

Apply Logistic Regression to Analyze   
Singapore Workplace Injury Data

EBS5101 Foundation of Business Analytics – Assignment 1

Submitted by:

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

Pradeep Kumar (A0163453H), Zheng Weiyu (A0163412R)

[Objective 3](#_Toc474362811)

[Problem Statement 3](#_Toc474362812)

[Exploratory Analysis 4](#_Toc474362813)

[Determination of Key Factors 5](#_Toc474362814)

[Iteration 1 5](#_Toc474362815)

[Observations 5](#_Toc474362816)

[Iteration 2 6](#_Toc474362817)

[Observations 6](#_Toc474362818)

[Iteration 3 7](#_Toc474362819)

[Covariance Test 7](#_Toc474362820)

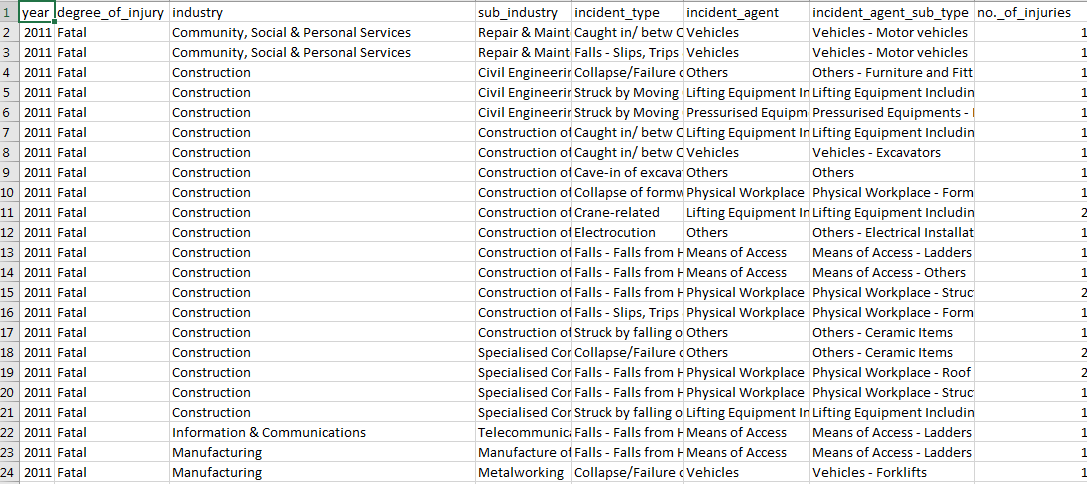
[Observations 8](#_Toc474362821)

[Conclusion 8](#_Toc474362822)

# Objective

The objective of this report is to explain the team work done to apply data exploration learning technique. 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

After loading the dataset from csv, 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 injury type based on statisctically signficant variables?

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

* 0: Represents 1 or 2 people involved in accident
* 1: Represents more than 2 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 * Next we bar plotted the different factors variable against the “Number of Injury” as you can see some examples on the right and below. |  |

Figure 1: Scatterplot of different variables with injury\_count

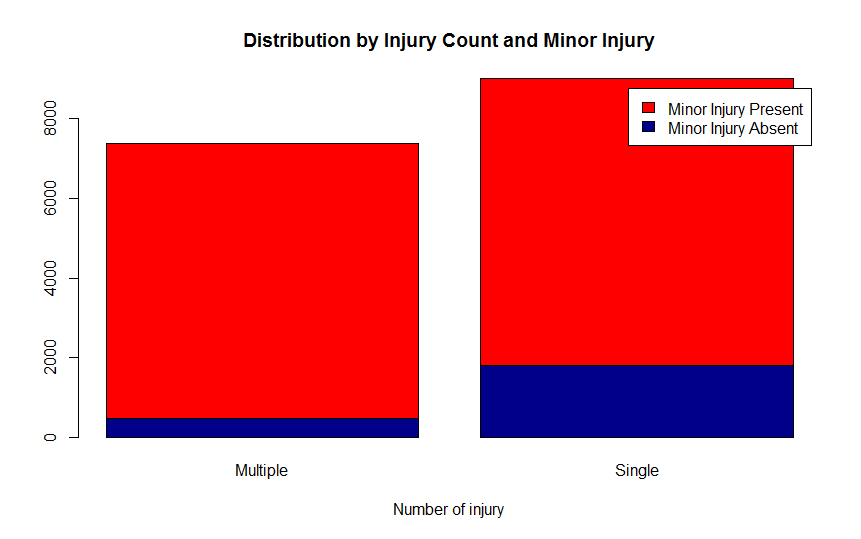
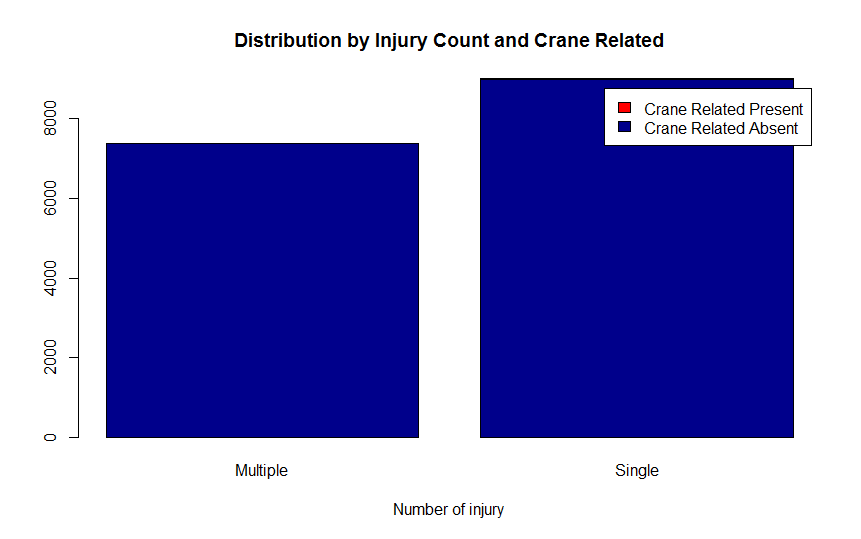


Figure 2: Number of Injuries vs Minor Injury Figure 3: Number of Injuries vs Crane Related injuries

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 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

We ran the logit function on our data. Below is the summary of our logit run:

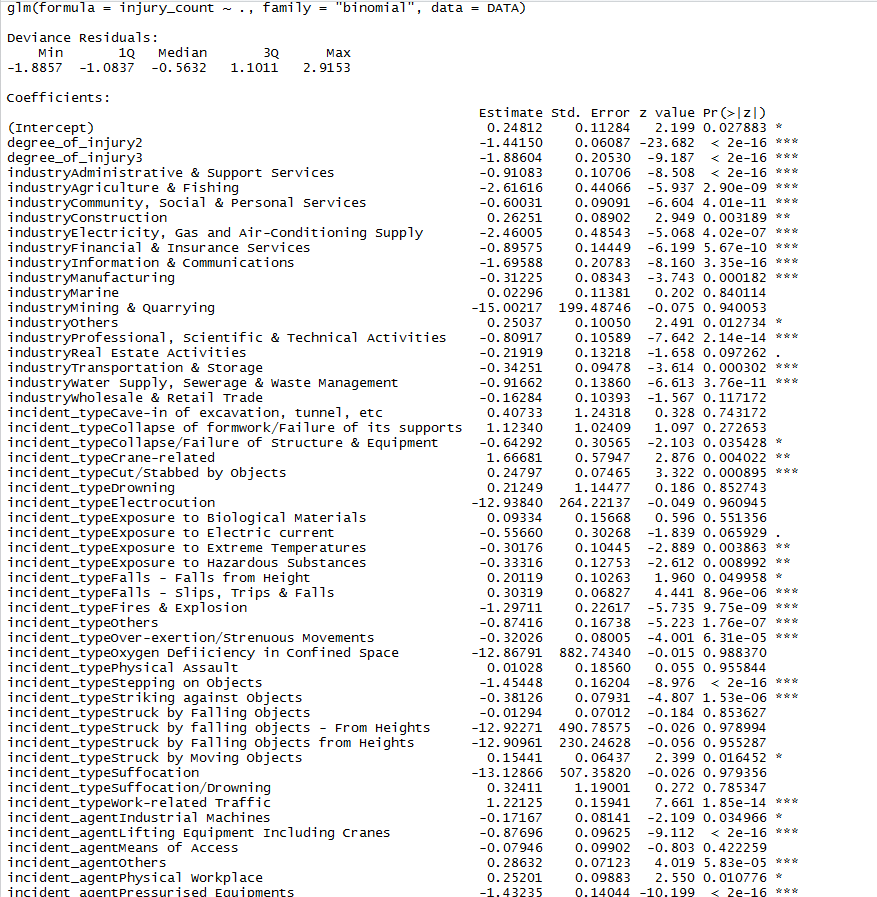


Figure 4: First Iteration of our model

## 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 observe the output as generated by R

## Iteration 2

Below Variables were taken in to consideration:

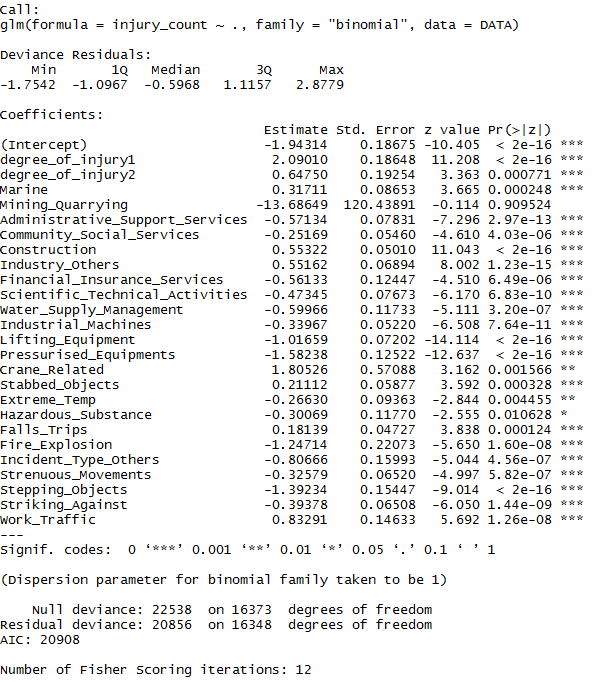


Figure 5: Second Iteration of our model

## Observations

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

## Iteration 3

Below Variables were taken in to consideration:

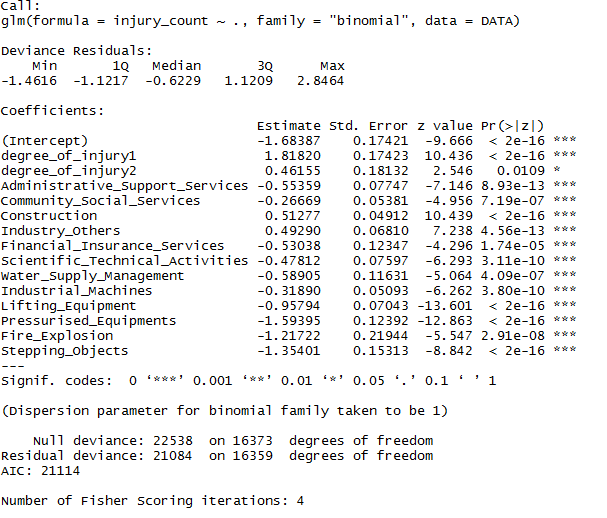


Figure 6: Third Iteration of our model

## 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:

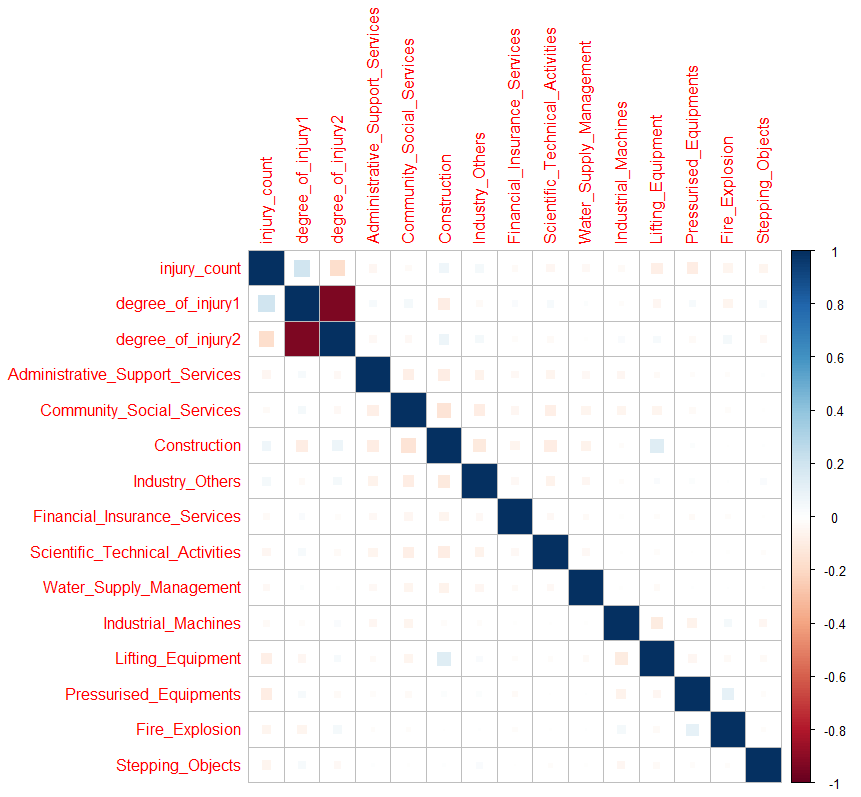


Figure (7): Correlation Matrix between predictor variable

1. We can ignore the injury\_count value as they are really not predictor value
2. Degree\_of\_injury1 and degree\_of\_injury2 shows negative relationship but from the data we know they are mutually exclusive

Hence we didn’t drop any predictor variable and ran the model.

## Observations

Below is the Confusion Matrix for this model:

predict

0 1

0 5474 3525 (approx. 60%)

1 3005 4370 (approx. 40%)

# Conclusion

The third iteration of our model showed the better results of all other iterations. To further improve the model, we would require more sample data and fine tune accordingly.