Regression Analysis for Industrial Manufacturing Company

Jing Heng Lim

September 6, 2019

Background and Objective

A recent unfortunate accident in South America has put your company's safety record under the microscope. Journalists have been sending your CEO difficult questions about your company's workplace safety practices, which have been identified as highly variable across your network. Perform some analysis on workplace injury data to help inform your company's response to this growing crisis.

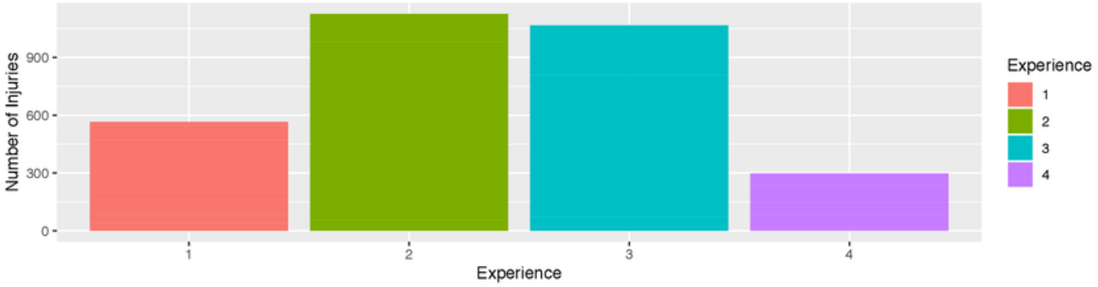
From the data the variables are:

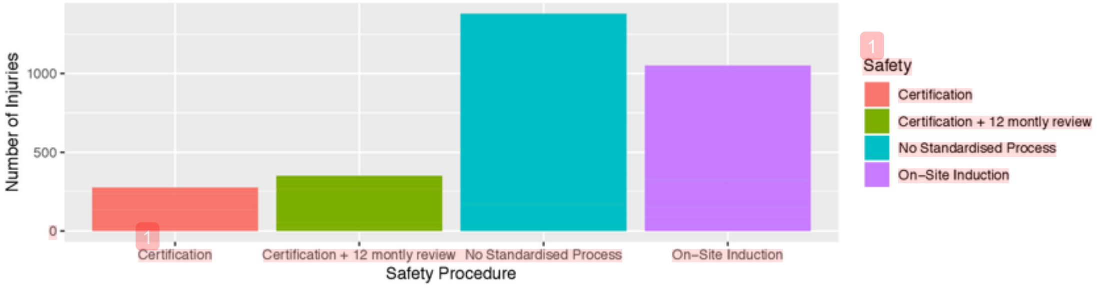
```
Injuries - number of injuries in the group

Safety - the safety regime in place for the group

Hours - total hours worked by this group

Experience - the experience level in years of the group
```

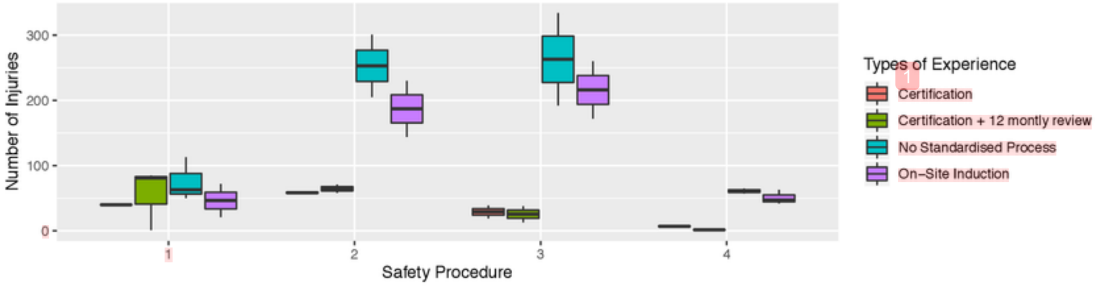
```
##  Experience                               Safety     Injuries
##  1:10       Certification                    :9   Min.   :  1.00
##  2: 8       Certification + 12 montly review:9   1st Qu.: 33.75
##  3: 8       No Standardised Process          :9   Median : 57.50
##  4:10       On-Site Induction                :9   Mean   : 84.92
##                                                   3rd Qu.: 92.00
##                                                   Max.   :334.00
##      Hours
##  Min.   :  34574
##  1st Qu.: 130272
##  Median : 302879
##  Mean   : 549996
##  3rd Qu.: 813381
##  Max.   :2135146
```

There were 37 data being recorded in this data set. There were 84 worker injured in average while working and the highest injuries were 334 worker injured. Experience and Safety variable (factor) is balanced into group.

Exploratary Data







This box plot shows that worker who had experience 2 and 3 have a higher chance of causing themselves injured. Whereas, worker who sit in Experience4 is less likely getting injured.

On safety procedure point of view in the second boxplot, no standardised process and on-site induction are not an ideal safety procedure to prevent injuries. From the 3rd boxplot,it indicates that even worker that has a significant amount of experience would most likely injured themselves if a 'formal' safety regime (certification and 12month review) is not implemented.

Poisson Model Choice

Poisson generalised linear model would be a perfect start to build a model for our data since our outcomes is a count data. A stepwise selection based on Akaike Information Criterion (AIC) will be implemented to determine the best model for us by including significant variables. Both backward and forward selection were used to find an optimal model.

Inspect AIC results for both forward and backward model

Posisson backward model:

```
formula(backward_model)
```

```
## Injuries ~ Experience * Safety + offset(log(Hours))
```

```
AIC(backward_model)
```

```
## [1] 682.1231
```

Poisson forward model:

```
formula(forward_model)
```

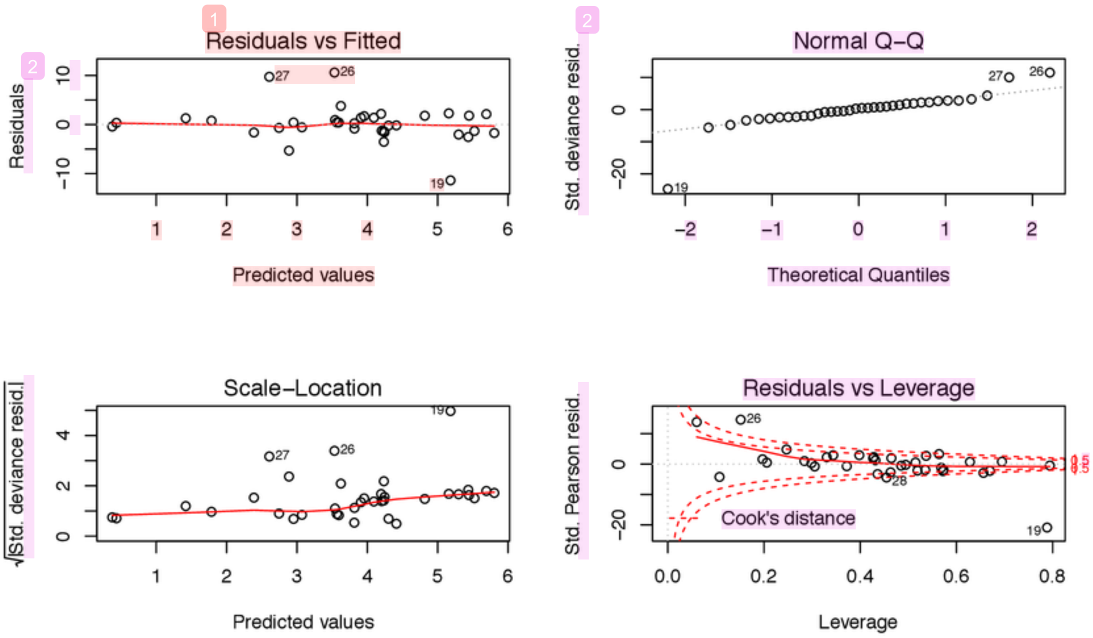
```
## Injuries ~ Experience + Safety + Experience:Safety + offset(log(Hours))
```

```
AIC(forward_model)
```

```
## [1] 682.1231
```

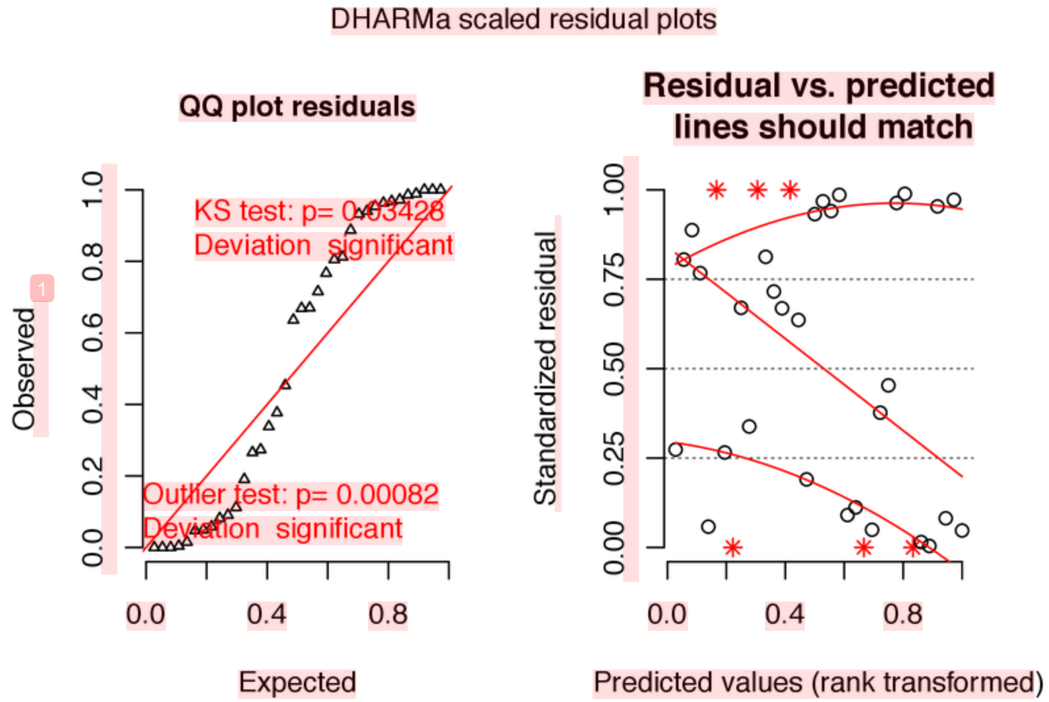
Both backward and forward selection model appears to be the exact same model. We can now inspect the residual plots using the simulation by DHARMa package.

Checking for any outlier for the Poisson GLM



Based on Cook's distance, it appears that row 19 is an outlier for the model. Hence, it will be removed.

Simulate Residuals from the Poisson GLM



It is clear that overdispersion occer based on QQplot as the distribution between redisuals and expected does not match.

Overdispersion Test for Poisson GLM

```
##
##  Overdispersion test
##
## data:  backward_model
## z = 1.9973, p-value = 0.02289
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   15.26074
```

The p-value for the test of dispersion is highly significant (p-value= 0.02289), this indicates that the data is more variable than expected under Poisson GLM model. Hence, it is overdisperse.

Therefore, a Quasi-poisson model will be considered as quasi-likelihood estimation is one way of allowing for overdispersion.

Quasi-poisson model

```
## 
## Call:
## glm(formula = Injuries ~ . - Hours + offset(log(Hours)), family = quasipoisson, 
##     data = injuries)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -5.2906  -1.2597  -0.2483   1.5257   2.9520  
## 
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                               -7.98819    0.14260 -56.019
## Experience2                               -0.52335    0.11068  -4.729
## Experience3                               -1.08325    0.11439  -9.470
## Experience4                               -2.02253    0.15367 -13.161
## SafetyCertification + 12 montly review     0.02917    0.16387   0.178
## SafetyNo Standardised Process              0.20427    0.13449   1.519
## SafetyOn-Site Induction                    0.36800    0.13842   2.659
##                                           Pr(>|t|)    
## (Intercept)                                < 2e-16 ***
## Experience2                               5.83e-05 ***
## Experience3                               3.16e-10 ***
## Experience4                               1.63e-13 ***
## SafetyCertification + 12 montly review      0.8600    
## SafetyNo Standardised Process               0.1400    
## SafetyOn-Site Induction                     0.0128 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for quasipoisson family taken to be 4.071237)
## 
##     Null deviance: 1098.20  on 34  degrees of freedom
## Residual deviance:  125.79  on 28  degrees of freedom
## AIC: NA
## 
## Number of Fisher Scoring iterations: 4
```

The dispersion parameter is approximately 4.07 in this case which is more than 1, This indicates that overdispersion still occur in the data.

Chi-square Test for Quasi-poisson model

Chi-sqr test:

```
qchisq(0.95, df=quasi_model$df.residual)
```

```
## [1] 41.33714
```

Deviance of Quasi poisson model:

```
deviance(quasi_model)
```

```
## [1] 125.7905
```

It shows that the deviance of the model is larger than the chi-squared test. Hence, it indicates that the model doesn't fit well form the data. Since, Quasi-poisson does not fit well here. We shall therefore consider implementing a Negative-Binomial model instead, which is more flexible.

Negative Binomial Model

Negative binomial model is another modelling count variables and is widely used for over-dispersed count outcome variables. It can be used when the conditional variance exceed the conditional mean [Var(x)>E(x)]. Similar approach from Poisson GLM is used for AIC forward and backward selection.

Mean-Variance relationship

E(x):

```
mean_x<-mean(Injuries)
variance_x<-var(Injuries)
mean_x
```

```
## [1] 84.91667
```

Var(x):

```
variance_x
```

```
## [1] 7669.85
```

Therefore, negative binomial regression can be used for our data since Var(x)>E(x).

Model Selection

Negative-binomial forward model:

```
formula(NB_forward_sel)
```

```
## Injuries ~ (Experience + Safety + Hours) - Hours + offset(log(Hours))
```

Negative-binomial backward model:

```
formula(NB_backward_sel)
```

```
## Injuries ~ Experience + Safety + offset(log(Hours))
```

AIC results for both model

```
AIC(NB_forward_sel)
```

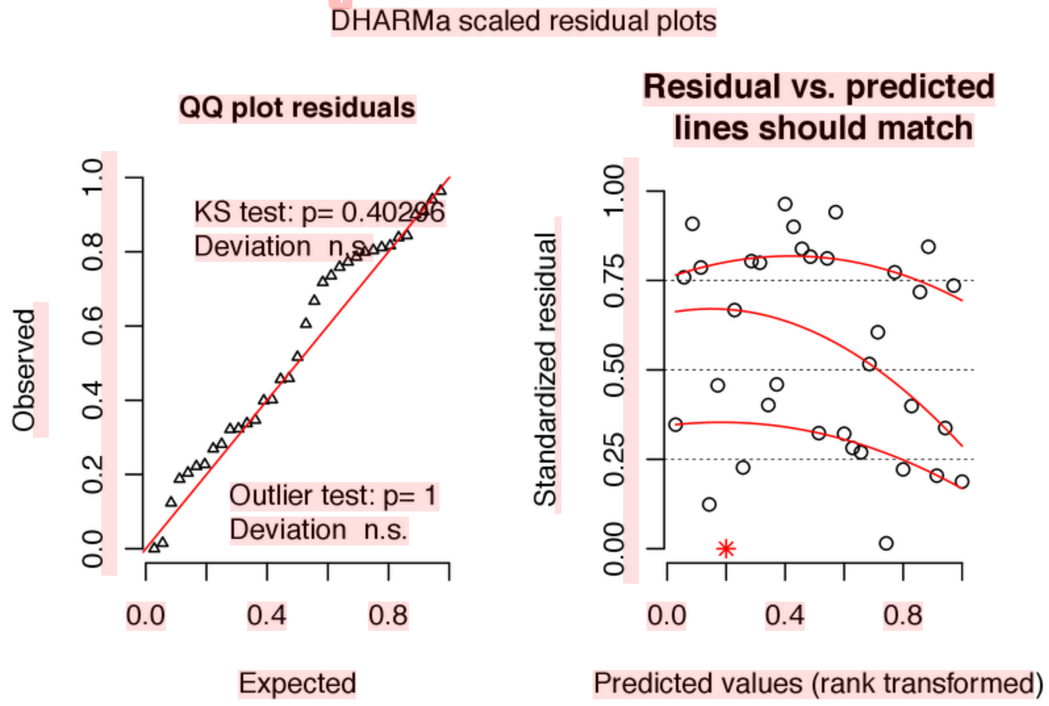
```
## [1] 300.1648
```

```
AIC(NB_backward_sel)
```

```
## [1] 300.1648
```

As we obeserved the AIC results by using Negative-Binomial model (NB) had relatively decreased to AIC = [1] 300.1648 compared to the Poisson and Quassi-Poisson model.

```
nb_residuals<-simulateResiduals(NB_backward_sel)
plot(nb_residuals)
```



From the QQplot it seems slightly better here, even the points are not well distributed at the tail. However this could probably due to small amount of data that we obtained but theres is definately a better fit compare to previous model. Since the AIC results for forward and backward selection are the same we can pick either as our final model. In this case would be, Injuries ~ Experience + Safety + offset(log(Hours)).

Scaled deviance test for negative binomial backward selection

Chi-sqr test:

```
## [1] 41.33714
```

Deviance of Negative-Binomial model

```
deviance(NB_backward_sel)
```

```
## [1] 45.69852
```

Besides, through scaled deviance test the deviance value (45.69852) is slightly higher than the chi-square dist value (45.69852) this indicates that the data is still slightly under-fit for the model. However, NB definately has a better fit compare to Poisson and Quassi-Poisson.

```
## 
## Call:
## glm.nb(formula = Injuries ~ Experience + Safety + offset(log(Hours)),
##     data = injuries, init.theta = 25.59005637, link = "log")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3926  -0.5611  -0.1710   0.7442   1.5650
##
## Coefficients:
##                                        Estimate Std. Error z value
## (Intercept)                            -7.98975    0.11889 -67.205
## Experience2                            -0.48480    0.11439  -4.238
## Experience3                            -1.11907    0.11885  -9.416
## Experience4                            -1.97729    0.13273 -14.897
## SafetyCertification + 12 montly review -0.02443    0.13587  -0.180
## SafetyNo Standardised Process           0.24421    0.12308   1.984
## SafetyOn-Site Induction                 0.36235    0.12342   2.936
##                                        Pr(>|z|)
## (Intercept)                             < 2e-16 ***
## Experience2                            2.25e-05 ***
## Experience3                             < 2e-16 ***
## Experience4                             < 2e-16 ***
## SafetyCertification + 12 montly review  0.85729
## SafetyNo Standardised Process           0.04724 *
## SafetyOn-Site Induction                 0.00333 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(25.5901) family taken to be 1)
##
##     Null deviance: 300.730  on 34  degrees of freedom
## Residual deviance:  45.699  on 28  degrees of freedom
## AIC: 300.16
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  25.6
##           Std. Err.:  10.8
##
##  2 x log-likelihood:  -284.165
```

The increase of 1 worker in experience 2 will reduce approx 0.48 times of injuries, experience 3 will reduce aprrox 1.12 and experience 4 will reduce approx 1.977 respectively. For safety procedure, a increase of 1 worker that had safe certification with 12 month review would reduce 0.02 times of injuries. However, no standardised process and on-site induction will increase the number of injuries by 0.24 and 0.36 respectively. The standard error tell us that we have approximately 0.11-0.13% of variation in our model based on all the variables used.

Confidence Interval for covariate coefficients

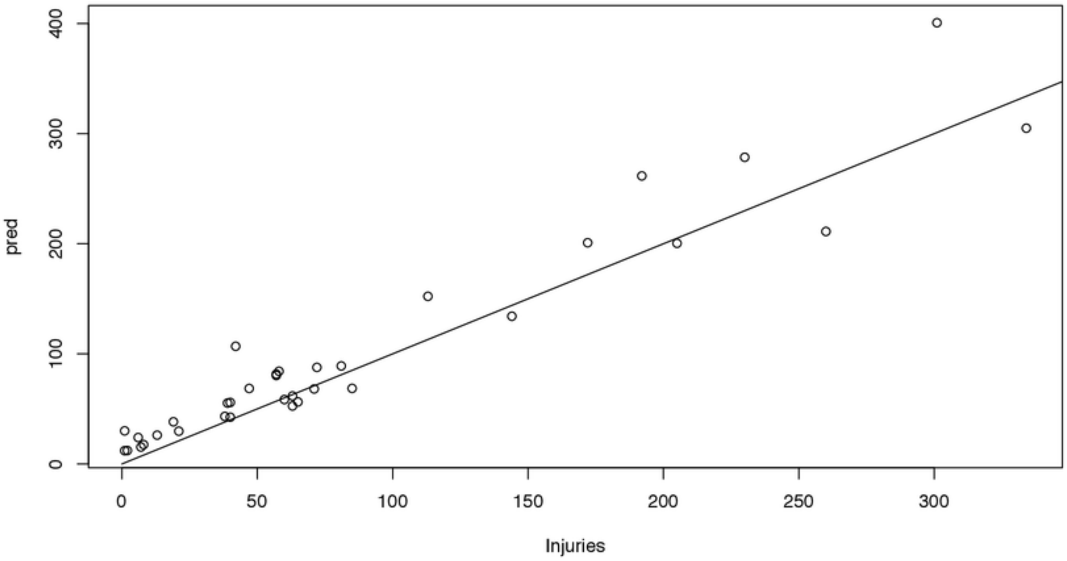
```
##                                           2.5 %     97.5 %
## (Intercept)                          -8.220809299 -7.7565725
## Experience2                          -0.708429159 -0.2605011
## Experience3                          -1.351060415 -0.8867063
## Experience4                          -2.238358422 -1.7171381
## SafetyCertification + 12 montly review -0.289142621  0.2405599
## SafetyNo Standardised Process         0.002152297  0.4877352
## SafetyOn-Site Induction               0.120522523  0.6054524
```

We are 95% confident that the covariates value lies between the 2 values above respectively.

Assess the performance of the model and validating the model

Cross validation of the model by leave one out

```
## The following objects are masked from injuries (pos = 11):
##
##     Experience, Hours, Injuries, Safety
```



From the plot above, it shows that the predicted injuries and the actual injuries are quite linear. It indicastes that the model holds a high perfomance predictabilty for future unforseen number of injuries that would occur.

In general, the Possion GLM and Quasi-Poisson were a poor fit to the data, as the data was overdispersed with respect to it. Additionally, the AIC for the best Poisson GLM was 682.1231 and , and the AIC for the best negative binomial GLM was 300.1648, which is substantially lower and indicates the negative binomial is a much better fit to the data.

In conclusion to justify CEO's concerns, a certification or a certification with 12 monthly review is recommended as a international standard safety regime for the company based on the exploratory analysis by plotting number of injuries aginst different safety regime. Futhermore the boxplot shows that safety regime is more important than experiences when it comes to preventing injuries as it tends to has a lower count of injuries throught out 4 level of experience of workers. Furthermore, workers tends to injured themselves when they had no standardised process and on-site induction safety regime.