Workplace Injury Analysis

Yonten Loday

2024-09-12

Introduction

This report analyses the workplace injury data to address concerns raised by the CEO regarding injury rates across various safety regimes and associated factors. The data pertains to a global industrial manufacturing company employing over 10,000 workers. Following a recent incident in South America, the company's safety practices have come under scrutiny. The aim of this analysis is to determine the most effective safety regime in terms of injury prevention, assess the influence of employee experience on injury rates, and evaluate whether bonuses, external safety training, and formal qualifications have any impact on injury counts.

Research Questions

The CEO has posed the following questions:

1. Of the various safety regimes in place, which one would you recommend become the international standard based solely on injury prevention performance?
2. Is industry experience more important than the safety regime in preventing injuries?
3. Are injuries related to the following:
    - 3.1 Annual bonuses a proportion of employees receive?
    - 3.2 Whether staff have received formal external qualifications, e.g., external safety training or university degrees?

Summary of available data

- Injuries: Count of injuries in each group.
- Safety: Safety regime in place for each group (1 to 4).
- Hours: Total hours worked by each group.
- Experience: The average experience level in each group (1 to 4).
- Bonus: Proportion of the group who received an annual bonus.
- Training: Proportion of the group who completed external safety training.
- University: Proportion of the group who have at least one university degree.

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(MASS)
library(ggpubr)
library(gridExtra)
```

```r
library(DHARMa)
library(AER)
```

Data Processing

No extensive preprocessing was necessary for this analysis, as the dataset was well-structured, clean, and free from missing values, which allowed for a straightforward modeling process. The continuous variables, such as Hours and the proportion-based variables like Bonus, Training, and University qualifications, were retained in their original form since they did not require any transformation or imputation. In contrast, the categorical variables, Safety and Experience , were converted into factors to ensure they were appropriately handled in both the statistical models and visualizations. Factoring these variables allows the model to interpret each level of Safety and Experience as distinct categories rather than continuous data, which is crucial for evaluating their specific effects on injury counts. This straightforward preprocessing ensured that the dataset was ready for both the exploratory data analysis (EDA) and the modeling phases, with minimal data transformation required to preserve the integrity of the original variables.

```r
# Load the dataset
injury_data <- read.csv("injury.csv")

# Preview the dataset
head(injury_data)
```

```
##   X Injuries Safety Experience  Hours bonus training university
## 1 1        6      1          4 231437  0.27     0.35       0.73
## 2 2        8      1          4 126655  0.37     0.33       0.45
## 3 3        7      1          4  87847  0.57     0.48       0.18
## 4 4       19      1          3 222970  0.91     0.89       0.75
## 5 5       39      1          3 376438  0.20     0.86       0.10
## 6 6       57      1          2 316462  0.90     0.39       0.86
```

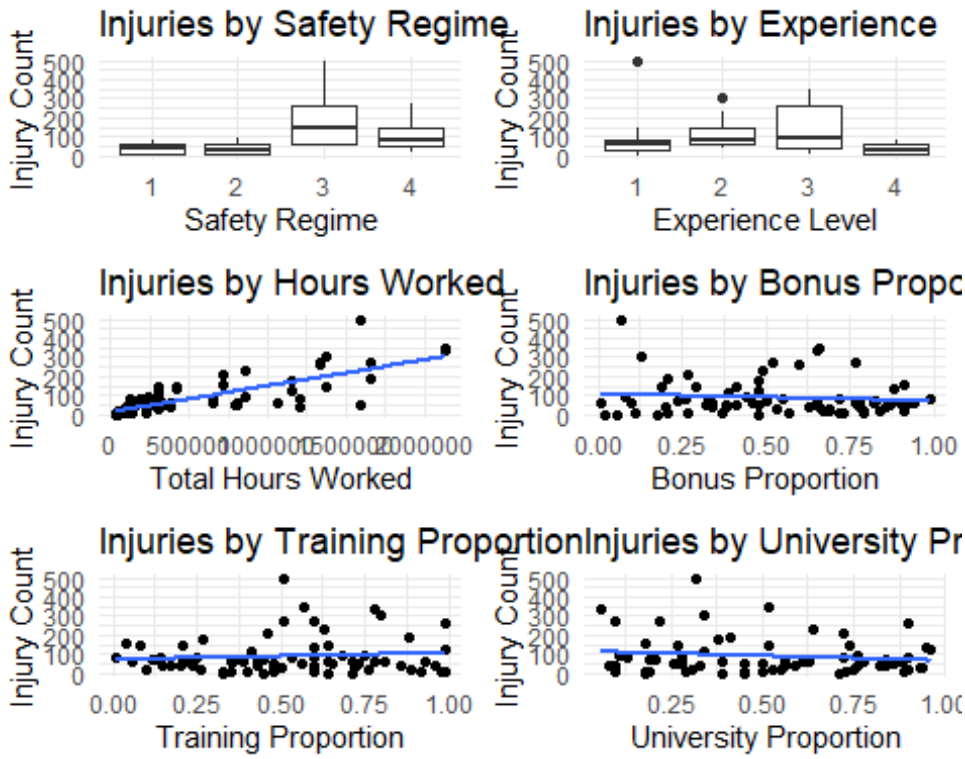
```r
# Summary of the dataset
summary(injury_data)
```

```
##        X            Injuries          Safety       Experience
##  Min.   : 1.00   Min.   :  0.00   Min.   :1.00   Min.   :1.0
##  1st Qu.:18.75   1st Qu.: 25.00   1st Qu.:1.75   1st Qu.:1.0
##  Median :36.50   Median : 57.50   Median :2.50   Median :2.5
##  Mean   :36.50   Mean   : 87.46   Mean   :2.50   Mean   :2.5
##  3rd Qu.:54.25   3rd Qu.: 96.50   3rd Qu.:3.25   3rd Qu.:4.0
##  Max.   :72.00   Max.   :491.00   Max.   :4.00   Max.   :4.0
##      Hours             bonus           training        university
##  Min.   :  34574   Min.   :0.0100   Min.   :0.0100   Min.   :0.0600
##  1st Qu.: 130272   1st Qu.:0.3125   1st Qu.:0.2675   1st Qu.:0.2800
##  Median : 302879   Median :0.4950   Median :0.5050   Median :0.5200
##  Mean   : 549996   Mean   :0.5140   Mean   :0.5096   Mean   :0.5189
##  3rd Qu.: 813381   3rd Qu.:0.7400   3rd Qu.:0.7125   3rd Qu.:0.7600
##  Max.   :2135146   Max.   :0.9900   Max.   :0.9900   Max.   :0.9600
```

```r
# Convert the Safety and Experience variables to a factor
injury_data$Safety <- as.factor(injury_data$Safety)
injury_data$Experience <- as.factor(injury_data$Experience)
```

Exploratory data analysis (EDA)

```r
p1 <- ggplot(injury_data, aes(x = Safety, y = Injuries)) +
  geom_boxplot() +
  labs(title = "Injuries by Safety Regime", x = "Safety Regime", y = "Injury
Count") +
  theme_minimal()

p2 <- ggplot(injury_data, aes(x = factor(Experience), y = Injuries)) +
  geom_boxplot() +
  labs(title = "Injuries by Experience", x = "Experience Level", y = "Injury
Count") +
  theme_minimal()

p3 <- ggplot(injury_data, aes(x = Hours, y = Injuries)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Injuries by Hours Worked", x = "Total Hours Worked", y =
"Injury Count") +
  theme_minimal()

p4 <- ggplot(injury_data, aes(x = bonus, y = Injuries)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Injuries by Bonus Proportion", x = "Bonus Proportion", y =
"Injury Count") +
  theme_minimal()

p5 <- ggplot(injury_data, aes(x = training, y = Injuries)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Injuries by Training Proportion", x = "Training Proportion",
y = "Injury Count") +
  theme_minimal()

p6 <- ggplot(injury_data, aes(x = university, y = Injuries)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Injuries by University Proportion", x = "University
Proportion", y = "Injury Count") +
  theme_minimal()

# Arrange the plots in a grid
grid.arrange(p1, p2, p3, p4, p5, p6, ncol = 2)
```



The visualizations show key insights into the relationships between injury counts and various factors such as safety regime, experience, hours worked, bonuses, training, and university qualifications. From the box plot of safety regimes, it is evident that **Safety Regime 3** exhibits a wider range of injuries, suggesting a higher injury risk compared to the other regimes, particularly **Safety Regime 1**, which has a lower median injury count. The **Experience** plot indicates that higher experience levels, especially **Experience Level 1 and 4**, are associated with fewer injuries, while **Experience Level 3** shows a wider distribution of injury counts.

The scatter plot of **Hours Worked** reveals a positive relationship between hours worked and injuries, indicating that longer working hours are associated with higher injury counts. However, **Bonus Proportion, Training Proportion, and University Proportion** do not show clear trends in relation to injury counts. The flat trend lines in these scatter plots suggest that these variables have minimal impact on the injury rates, consistent with the findings from the regression analysis.

Modelling Approach and Justification

Given the nature of the data, which consists of count data (Injuries), the first step is to implement a **Poisson regression model**. Poisson regression is commonly used for count data where the mean and variance of the counts are assumed to be equal. However, this assumption may not always hold, particularly when there is more variability than expected (i.e., overdispersion). Overdispersion can lead to the underestimation of standard errors in the Poisson model, potentially resulting in misleading conclusions about the significance of the predictors.

To determine if overdispersion is present, diagnostic tests such as a **dispersion test** and **residual checks** will be performed using the DHARMa package. DHARMa provides a simulation-based approach for testing overdispersion and model fit, allowing for both visual inspection and statistical testing of residuals. The **Kolmogorov-Smirnov (KS) test** and the dispersion test will be used to detect any deviations from the expected distribution of residuals, which may suggest overdispersion or other model assumption violations. If significant overdispersion is detected, the Poisson model will be replaced with a **Negative Binomial regression model**, which extends the Poisson model by accounting for overdispersion through an additional parameter that allows the variance to exceed the mean. This two-step approach ensures that the simplest model is used while maintaining accuracy and reliability in the final results.

After fitting the appropriate model, **stepwise variable selection** will be employed to check the significance of predictors. However, if this method is not to refine the model (at least for this analysis) because to address all concerns, it is mandatory to include all variables as each variable is related to at least one concern of the CEO.

```
full_model <- glm(data = injury_data,
                  formula = Injuries ~ Safety + Experience + Hours+ bonus +
training + university,
                  family = poisson(link = "log"))
# Perform stepwise selection
backward_sel_model <- stepAIC(full_model, direction = "backward",trace = 0)

#Model with no variables present for forwards selection:
null_model <- glm(data = injury_data,formula = Injuries ~ 1 ,family =
poisson(link = "log"))
forward_sel_model <- stepAIC(null_model,scope = formula(full_model),
direction = "forward", trace = 0)
formula(backward_sel_model)

## Injuries ~ Safety + Experience + Hours + training + university

formula(forward_sel_model)

## Injuries ~ Hours + Experience + Safety + training + university

AIC(backward_sel_model)

## [1] 1753.136

AIC(forward_sel_model)

## [1] 1753.136
```

As we can see in the predictors defined by stepwise variable selection, it did not identify bonus as signifcant predictor but we must enforce bonus variable due to the CEO's specific question about the relation between bonus propotion and injury counts. So, therefore, the full model with bonus variable is also used even though it is not identified by step wise method.

```r
summary(full_model)
```

```
##
## Call:
## glm(formula = Injuries ~ Safety + Experience + Hours + bonus +
##     training + university, family = poisson(link = "log"), data =
injury_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.606e+00  7.282e-02  49.519  < 2e-16 ***
## Safety2     -6.596e-02  5.777e-02  -1.142  0.25351
## Safety3      5.734e-01  6.105e-02   9.391  < 2e-16 ***
## Safety4      6.809e-01  5.139e-02  13.249  < 2e-16 ***
## Experience2  1.494e-01  3.534e-02   4.227 2.37e-05 ***
## Experience3 -2.515e-01  4.842e-02  -5.194 2.06e-07 ***
## Experience4 -1.135e+00  4.875e-02 -23.285  < 2e-16 ***
## Hours        9.021e-07  3.630e-08  24.853  < 2e-16 ***
## bonus       -3.201e-03  5.726e-02  -0.056  0.95542
## training     1.567e-01  5.602e-02   2.798  0.00515 **
## university  -1.431e-01  5.507e-02  -2.599  0.00936 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 6213.8  on 71  degrees of freedom
## Residual deviance: 1331.4  on 61  degrees of freedom
## AIC: 1755.1
##
## Number of Fisher Scoring iterations: 5
```

The full Poisson regression model reveals key factors influencing injury counts. Safety regimes show varying effects, with Safety3 and Safety4 associated with significantly higher injury counts compared to the reference category (Safety1), indicating these regimes may be less effective in reducing injuries. In terms of employee experience, Experience2 leads to higher injury counts, while Experience3 and Experience4 significantly reduce injury rates, showing that greater experience generally correlates with fewer injuries. Additionally, hours worked is positively linked to injury counts, suggesting that more hours worked increases injury risk. Other significant variables include training, which is associated with higher injury counts, and university qualifications, which reduce injury rates. However, bonus does not appear to have a significant impact on injury counts with p-value 0.955.

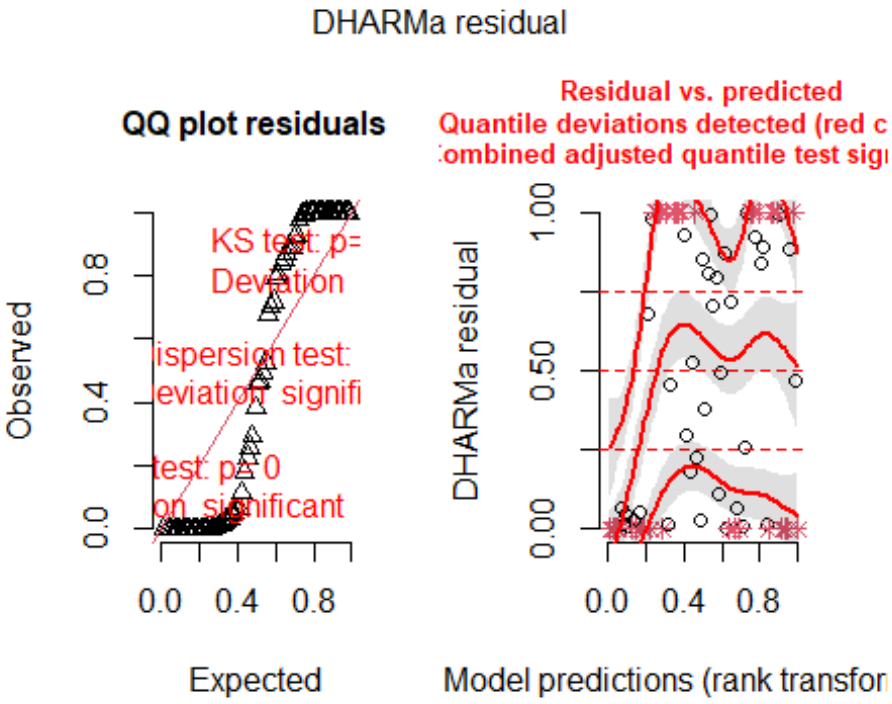
Despite the insights provided by the Poisson model, one of its key assumptions of Poisson model is that the variance of the outcome variable should be equal to its mean. To check whether this assumption holds, a dispersion test will be conducted.

```r
disp_result <- dispersiontest(full_model)
print(disp_result)
```

```
##
##  Overdispersion test
##
## data:  full_model
## z = 4.5213, p-value = 3.072e-06
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   16.47374
```

The results of the overdispersion test indicate significant overdispersion in the data, as shown by a high z-value (4.5188) and a very small p-value (3.109e-06). The estimated dispersion of 16.48 confirms that the variance far exceeds the mean, violating the Poisson model's assumption of equal mean and variance.

```
poisson_residuals = simulateResiduals(full_model)
plot(poisson_residuals)
```



The DHARMa residual diagnostics indicate significant deviations from the expected distribution for the Poisson model. The Kolmogorov-Smirnov (KS) test shows a p-value of 0, indicating that the observed residuals significantly deviate from the expected distribution. Similarly, the dispersion test with a p-value of 0 confirms that overdispersion is present in the data, meaning that the variance is greater than the mean, violating the assumptions of the Poisson model. Additionally, the outlier test suggests the presence of outliers. The right-hand plot, showing the residuals versus predictions, also reveals

quantile deviations (red curves), indicating that the model fails to capture the variability in the data adequately.

Both dispesion test and residual checks conform that Poisson model's key assumption about equal mean and its outcome variance is not satisfied. Therefore, a Negative Binomial regression model, which accounts for overdispersion, would be more appropriate for this dataset.

```
nb_full_model <- glm.nb(data = injury_data,
                  formula = Injuries ~ Safety + Experience + Hours +bonus +
training + university,
                  link = "log")
nb_null_model <- glm.nb(data = injury_data,
                        formula = Injuries ~ 1,
                        link = "log")
#backward and forward selction
nb_backward_sel_model <- stepAIC(object = nb_full_model,direction =
"backward",trace = 0)
nb_forward_sel_model <- stepAIC(nb_null_model,scope = formula(nb_full_model),
direction = "forward",trace = 0)

formula(nb_backward_sel_model)

## Injuries ~ Safety + Experience + Hours

formula(nb_forward_sel_model)

## Injuries ~ Hours + Experience + Safety
```

As we can see in the predictors defined by stepwise variable selection, just like in Poisson model, it identified only Safety + Experience + Hours. However, to answer all questions with one fitted model, all predictors are necessary.

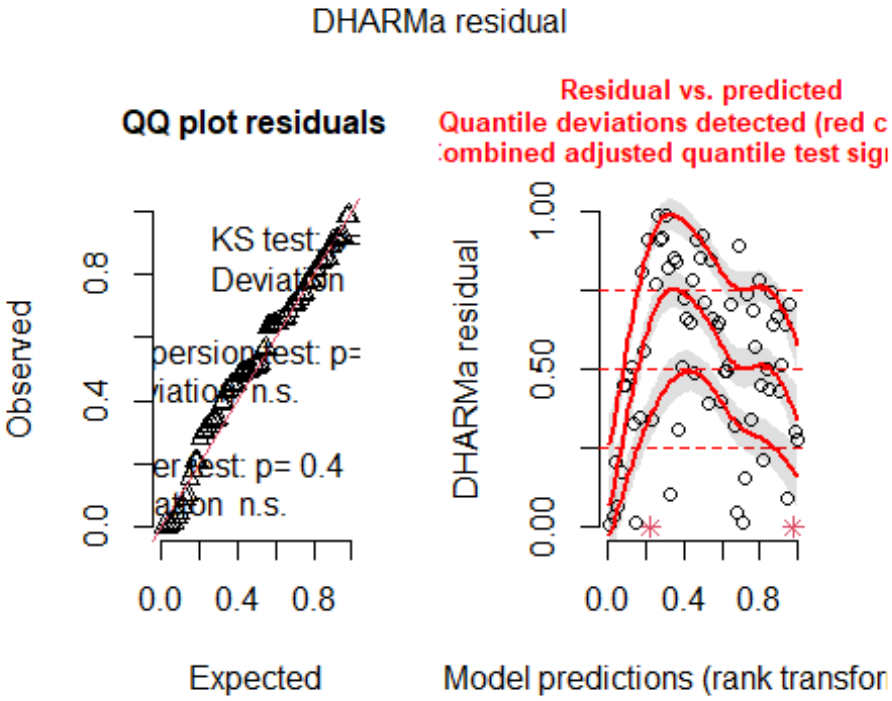
```
summary(nb_full_model)

##
## Call:
## glm.nb(formula = Injuries ~ Safety + Experience + Hours + bonus +
##     training + university, data = injury_data, link = "log",
##     init.theta = 3.139459534)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.634e+00  3.315e-01  10.962  < 2e-16 ***
## Safety2     -2.476e-01  2.108e-01  -1.174  0.24027
## Safety3      5.071e-01  2.844e-01   1.783  0.07459 .
## Safety4      6.516e-01  2.323e-01   2.805  0.00503 **
## Experience2 -1.710e-02  2.057e-01  -0.083  0.93374
## Experience3 -5.696e-01  2.533e-01  -2.249  0.02450 *
## Experience4 -1.576e+00  2.019e-01  -7.806 5.90e-15 ***
## Hours        1.272e-06  2.184e-07   5.824 5.76e-09 ***
```

```
## bonus         2.679e-01  2.848e-01   0.941  0.34683
## training     -2.882e-01  2.828e-01  -1.019  0.30806
## university    4.540e-03  2.745e-01   0.017  0.98680
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(3.1395) family taken to be 1)
##
##     Null deviance: 289.001  on 71  degrees of freedom
## Residual deviance:  81.152  on 61  degrees of freedom
## AIC: 711.26
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  3.139
##           Std. Err.:  0.588
##
##  2 x log-likelihood:  -687.258
```

The Negative Binomial (NB) model has been fitted to account for overdispersion in the injury data, which was identified earlier with Poisson model. The NB model's intercept of 3.634 indicates the baseline log injury rate when all other variables are at their reference levels (defaults). Among the predictor variables, **Safety Regime 4** (p-value = 0.00503) shows a significant and positive effect, meaning it is associated with an increase in injury counts compared to the reference level (Safety Regime 1). **Safety Regime 3** (p-value = 0.07459) has a marginal effect, indicating that it may also be linked to higher injury counts, though the evidence is less strong. On the other hand, **Safety Regime 2** (p-value = 0.24027) is not statistically significant, suggesting that it does not differ significantly from Safety Regime 1 in terms of injury rates.

Regarding experience, **Experience Level 4** (p-value < 2e-16) and **Experience Level 3** (p-value = 0.02450) are significantly associated with **lower injury rates** compared to Experience Level 1, showing that more experienced employees tend to have fewer injuries. However, **Experience Level 2** (p-value = 0.93374) is not statistically significant, implying that it has a similar injury count to Experience Level 1. Additionally, **Hours worked is highly significant** (p-value < 5.76e-09), indicating a strong positive correlation between hours worked and injury counts. On the other hand, **bonus, training, and university qualifications** do not significantly influence injury counts in this model. The dispersion parameter (Theta = 3.139) shows that the Negative Binomial model successfully addresses overdispersion, with improved residual deviance (81.152) and AIC (711.26) compared to the Poisson model.

```
nb_all_residuals = simulateResiduals(nb_full_model)
plot(nb_all_residuals)
```



The residual plot from the Negative Binomial (NB) model (first image) shows a good fit, with the residuals closely following the expected line in the QQ plot. The KS test (p-value = 0.48495) and the dispersion test (p-value = 0.616) indicate no significant deviations from the expected distribution, suggesting that the NB model adequately addresses overdispersion and produces a good fit for the data. This is further supported by the outlier test (p-value = 0.42), which shows no significant deviation from expected values.

In comparison, the residual plot from the Poisson model showed significant deviations in both the QQ plot and the residuals vs. predicted plot. The KS test and dispersion test both yield p-values of 0, indicating significant deviations, which confirm overdispersion in the Poisson model. The red quantile curves in the residual vs. predicted plot further highlighted these deviations, showing that the Poisson model struggles to capture the variability in the data accurately. This clearly demonstrates that the NB model is a better fit for the injury data.

Validity of model and modelling results

The Negative Binomial model was chosen based on its ability to handle overdispersion which Poisson model struggles to address, as indicated by the results of the DHARMa residuals and dispersion tests. The final model showed:

- Experience Level 3 (p-value = 0.02450): Shows a negative relationship with injury rates, meaning that more experienced workers have fewer injuries.
- Experience Level 4 (p-value < 2e-16): Also significantly decreases injury rates.

- Hours Worked (p-value = 5.76e-09): Positively associated with injury rates, indicating that more hours worked increases the likelihood of injuries.
- Safety Regime 4 (p-value = 0.00503): Positively related to injury counts, suggesting that this regime may lead to more injuries compared to others.
- Safety Regime 2 (p-value = 0.24027) and Experience Level 2 (p-value = 0.93374) show no significant effect on injuries.
- Bonus, Training, and University qualifications are not statistically significant in predicting injury counts.

Recommendations and conclusions

1. Recommendation for International Standard Based on Safety Regimes:

Based on the Negative Binomial (NB) model, Safety Regime 4 showed a significant positive relationship with injury counts (p-value = 0.00503), indicating that it is associated with higher injury rates. Safety Regime 3 also had some evidence of a positive association with injuries (p-value = 0.07459), while Safety Regime 2 did not show a significant effect on injuries (p-value = 0.24027).

Recommendation: Safety Regime 1 or Safety Regime 2 should be considered as the international standard for injury prevention, as both showed fewer injuries in comparison to the other regimes. Safety Regime 4, in particular, should be avoided due to its association with increased injuries.

2. Is Industry Experience More Important than Safety Regime?

As evident from the NB model results, Experience levels 3 and 4 are highly significant in reducing injury rates, with Experience Level 4 having a p-value < 2e-16, and Experience Level 3 having a p-value of 0.02450. This indicates that workers with more experience have significantly fewer injuries. However, Experience Level 2 was not statistically significant (p-value = 0.93374) in comparison to Experience level 1, meaning it does not impact injury rates.

Conclusion: Both Experience Level 3 and Experience Level 4 significantly reduce injuries, while certain safety regimes (like Safety Regime 4) are associated with more injuries. However, overall, Experience seems to have a stronger and more consistent impact on injury prevention compared to Safety Regimes, particularly for highly **experienced workers**.

3.1 Relationship Between Injuries and Bonuses:

As evident from the NB model the coefficient for bonus is not statistically significant (p-value = 0.34683), meaning there is no evidence to suggest that annual bonuses are related to injury rates.

Conclusion: Injuries are not related to the proportion of employees receiving annual bonuses based on the available data.

3.2 Relationship Between Injuries and Formal Qualifications (Training and University Degrees):

- **Training:** The model shows that training is not statistically significant in predicting injury counts (p-value = 0.30806). This suggests that external safety training has no measurable impact on reducing injuries.
- **University Degrees:** The coefficient for university is also not statistically significant (p-value = 0.98680), meaning formal education, such as a university degree, has no relationship with injury rates.

Conclusion: There is no evidence that receiving formal external qualifications, such as external safety training or university degrees, has any significant effect on injury rates based on the available data.

Overall Summary

In summary, the analysis of workplace injury data has provided important insights into the factors that influence injury rates within the company. The Negative Binomial model was found to be a better fit for the data compared to the Poisson model, primarily due to the presence of overdispersion. The results indicate that certain safety regimes, particularly Safety Regime 4, are associated with higher injury rates, while Safety Regime 1 and Safety Regime 2 are preferable for minimizing injuries. This suggests that a more cautious approach should be taken when implementing certain safety regimes.

Industry experience also plays a critical role in injury prevention. Employees with higher levels of experience, particularly those in Experience Level 4 and Experience Level 3, demonstrate significantly lower injury rates. This highlights the importance of retaining experienced employees and possibly focusing on targeted training programs that build expertise over time.

Finally, the analysis found no significant relationship between injury counts and factors such as bonuses, formal external safety training, or university qualifications. This suggests that while financial incentives and formal education are important, they do not directly contribute to reducing workplace injuries in the dataset studied. Therefore, the company may benefit more from focusing on improving safety regimes and leveraging industry experience to enhance overall safety.