

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)

ContenTs

[Objective 3](#_Toc474080297)

[Problem Dataset 3](#_Toc474080298)

[Exploratory Analysis 4](#_Toc474080299)

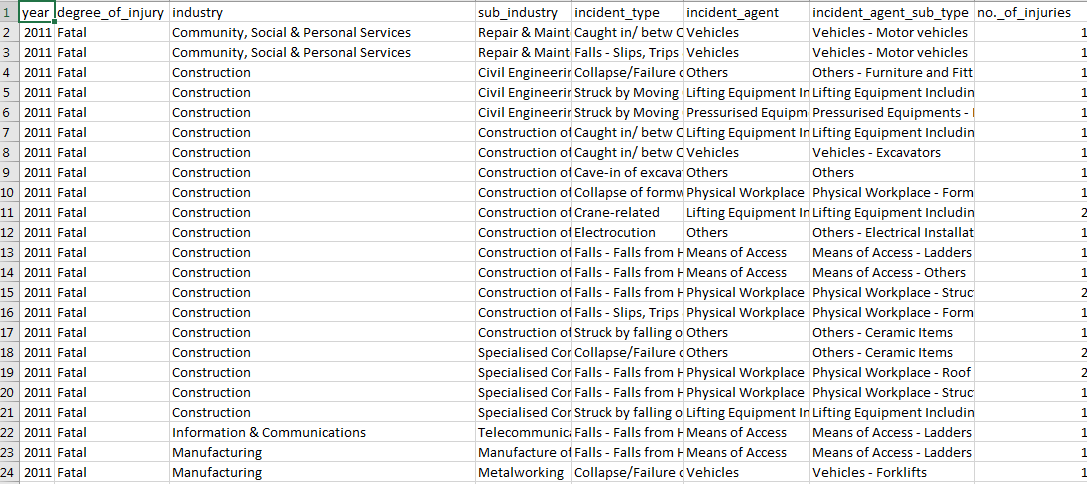
[Determination of Key Factors 5](#_Toc474080300)

[Observations 5](#_Toc474080301)

# 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 Dataset

By simply loading the dataset, we get the 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

Looking at this data we were inquisitive to know:

1. Is there any relation between single injury or group injury with other factors?
2. Can we predict based on DEGREE\_OF\_INJURY and other factors if two or more people were involved in the accident?

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. Scatter plots were used to detect unusual patterns. 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 the variable * There are too many unique attributes in the sub types which can lead to inaccuracies of the model. Those were dropped * Degree\_of\_Injury(DI), Industry(IND), and Incident\_agent(IA) were found suitable   As you can see in figure (1) we plotted the different accepted variables against the injury\_count. It doesn’t provide a very clear picture but definitely indicates that for some factors, injury\_count was quite low whereas of others had a huge count. |  |

Figure 1: Scatterplot of different variables with injury\_count

## Determination of Key Factors

We used random sampling to select 70% of the data for training and 30% for prediction. To determine the factors influencing the injury count we used logistic regression in language ‘R’. The first model summary of our data was:

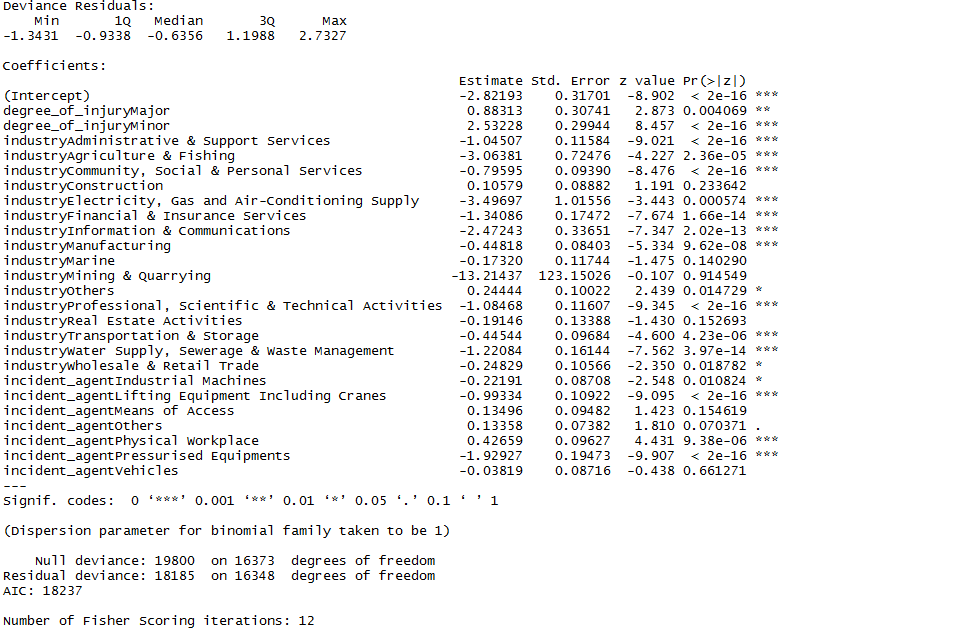


Figure 2: First Iteration of our model

## Observations

* Fatal injury type has been filtered in ‘R’ output as it has low significance in predicting the group injury (greater than 2)
* We identified few more factors like industry\_construction, industry\_marine etc. which have high P value, hence can be dropped from the model.

After filtering out the unimportant factors we ran the iteration 2 and following are the results: