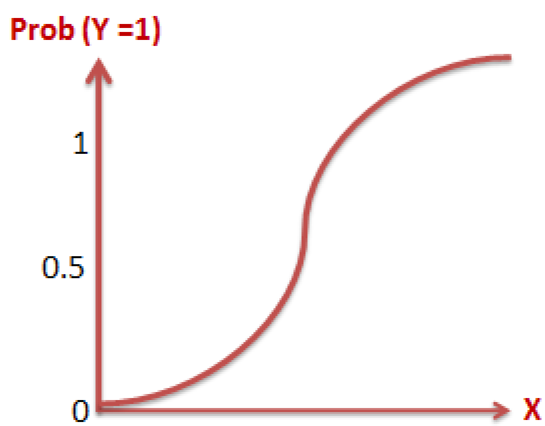

Apply Logistic Regression to Analyze Singapore Workplace Injury Data

EBS5101 Foundation of Business Analytics – Assignment 1

Submitted by:

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

Pradeep Kumar (A0163453H), Zheng Weiyu (A0163412R)

Objective

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:

1	year	degree_of_injury	industry	sub_industry	incident_type	incident_agent	incident_agent_sub_type	no._of_injuries
2	2011	Fatal	Community, Social & Personal Services	Repair & Maint	Caught in/ betw C	Vehicles	Vehicles - Motor vehicles	1
3	2011	Fatal	Community, Social & Personal Services	Repair & Maint	Falls - Slips, Trips	Vehicles	Vehicles - Motor vehicles	1
4	2011	Fatal	Construction	Civil Engineerir	Collapse/Failure c	Others	Others - Furniture and Fitt	1
5	2011	Fatal	Construction	Civil Engineerir	Struck by Moving	Lifting Equipment Ir	Lifting Equipment Includin	1
6	2011	Fatal	Construction	Civil Engineerir	Struck by Moving	Pressurised Equipm	Pressurised Equipments -	1
7	2011	Fatal	Construction	Construction of	Caught in/ betw C	Lifting Equipment Ir	Lifting Equipment Includin	1
8	2011	Fatal	Construction	Construction of	Caught in/ betw C	Vehicles	Vehicles - Excavators	1
9	2011	Fatal	Construction	Construction of	Cave-in of excava	Others	Others	1
10	2011	Fatal	Construction	Construction of	Collapse of formw	Physical Workplace	Physical Workplace - Form	1
11	2011	Fatal	Construction	Construction of	Crane-related	Lifting Equipment Ir	Lifting Equipment Includin	2
12	2011	Fatal	Construction	Construction of	Electrocution	Others	Others - Electrical Installat	1
13	2011	Fatal	Construction	Construction of	Falls - Falls from H	Means of Access	Means of Access - Ladders	1
14	2011	Fatal	Construction	Construction of	Falls - Falls from H	Means of Access	Means of Access - Others	1
15	2011	Fatal	Construction	Construction of	Falls - Falls from H	Physical Workplace	Physical Workplace - Struc	2
16	2011	Fatal	Construction	Construction of	Falls - Slips, Trips	Physical Workplace	Physical Workplace - Form	1
17	2011	Fatal	Construction	Construction of	Struck by falling o	Others	Others - Ceramic Items	1
18	2011	Fatal	Construction	Specialised Cor	Collapse/Failure c	Others	Others - Ceramic Items	2
19	2011	Fatal	Construction	Specialised Cor	Falls - Falls from H	Physical Workplace	Physical Workplace - Roof	2
20	2011	Fatal	Construction	Specialised Cor	Falls - Falls from H	Physical Workplace	Physical Workplace - Struc	1
21	2011	Fatal	Construction	Specialised Cor	Struck by falling o	Lifting Equipment Ir	Lifting Equipment Includin	1
22	2011	Fatal	Information & Communications	Telecommunic	Falls - Falls from H	Means of Access	Means of Access - Ladders	1
23	2011	Fatal	Manufacturing	Manufacture of	Falls - Falls from H	Means of Access	Means of Access - Ladders	1
24	2011	Fatal	Manufacturing	Metalworking	Collapse/Failure c	Vehicles	Vehicles - Forklifts	1

Source: data.gov.sg

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.

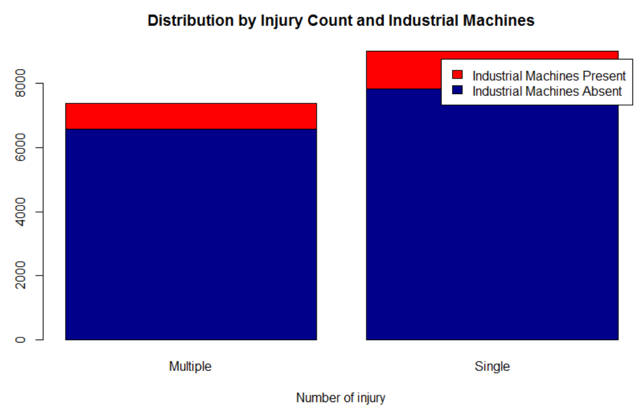

Figure 1: Number of Injuries vs Industrial Machine

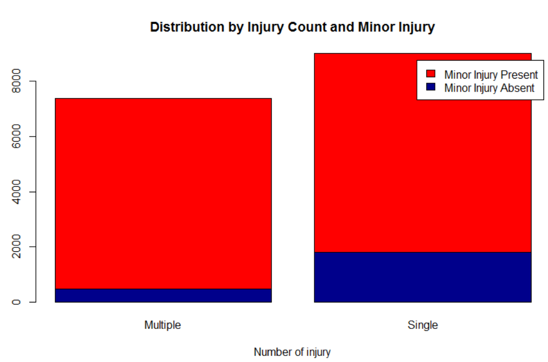

Figure 2: Number of Injuries vs Minor Injury

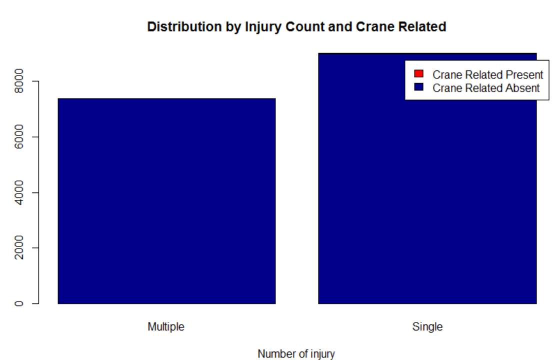

Figure 3: Number of Injuries vs Crane Related injuries

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:

```
DATA <- DATA[,list(injury_count=injury_count,
                   degree_of_injury=degree_of_injury,
                   industry=industry,
                   incident_type=incident_type,
                   incident_agent=incident_agent
                   )]

mylogit <- glm(injury_count ~ ., data = DATA, family = "binomial")
```

please note ***injury_count ~ .* above :- "." Represents all other variables in datatable

Below is the output of our model run:

```
glm(formula = injury_count ~ ., family = "binomial", data = DATA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8857  -1.0837  -0.5632   1.1011   2.9153

Coefficients:
                                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                                         0.24812    0.11284   2.199 0.027883 *
degree_of_injury2                                  -1.44150    0.06087 -23.682  < 2e-16 ***
degree_of_injury3                                  -1.88604    0.20530  -9.187  < 2e-16 ***
industryAdministrative & Support Services          -0.91083    0.10706  -8.508  < 2e-16 ***
industryAgriculture & Fishing                      -2.61616    0.44066  -5.937 2.90e-09 ***
industryCommunity, Social & Personal Services      -0.60031    0.09091  -6.604 4.01e-11 ***
industryConstruction                                0.26251    0.08902   2.949 0.003189 **
industryElectricity, Gas and Air-Conditioning Supply -2.46005  0.48543  -5.068 4.02e-07 ***
industryFinancial & Insurance Services             -0.89575    0.14449  -6.199 5.67e-10 ***
industryInformation & Communications               -1.69588    0.20783  -8.160 3.35e-16 ***
industryManufacturing                              -0.31225    0.08343  -3.743 0.000182 ***
industryMarine                                      0.02296    0.11381   0.202 0.840114
industryMining & Quarrying                        -15.00217  199.48746  -0.075 0.940053
industryOthers                                      0.25037    0.10050   2.491 0.012734 *
industryProfessional, Scientific & Technical Activities -0.80917 0.10589 -7.642 2.14e-14 ***
industryReal Estate Activities                     -0.21919    0.13218  -1.658 0.097262 .
industryTransportation & Storage                   -0.34251    0.09478  -3.614 0.000302 ***
industryWater Supply, Sewerage & Waste Management  -0.91662    0.13860  -6.613 3.76e-11 ***
industryWholesale & Retail Trade                   -0.16284    0.10393  -1.567 0.117172
incident_typeCave-in of excavation, tunnel, etc     0.40733    1.24318   0.328 0.743172
incident_typeCollapse of formwork/Failure of its supports 1.12340 1.02409 1.097 0.272653
incident_typeCollapse/Failure of Structure & Equipment -0.64292 0.30565 -2.103 0.035428 *
incident_typeCrane-related                          1.66681    0.57947   2.876 0.004022 **
incident_typeCut/Stabbed by Objects                 0.24797    0.07465   3.322 0.000895 ***
incident_typeDrowning                               0.21249    1.14477   0.186 0.852743
incident_typeElectrocution                        -12.93840  264.22137  -0.049 0.960945
incident_typeExposure to Biological Materials       0.09334    0.15668   0.596 0.551356
incident_typeExposure to Electric current          -0.55660    0.30268  -1.839 0.065929 .
incident_typeExposure to Extreme Temperatures      -0.30176    0.10445  -2.889 0.003863 **
incident_typeExposure to Hazardous Substances      -0.33316    0.12753  -2.612 0.008992 **
incident_typeFalls - Falls from Height              0.20119    0.10263   1.960 0.049958 *
incident_typeFalls - Slips, Trips & Falls           0.30319    0.06827   4.441 8.96e-06 ***
incident_typeFires & Explosion                     -1.29711    0.22617  -5.735 9.75e-09 ***
incident_typeOthers                                -0.87416    0.16738  -5.223 1.76e-07 ***
incident_typeOver-exertion/Strenuous Movements     -0.32026    0.08005  -4.001 6.31e-05 ***
incident_typeOxygen Defiiciency in Confined Space -12.86791  882.74340  -0.015 0.988370
incident_typePhysical Assault                       0.01028    0.18560   0.055 0.955844
incident_typeStepping on Objects                   -1.45448    0.16204  -8.976  < 2e-16 ***
incident_typeStriking against Objects              -0.38126    0.07931  -4.807 1.53e-06 ***
incident_typeStruck by Falling Objects             -0.01294    0.07012  -0.184 0.853627
incident_typeStruck by falling objects - From Heights -12.92271 490.78575 -0.026 0.978994
incident_typeStruck by Falling Objects from Heights -12.90961 230.24628 -0.056 0.955287
incident_typeStruck by Moving Objects               0.15441    0.06437   2.399 0.016452 *
incident_typeSuffocation                          -13.12866  507.35820  -0.026 0.979356
incident_typeSuffocation/Drowning                   0.32411    1.19001   0.272 0.785347
incident_typeWork-related Traffic                   1.22125    0.15941   7.661 1.85e-14 ***
incident_agentIndustrial Machines                  -0.17167    0.08141  -2.109 0.034966 *
incident_agentLifting Equipment Including Cranes   -0.87696    0.09625  -9.112  < 2e-16 ***
incident_agentMeans of Access                      -0.07946    0.09902  -0.803 0.422259
incident_agentOthers                                0.28632    0.07123   4.019 5.83e-05 ***
incident_agentPhysical Workplace                    0.25201    0.09883   2.550 0.010776 *
incident_agentPressurised Equipments               -1.43235    0.14044 -10.199  < 2e-16 ***
```

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:

```
DATA <- DATA[,list(injury_count=injury_count,
                   degree_of_injury1=degree_of_injury1,
                   degree_of_injury2=degree_of_injury2,
                   Marine=Marine,
                   Mining_Quarrying=Mining_Quarrying,
                   Administrative_Support_Services=Administrative_Support_Services,
                   Community_Social_Services=Community_Social_Services,
                   Construction=Construction,
                   Industry_Others=Industry_Others,
                   Financial_Insurance_Services=Financial_Insurance_Services,
                   Scientific_Technical_Activities=Scientific_Technical_Activities,
                   Water_Supply_Management=Water_Supply_Management,
                   Industrial_Machines=Industrial_Machines,
                   Lifting_Equipment=Lifting_Equipment,
                   Pressurised_Equipments=Pressurised_Equipments,
                   Crane_Related=Crane_Related,
                   Stabbed_Objects=Stabbed_Objects,
                   Extreme_Temp=Extreme_Temp,
                   Hazardous_Substance=Hazardous_Substance,
                   Falls_Trips=Falls_Trips,
                   Fire_Explosion=Fire_Explosion,
                   Incident_Type_Others=Incident_Type_Others,
                   Strenuous_Movements=Strenuous_Movements,
                   Stepping_Objects=Stepping_Objects,
                   Striking_Against=Striking_Against,
                   Work_Traffic=Work_Traffic)]

mylogit <- glm(injury_count ~ ., data = DATA, family = "binomial")
```

Here is the output of our model run:

```
Call:
glm(formula = injury_count ~ ., family = "binomial", data = DATA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7542  -1.0967  -0.5968   1.1157   2.8779

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -1.94314    0.18675 -10.405  < 2e-16 ***
degree_of_injury1                 2.09010    0.18648  11.208  < 2e-16 ***
degree_of_injury2                 0.64750    0.19254   3.363 0.000771 ***
Marine                            0.31711    0.08653   3.665 0.000248 ***
Mining_Quarrying                -13.68649  120.43891  -0.114 0.909524
Administrative_Support_Services  -0.57134    0.07831  -7.296 2.97e-13 ***
Community_Social_Services        -0.25169    0.05460  -4.610 4.03e-06 ***
Construction                      0.55322    0.05010  11.043  < 2e-16 ***
Industry_Others                   0.55162    0.06894   8.002 1.23e-15 ***
Financial_Insurance_Services     -0.56133    0.12447  -4.510 6.49e-06 ***
Scientific_Technical_Activities  -0.47345    0.07673  -6.170 6.83e-10 ***
Water_Supply_Management          -0.59966    0.11733  -5.111 3.20e-07 ***
Industrial_Machines              -0.33967    0.05220  -6.508 7.64e-11 ***
Lifting_Equipment                -1.01659    0.07202 -14.114  < 2e-16 ***
Pressurised_Equipments           -1.58238    0.12522 -12.637  < 2e-16 ***
Crane_Related                     1.80526    0.57088   3.162 0.001566 **
Stabbed_Objects                   0.21112    0.05877   3.592 0.000328 ***
Extreme_Temp                     -0.26630    0.09363  -2.844 0.004455 **
Hazardous_Substance              -0.30069    0.11770  -2.555 0.010628 *
Falls_Trips                       0.18139    0.04727   3.838 0.000124 ***
Fire_Explosion                   -1.24714    0.22073  -5.650 1.60e-08 ***
Incident_Type_Others             -0.80666    0.15993  -5.044 4.56e-07 ***
Strenuous_Movements              -0.32579    0.06520  -4.997 5.82e-07 ***
Stepping_Objects                 -1.39234    0.15447  -9.014  < 2e-16 ***
Striking_Against                 -0.39378    0.06508  -6.050 1.44e-09 ***
Work_Traffic                      0.83291    0.14633   5.692 1.26e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22538  on 16373  degrees of freedom
Residual deviance: 20856  on 16348  degrees of freedom
AIC: 20908

Number of Fisher Scoring iterations: 12
```

