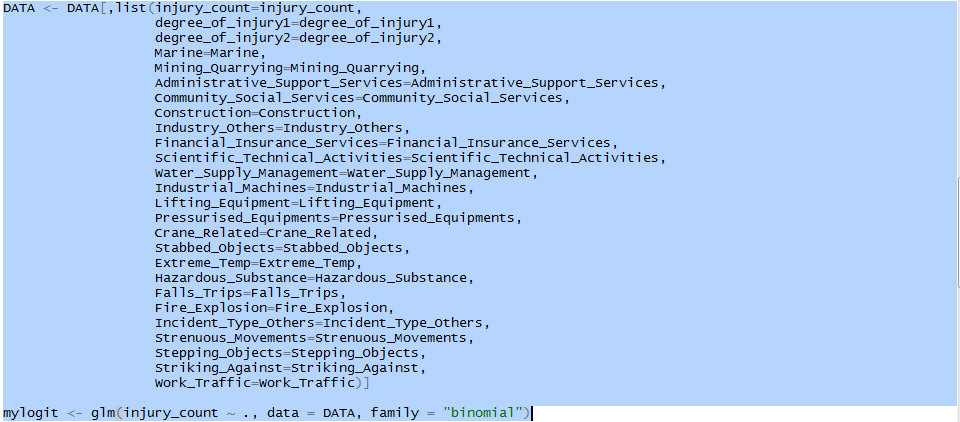
## Observations

* “Fatal” injury type has been filtered in ‘R’ output as it has very low significance in predicting the group injury
* There are multiple other factors such as “industryMining & quarrying”, industryMarine etc. which have very low significance based on alpha levels, hence they can also be dropped from the model.

Now After removing the factors with low significance let’s re-run the model.

## Iteration 2

Below Variables were taken in to consideration in our second iteration. We converted all the factors under columns to separate predictor variables as shown in the sample Code:



Here is the output of our model run:

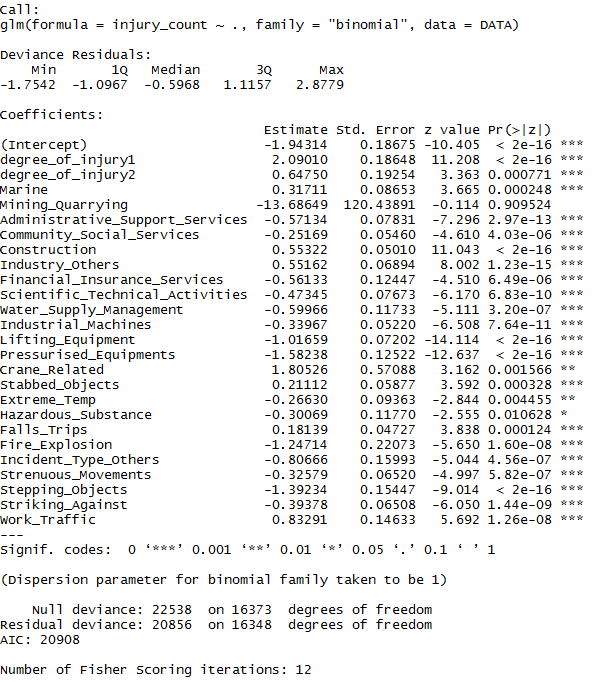


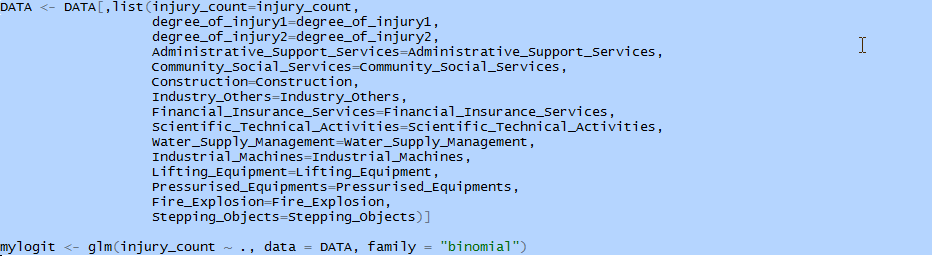
Figure 5: Second Iteration of our model

## Observations

* Based on the P Value we still have some parameters which have low significance and could be dropped further from our model.

## Iteration 3

Below Variables were taken in to consideration:



Below is the output of our model in third iteration:

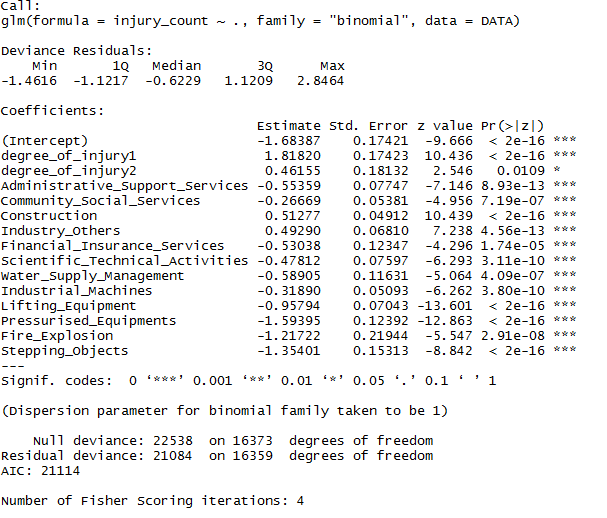


Figure 6: Third Iteration of our model

## Covariance Test

After we sort out the variables on the basis of significance, we also ran the Covariance test to identify if there is any interrelation exist between the predictor variables

Figure (7) below shows the Correlation Matrix chart:

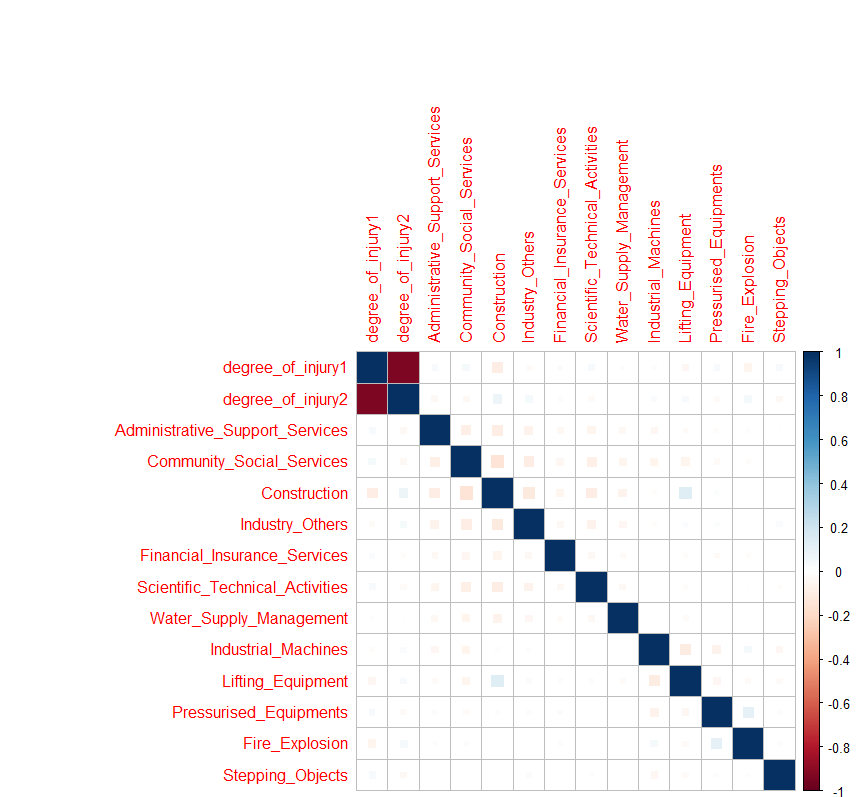


Figure (7): Correlation Matrix between predictor variable

We can observe in the matrix that dgree\_of\_injury1 and degree\_of\_injury2 shows inverse relationship. So we tried the different run by dropping one of them. The results were not affected by this.

Finally, we randomly divided the data in to **train\_set**,**test\_set in** 70:30 ratio andran our model.

Below shows the Confusion Matrix for both.

## Observations

Confusion Matrix (Train Set)

0 1

0 3802 2497 (accuracy approx. 60.36%)

1 2096 3066 (accuracy approx. 59.39%)

Confusion Matrix (Validation Set)

0 1

0 1672 1028 (accuracy approx. 61.92%)

1 0909 1304 (accuracy approx. 58.92%)

# 

# Conclusion

The third iteration of our model showed the better results of all other iterations. We used the train set and test set to validate our model.

To further improve the model, we would require more sample data and fine tune predictor variables accordingly.

# References

1. Lecture notes @ [ivle](https://ivle.nus.edu.sg/v1/File/Student/default.aspx?CourseID=1a7b5b36-ba60-41dc-9aeb-54969d887b23&WorkbinID=1a3a22a0-6283-4dd2-82f9-0cf9cfb5df04&FolderID=34239dca-84c0-4108-bea7-23c24ca598ed)
2. Logistic Regression - <https://www.youtube.com/watch?v=zAULhNrnuL4>