Assignment

Kevin Nguyen

2023-04-17

Introduction

We have been tasked with performing an analysis on workplace injury data to help inform the company's response to a growing crisis.

Research Question

1. Recommend an existing safety regime, based on injury rate, to implement as the international standard company-wide.
2. Find supporting evidence to suggest whether experience is more important than safety regime in reducing injury rate.

Including Plots

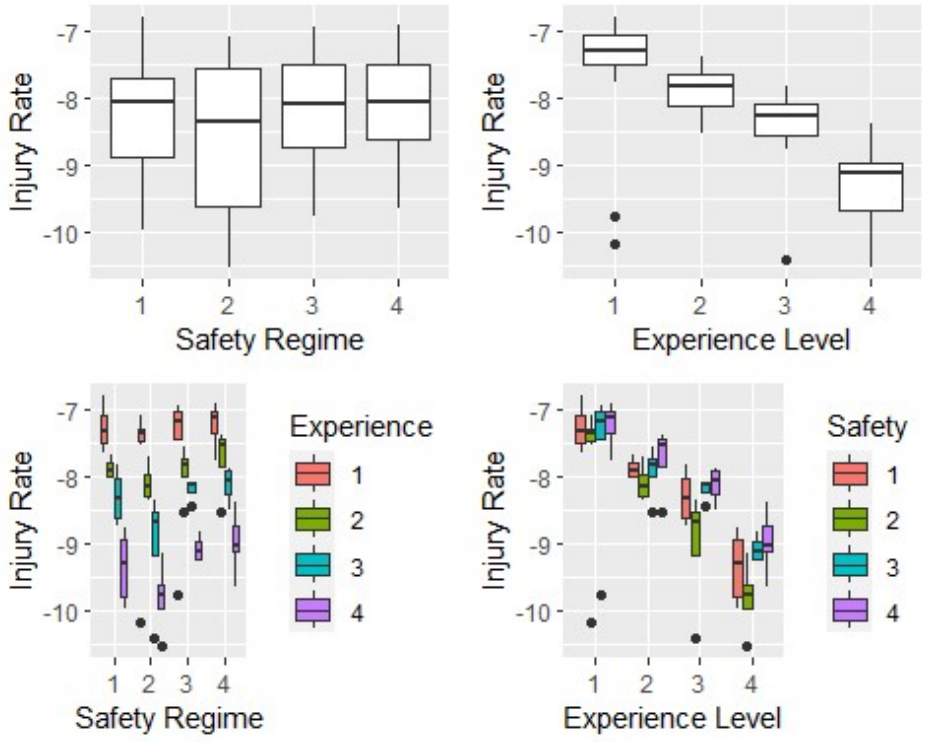
Analysis of Data

```
##     Injuries      Safety Experience     Hours                ID
##  Min.   :  0.0   1:18   1:20       Min.   :  34574   Min.   : 1.00
##  1st Qu.: 46.0   2:18   2:16       1st Qu.: 130272   1st Qu.:18.75
##  Median :106.5   3:18   3:16       Median : 302879   Median :36.50
##  Mean   :162.2   4:18   4:20       Mean   : 549996   Mean   :36.50
##  3rd Qu.:179.2                     3rd Qu.: 813381   3rd Qu.:54.25
##  Max.   :913.0                     Max.   :2135146   Max.   :72.00
```

Observations of the data show that the data set contains information on 72 groups, with each group encountering 162 injuries per year on average. The predictors Safety and Experienced are roughly balanced by group as Safety has 18 group in each Safety Regime, and Experience ranging from 16 and 20, which are very similar frequencies. Hours appears to have a large range from the Min to the Max.

The response will be transformed into $log\left(\frac{Injuries+1}{Hours}\right)$. The Hours is in the denominator because our research question is to find the injury rate. Injury rate is a better measure for our solution than injury as injury rate treats workers equally irrespective how many hours they worked in 12 months. The data of Injuries has a minimum of 0. This causes an undefined answer when solving for log(0). Hence the addition of 1 to the numerator is to ensure numerical stability, and data points with Injuries = 0 are included in the modelling. The log is to transform the output away from very small numbers, as this could lead to truncating errors, and closer to numbers that are easier for analysts to comprehend and compare.

Exploratory Analysis



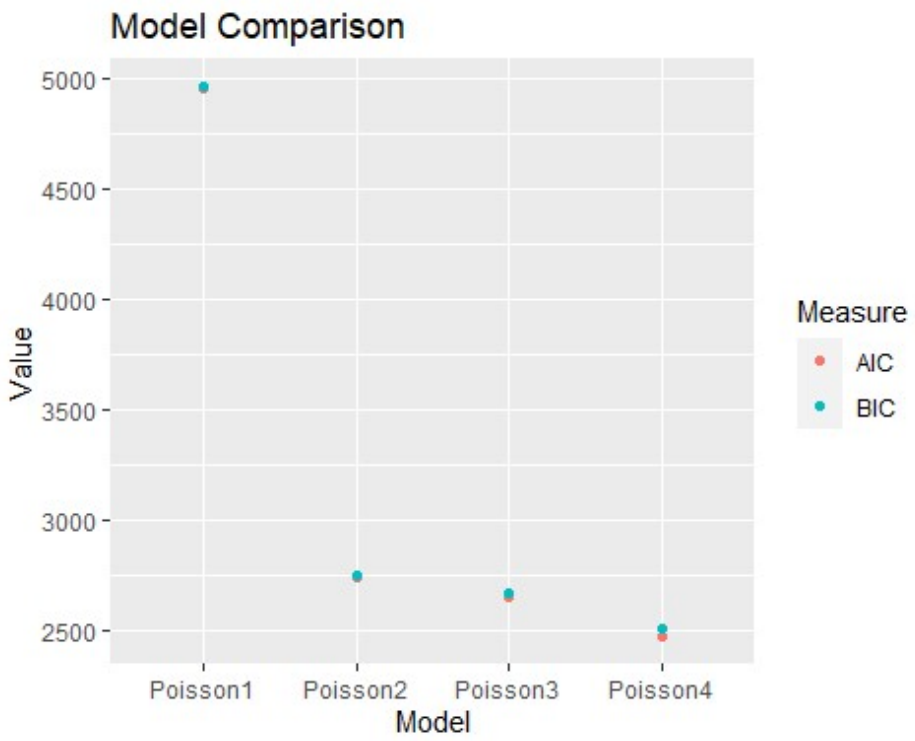
Observing the plots above, the injury rate of experience1 is larger than the injury rate of experience4. There are large differences in injury rate among different experience levels. This suggests that experience may have strong influence on injury rate. There is less difference on injury rate among Safety levels than experience. This may suggest that experience is more important than safety at reducing injury rate. The plots also suggest that safety regime 2 has the lowest injury rate outcome, possibly making it the most ideal safety regime. The plots did not include Hours effect on injury rate as analysing Hours is not relevant to the research questions.

Poisson

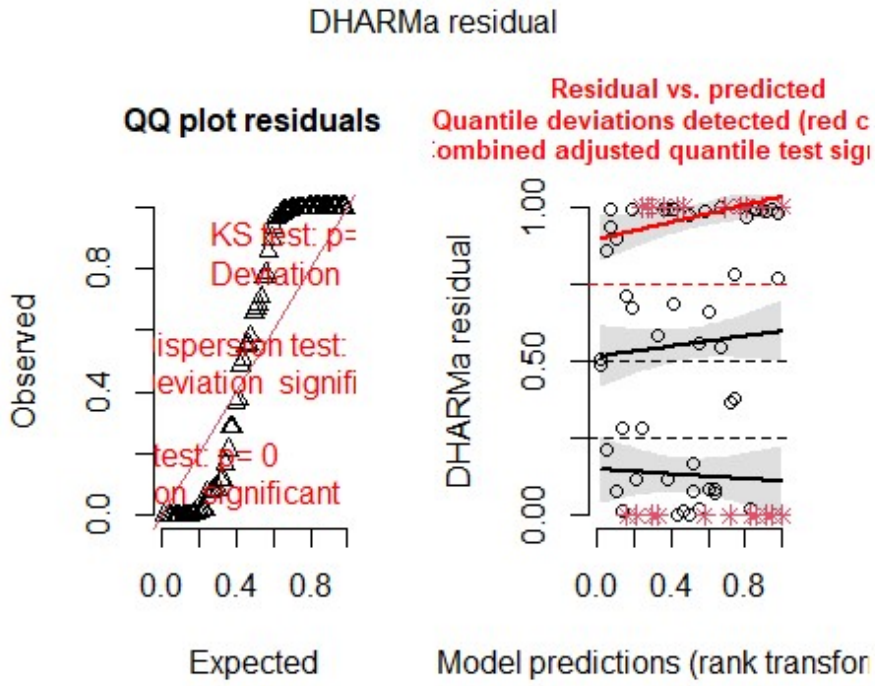
We will be fitting the data with the poisson regression model to model and predict the injury data. Poisson regression is ideal as our data uses a count dependent variable and categorical/continuous predictors. There are a total of 4 poisson models, each testing for an appropriate fit using different combinations of predictors. The response will be Injuries+1 due to the injury rate, and the predictors will be Safety, Experience and Safety Experience interaction. An additional predictor is the offset of log(Hours).

Model	AIC	BIC
Poisson1	4954.657	4963.764
Poisson2	2738.961	2748.068
Poisson3	2652.212	2668.149

Model	AIC	BIC
Poisson4	2472.477	2508.904



The best performing model is Poisson 4. This because it gave lowest AIC and BIC value compared to the other models. This means the model of choice is Poisson 4. Now we will check on the model fit.



The QQ plot shows that the poisson model has failed all 3 tests. The model is over dispersed as more residuals are in the tails of the distribution than in the center. The residuals are not normally distributed as they do not follow the line. Residuals vs predicted show that the 0.75 quantile residuals are not randomly distributed and has been flagged by the function. Overall, the DHARmMa residuals show that the Poisson is a weak fit for the data.

```
##
##  Overdispersion test
##
## data:  final_poisson_model
## z = 3.046, p-value = 0.001159
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   23.99997

## [1] TRUE
```

The above shows that the overdispersion test gave a value of dispersion=24, this is very far away from the appropriate value of 1, so we conclude by saying the Poisson model is over dispersed and not appropriate to fit the data. We will now find the Quasi-Poisson and Negative Binomial and see their suitability in fitting the data.

Quasi-Poisson

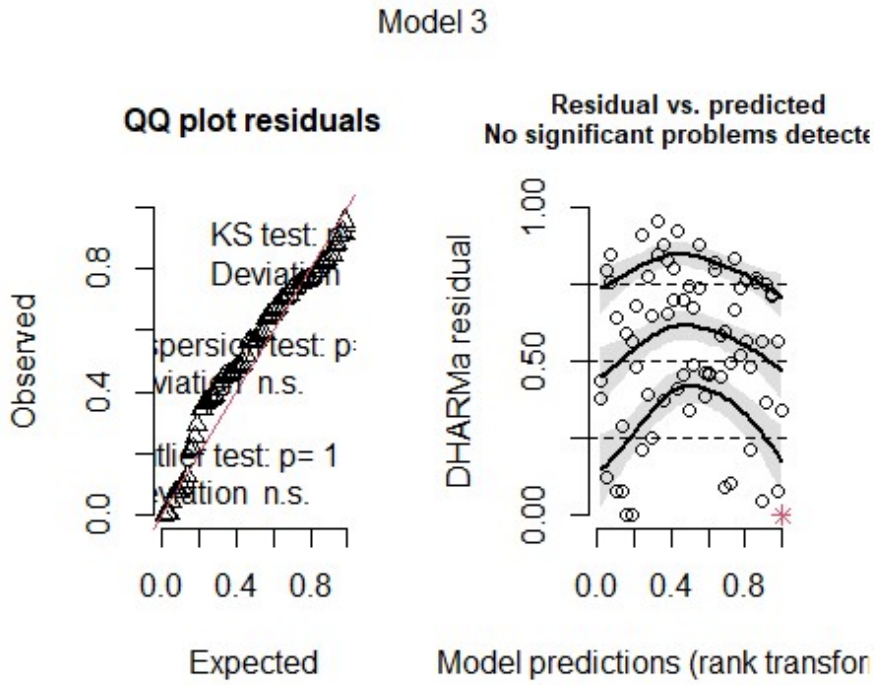
We create a Quasi-Poisson model and inspect later.

Negative Binomial

Build Negative Binomial models. We created 4 models using a combination of predictors.

Model	AIC	BIC
NB1	826.0854	837.4687
NB2	745.4422	756.8255
NB3	741.7381	759.9514
NB4	756.5899	795.2932

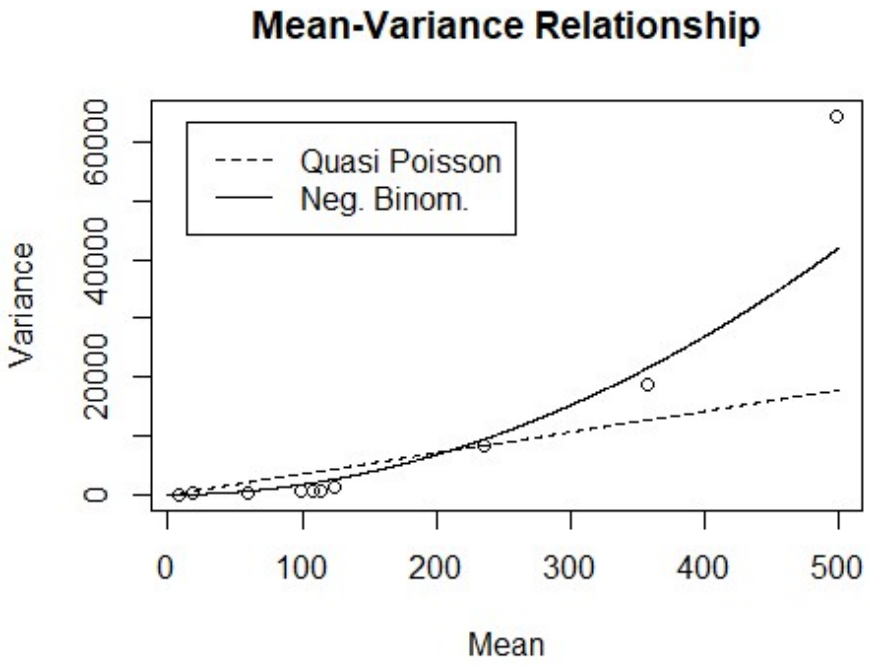
m3 performs the best in AIC, m2 has the lowest BIC. We choose to assess the models on AIC over the BIC as BIC penalises more complex models, however we have few predictors and want to keep as many as possible.



The Negative Binomial looks to be a good fit to the data. The residuals in the QQ plot are normally distributed, the residuals are roughly evenly distributed across tails and center, and all 3 tests have not been flagged. The Residuals vs predicted have not been flagged and look to be randomly distributed with no trends.

Model Comparison

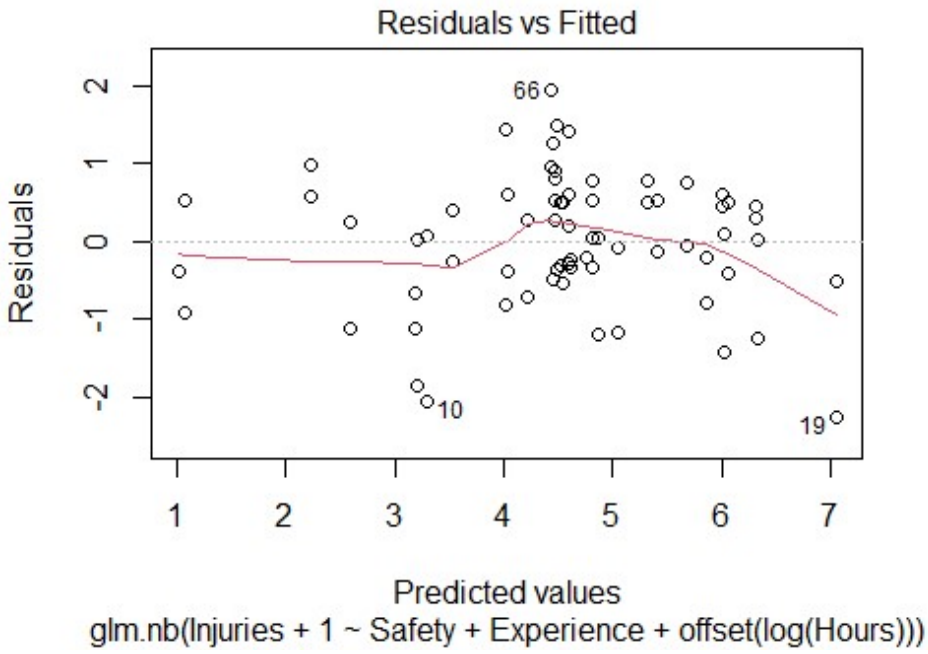
Having assessed that the Poisson model was a poor fit to the data, we now compare the quasi-poisson and negative binomial models find the best fit.

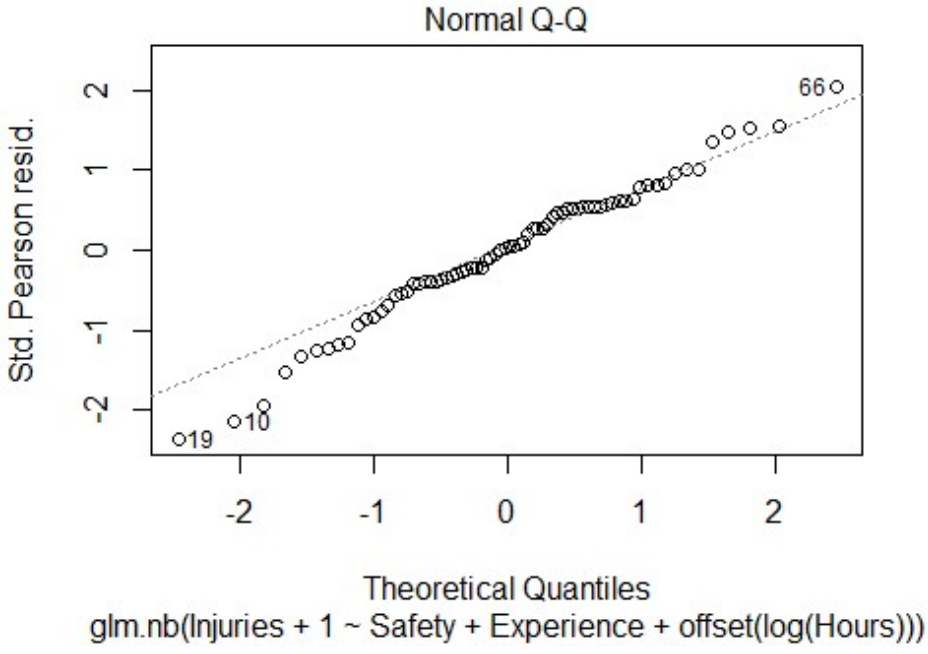


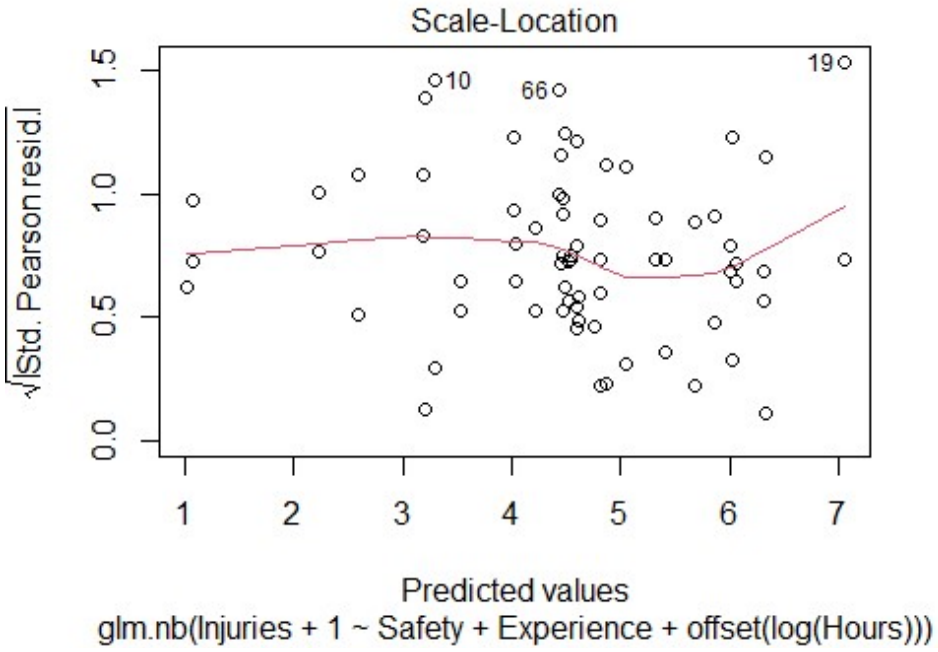
The Mean-Variance Relationship graph shows that the Quasi-Poisson does not fit the data very well. The Negative Binomial was a better fit. Therefore the Negative Binomial model was chosen as the best model.

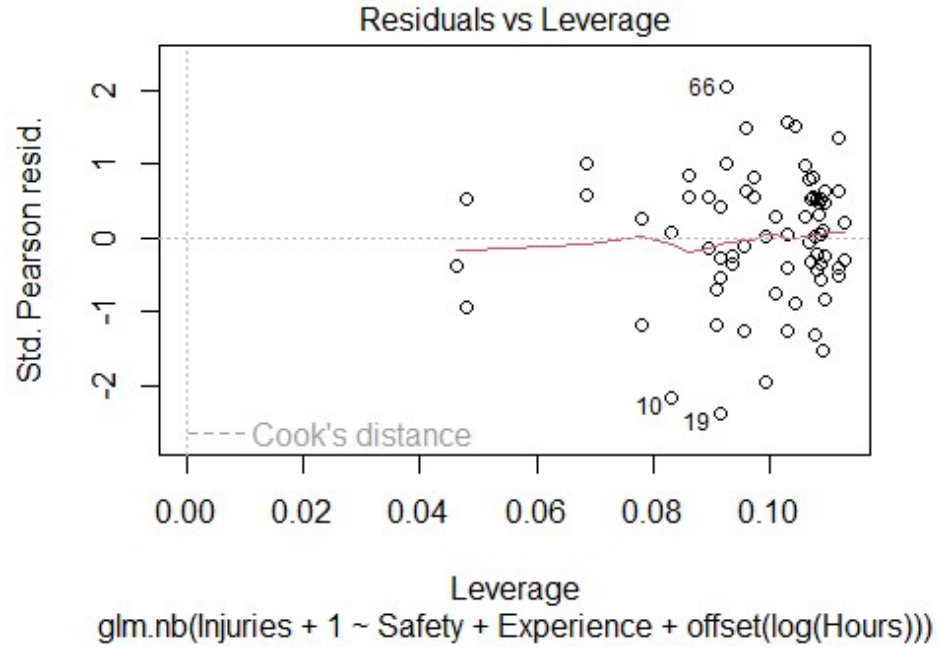
Checking Validity of final model

We do residual plots to test the validity of the model and whether any assumptions are violated.


Residuals vs Fitted
Residuals
Predicted values
glm.nb(Injuries + 1 ~ Safety + Experience + offset(log(Hours)))
66
10
19


Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
glm.nb(Injuries + 1 ~ Safety + Experience + offset(log(Hours)))
66
10
19





The Residuals vs Fitted plot show no clear linear trends and no constant variance fanning. The QQ plot is normally distributed, however there bottom left tail does not follow the normal distribution. This is only minor and so does not violate any glm assumptions. The square root standard deviation residuals vs predicted show no clear linear trends or

fanning, this suggests that there is no further significant variability that is not captured by the model. Therefore, the assumptions of the glm model have not been violated and the final Negative Binomial model is valid.

The summary of the final Negative Binomial model

```
##
## Call:
## glm.nb(formula = Injuries + 1 ~ Safety + Experience + offset(log(Hours)),
##     data = data, link = "log", init.theta = 6.033302632)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3446  -0.4154   0.0321   0.4853   1.5843
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.26831    0.13505 -53.821  < 2e-16 ***
## Safety2     -0.27817    0.15272  -1.821   0.0685 .
## Safety3      0.03641    0.14281   0.255   0.7987
## Safety4      0.20197    0.14321   1.410   0.1584
## Experience2 -0.57236    0.14079  -4.065 4.79e-05 ***
## Experience3 -1.02683    0.14191  -7.236 4.63e-13 ***
## Experience4 -1.89059    0.14504 -13.035  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(6.0333) family taken to be 1)
##
##     Null deviance: 248.493  on 71  degrees of freedom
## Residual deviance:  75.873  on 65  degrees of freedom
## AIC: 741.74
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  6.03
##           Std. Err.:  1.13
##
##  2 x log-likelihood:  -725.738
```

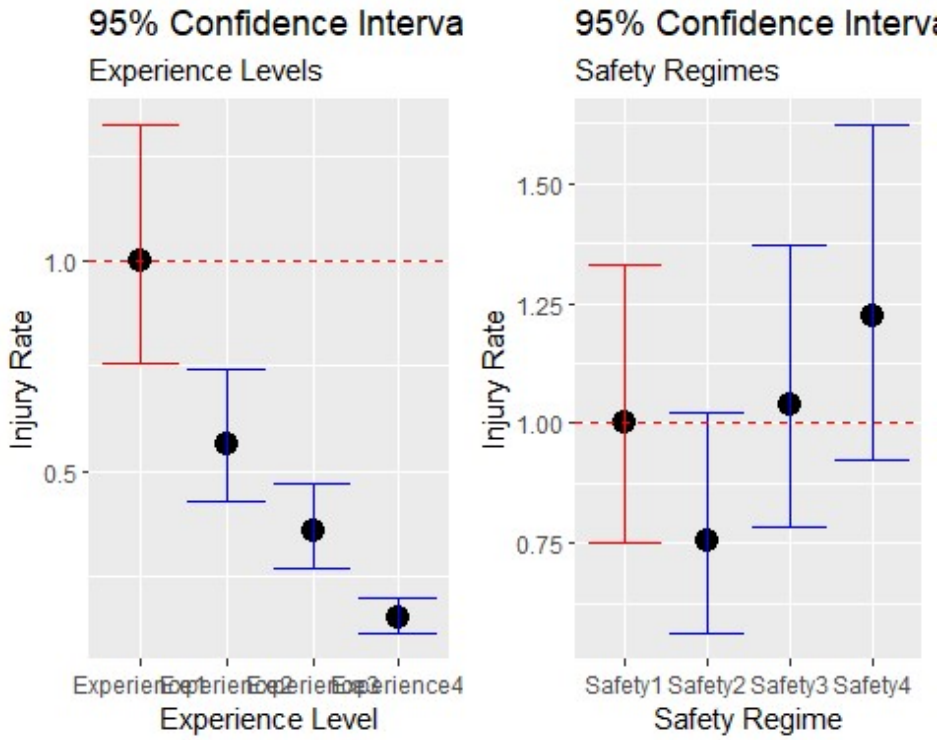
Results

Below is the best final model, given by the Negative Binomial

$$log\left(\frac{Injuries+1}{Hours}\right) = -7.27 - 0.28Safety2 + 0.04Safety3 + 0.20Safety4 -$$

$$0.57Experience2 - 1.03Experience3 - 1.89Experience4$$

```
##    Predictors     Lower RateRatio     Upper
## 1     Safety1 0.7507795 1.0000000 1.3319490
## 2     Safety2 0.5612959 0.7571690 1.0213952
## 3     Safety3 0.7838860 1.0370835 1.3720645
## 4     Safety4 0.9242995 1.2238093 1.6203721
## 5 Experience1 0.7561929 1.0000000 1.3224139
## 6 Experience2 0.4281393 0.5641914 0.7434774
## 7 Experience3 0.2711786 0.3581390 0.4729854
## 8 Experience4 0.1136220 0.1509827 0.2006282
```



The 95% confidence interval plots shown above, compares the proportional differences in injury rate of Experience levels and Safety levels with Experience 1 and Safety 1 respectively. The Experience plot shows that:

Experience2 has 0.56% effect on Injury Rate as Experience1.

Experience3 has 0.35% effect on Injury Rate as Experience1.

Experience4 has 0.15% effect on Injury Rate as Experience1.

This means that experience4 is the most ideal level within the experience group as it has the lowest rate ratio. Someone in Experience4 has an average 0.85% less likely chance of being injuried per hour, compared to someone in Experience1, adjusted for Safety regime.

The Safety Regime plot shows that:

Safety2 has 0.76% effect on Injury Rate as Safety1.

Safety3 has 1.03% effect on Injury Rate as Safety1.

Safety4 has 1.22% effect on Injury Rate as Safety1.

This means that Safety2 is the most ideal level within the Safety group as it has the lowest rate ratio. Someone in Safety2 has an average 0.24% less likely chance of being injured per hour, compared to someone in Safety1, adjusted for Experience. However, Safety2 has a confidence interval of 1.02% to 0.56%, as the confidence interval still includes 1, we can not be sure whether Safety2 will be bigger or smaller than Safety1. Thus Safety2 is statistically insignificant.

Discussion

Discussing the Research Questions:

1. The given data and modelling suggest that Safety Regime 2 should be implemented as the international company-wide Safety standard as it produced the lowest Injury rate. Lowest injury rate is ideal as it suggests that employees are injuried less in the workplace, regardless of how many hours they work.

2. The summary of the final model suggests that experience is more important at reducing the injury rate than safety regime as the coefficients for Experience decrease the injury rate much more than the coefficients of safety. For the biggest reduction in Injury with experience, for every 1 unit increase in Experience4, a 1.89 unit decrease in Injury Rate occurs. For the smallest reduction in Injury Rate with experience, for every 1 unit increase in Experience2, a 0.57 unit decrease in Injury Rate occurs. This is in comparison to the coefficients of Safety's levels which are a closer to 0. Safety coefficients Safety2, Safety3, Safety4 were: -0.27, 0.04 and 0.2 respectively, were values close to 0. This means that for every 1 unit increase in Safety levels, a small change in Injury Rate occurs compared to Experience. This shows that Experience was more important to preventing Injury Rate than Safety Regime.