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:

```
DATA <- DATA[,list(injury_count=injury_count,
                   degree_of_injury1=degree_of_injury1,
                   degree_of_injury2=degree_of_injury2,
                   Administrative_Support_Services=Administrative_Support_Services,
                   Community_Social_Services=Community_Social_Services,
                   Construction=Construction,
                   Industry_Others=Industry_Others,
                   Financial_Insurance_Services=Financial_Insurance_Services,
                   Scientific_Technical_Activities=Scientific_Technical_Activities,
                   Water_Supply_Management=Water_Supply_Management,
                   Industrial_Machines=Industrial_Machines,
                   Lifting_Equipment=Lifting_Equipment,
                   Pressurised_Equipments=Pressurised_Equipments,
                   Fire_Explosion=Fire_Explosion,
                   Stepping_Objects=Stepping_Objects)]

mylogit <- glm(injury_count ~ ., data = DATA, family = "binomial")
```

Below is the output of our model in third iteration:

```
Call:
glm(formula = injury_count ~ ., family = "binomial", data = DATA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4616  -1.1217  -0.6229   1.1209   2.8464

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.68387    0.17421  -9.666  < 2e-16 ***
degree_of_injury1                1.81820    0.17423  10.436  < 2e-16 ***
degree_of_injury2                0.46155    0.18132   2.546   0.0109 *
Administrative_Support_Services -0.55359    0.07747  -7.146 8.93e-13 ***
Community_Social_Services       -0.26669    0.05381  -4.956 7.19e-07 ***
Construction                     0.51277    0.04912  10.439  < 2e-16 ***
Industry_Others                  0.49290    0.06810   7.238 4.56e-13 ***
Financial_Insurance_Services    -0.53038    0.12347  -4.296 1.74e-05 ***
Scientific_Technical_Activities -0.47812    0.07597  -6.293 3.11e-10 ***
Water_Supply_Management         -0.58905    0.11631  -5.064 4.09e-07 ***
Industrial_Machines             -0.31890    0.05093  -6.262 3.80e-10 ***
Lifting_Equipment               -0.95794    0.07043 -13.601  < 2e-16 ***
Pressurised_Equipments          -1.59395    0.12392 -12.863  < 2e-16 ***
Fire_Explosion                  -1.21722    0.21944  -5.547 2.91e-08 ***
Stepping_Objects                -1.35401    0.15313  -8.842  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22538  on 16373  degrees of freedom
Residual deviance: 21084  on 16359  degrees of freedom
AIC: 21114

Number of Fisher Scoring iterations: 4
```

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:

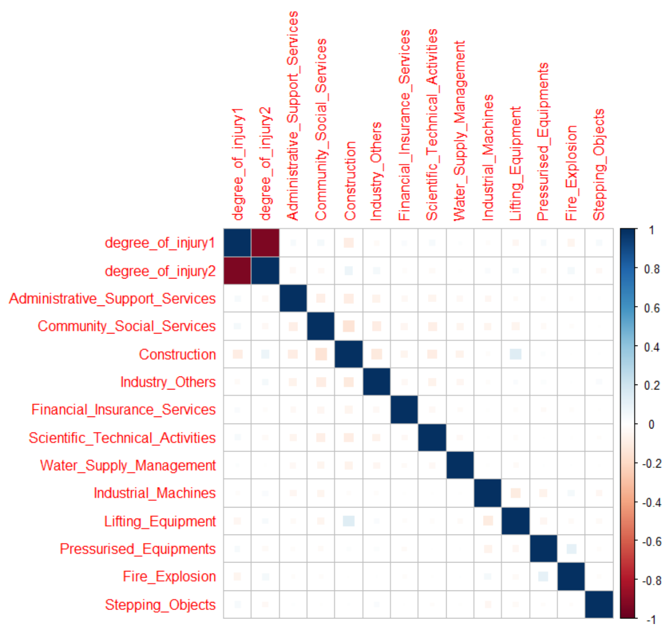

Figure (7): Correlation Matrix between predictor variable

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 and ran 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
2. Logistic Regression - https://www.youtube.com/watch?v=zAULhNrnuL4