Injury Statistical Analysis

Introduction

In our global industrial company, over 10,000 employees work across multiple factories worldwide. We have operations in Asia, Europe, and South America, each regulated by its own industrial standards and safety practices. A recent incident in South America has raised concerns about the effectiveness of our company's workplace practices.

Research questions

It is necessary to perform an analysis of workplace injury data to answer the following questions:

1. Of the various safety regimes in place across the company, which one would be recommended to become the international standard for our company, based solely on injury prevention performance?
2. It has been suggested by senior management that industry experience is more important than the safety regime when it comes to preventing injuries. The idea is that a policy should be developed that is directly related to lowering employee turnover will reduce injury rates. Do the available data support this assertion?
3. If there is any relationship between:

- Injuries and the annual bonuses a proportion of employees receive.
- Injuries and whether staff have received any formal external qualifications, e.g., external safety training or a university degree.

Summary of available data

The data for analysis is taken from the file 'injury.csv'. It contains counts of injuries and hours worked aggregated by the experience level of the workers and the workplace safety regime in place at their factory. The data are for the last 12 months of operation.

Specifically, the variables are:

- Injuries - count of injuries in group
- Safety - the safety regime in place for group
- Hours - total hours worked by group
- Experience - the experience level of group
- bonus - proportion of group who received an annual bonus last year
- training - proportion of group who have completed external safety training
- university - proportion of group who have at least one university degree

Data processing

Download necessary libraries:

```
library(tidyverse)
library(MASS)
library(ggpubr)
library(DHARMa)
library(AER)
```

Load the raw data and check if preprocessing is needed.

```
#Load the raw data:
injury_data <- read.csv(file = "~/Downloads/injury.csv")
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
```

```
#Check data structure and data types
str(injury_data)
```

```
## 'data.frame':    72 obs. of  8 variables:
##  $ X         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Injuries  : int  6 8 7 19 39 57 60 40 40 1 ...
##  $ Safety    : int  1 1 1 1 1 1 1 1 2 ...
##  $ Experience: int  4 4 4 3 3 2 2 1 1 1 ...
##  $ Hours     : int  231437 126655 87847 222970 376438 316462 235670 127361 96667 51156 ...
##  $ bonus     : num  0.27 0.37 0.57 0.91 0.2 0.9 0.94 0.66 0.63 0.06 ...
##  $ training  : num  0.35 0.33 0.48 0.89 0.86 0.39 0.78 0.96 0.43 0.71 ...
##  $ university: num  0.73 0.45 0.18 0.75 0.1 0.86 0.61 0.56 0.33 0.45 ...
```

```
#Check missing values
colSums(is.na(injury_data))
```

```
##          X   Injuries     Safety Experience      Hours      bonus   training
##          0          0          0          0          0          0          0
## university
##          0
```

To prepare the data for further analysis, the following steps need to be performed:

1. Convert 'Safety' to a Factor: Change the Safety variable to a factor to treat it as a categorical variable.
2. Convert 'Experience' to an Ordinal Factor: Convert the Experience variable to an ordinal factor to reflect the natural ordering of levels.

3. Calculate 'Injuries per Year': Create a new variable, Injuries_per_year, defined as (Injuries / Hours) * 40 * 52 to standardize the number of injuries per typical work year, which provides a more meaningful basis for next exploration analysis.

4. Exclude the column 'X' as it does not make sense for our analysis.

```
#Data preparation:
#Convert 'Safety' to factor
injury_data$Safety <- as.factor(injury_data$Safety)

#Convert 'Experience' to ordinal factor
injury_data$Experience <- factor(injury_data$Experience,
                      levels = c(1, 2, 3, 4),
                      ordered = TRUE)

#Create new column describing injuries per year (40 hours per week * 52 weeks)
injury_data <- injury_data %>%
  mutate(Injuries_per_year = (Injuries / Hours) * 40 * 52)

#Exclude column X from the dataset
injury_data <- injury_data %>%
  dplyr::select(-X)

head(injury_data)
```

```
##   Injuries Safety Experience  Hours bonus training university Injuries_per_year
## 1        6      1          4 231437  0.27     0.35       0.73        0.05392396
## 2        8      1          4 126655  0.37     0.33       0.45        0.13138052
## 3        7      1          4  87847  0.57     0.48       0.18        0.16574271
## 4       19      1          3 222970  0.91     0.89       0.75        0.17724358
## 5       39      1          3 376438  0.20     0.86       0.10        0.21549365
## 6       57      1          2 316462  0.90     0.39       0.86        0.37464214
```

Exploratory data analysis (EDA)

Calculate basic exploratory analysis for our data.

```
#Check data structure and data types
str(injury_data)
```

```
## 'data.frame':    72 obs. of  8 variables:
##  $ Injuries         : int  6 8 7 19 39 57 60 40 40 1 ...
##  $ Safety           : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 2 ...
##  $ Experience       : Ord.factor w/ 4 levels "1"<"2"<"3"<"4": 4 4 4 3 3 2 2 1 1 1 ...
##  $ Hours            : int  231437 126655 87847 222970 376438 316462 235670 127361 96667 51156 ...
##  $ bonus            : num  0.27 0.37 0.57 0.91 0.2 0.9 0.94 0.66 0.63 0.06 ...
##  $ training         : num  0.35 0.33 0.48 0.89 0.86 0.39 0.78 0.96 0.43 0.71 ...
##  $ university       : num  0.73 0.45 0.18 0.75 0.1 0.86 0.61 0.56 0.33 0.45 ...
##  $ Injuries_per_year: num  0.0539 0.1314 0.1657 0.1772 0.2155 ...
```

```r
#Calculate summary statistics
summary(injury_data)
```

```
##     Injuries      Safety Experience     Hours             bonus
##  Min.   :  0.00   1:18   1:20       Min.   :  34574   Min.   :0.0100
##  1st Qu.: 25.00   2:18   2:16       1st Qu.: 130272   1st Qu.:0.3125
##  Median : 57.50   3:18   3:16       Median : 302879   Median :0.4950
##  Mean   : 87.46   4:18   4:20       Mean   : 549996   Mean   :0.5140
##  3rd Qu.: 96.50                     3rd Qu.: 813381   3rd Qu.:0.7400
##  Max.   :491.00                     Max.   :2135146   Max.   :0.9900
##     training        university     Injuries_per_year
##  Min.   :0.0100   Min.   :0.0600   Min.   :0.0000
##  1st Qu.:0.2675   1st Qu.:0.2800   1st Qu.:0.1411
##  Median :0.5050   Median :0.5200   Median :0.3312
##  Mean   :0.5096   Mean   :0.5189   Mean   :0.4004
##  3rd Qu.:0.7125   3rd Qu.:0.7600   3rd Qu.:0.5983
##  Max.   :0.9900   Max.   :0.9600   Max.   :1.2575
```

Conclusion of summary statistics:

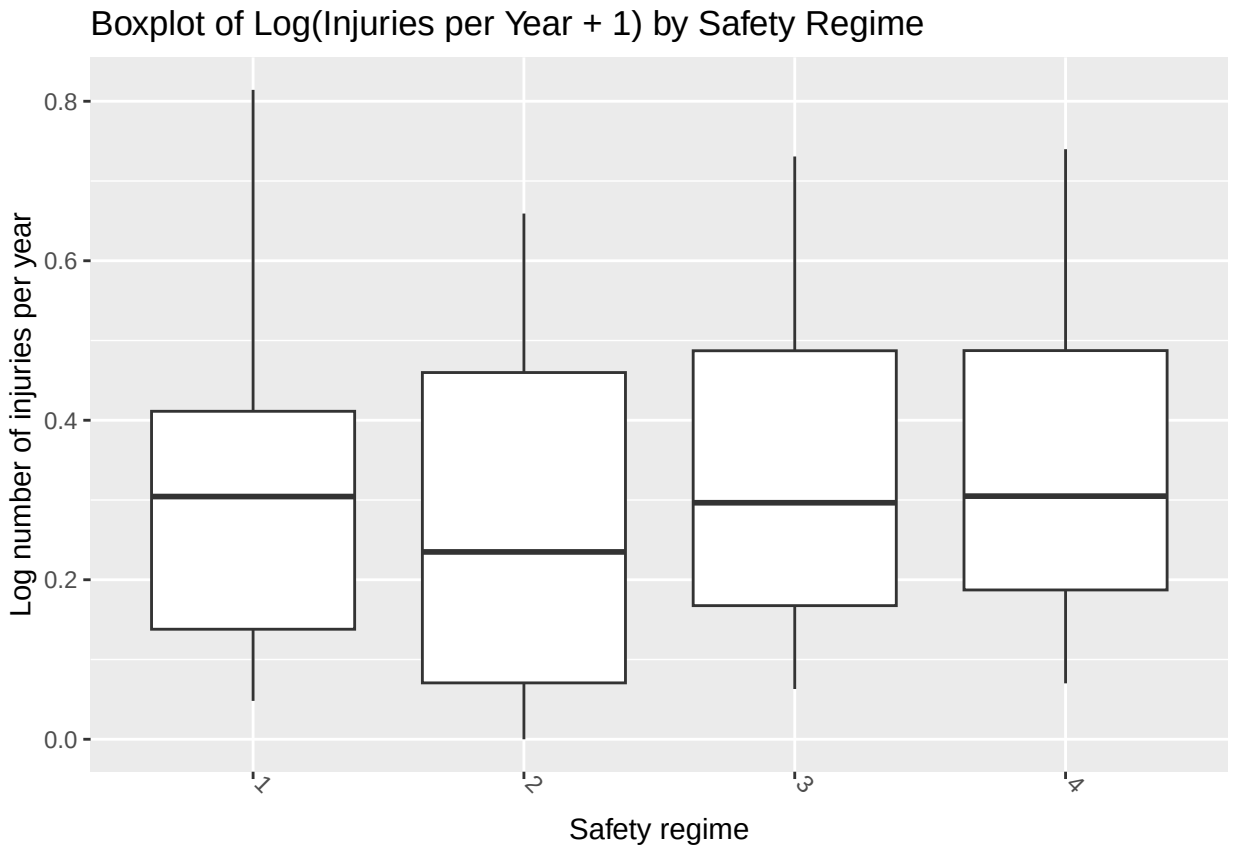
The basic exploratory analysis shows that we have 72 observations of 8 variables. The variable of interest, 'Injuries', has a mean of 87.46 and a range from 0 to 491. This variable is associated with the variable 'Hours'; generally, the more hours worked, the higher the number of injuries. The categorical variable 'Safety' is balanced, with an equal number of observations for each safety regime. The ordinal variable 'Experience' is slightly less balanced. The variables 'bonus', 'training', and 'university' represent the proportion of employees in a group with these characteristics, with values ranging from 0.01 to 0.99 for the first two, and from 0.06 to 0.96 for the last.

Create plots to visualize and explore the data.

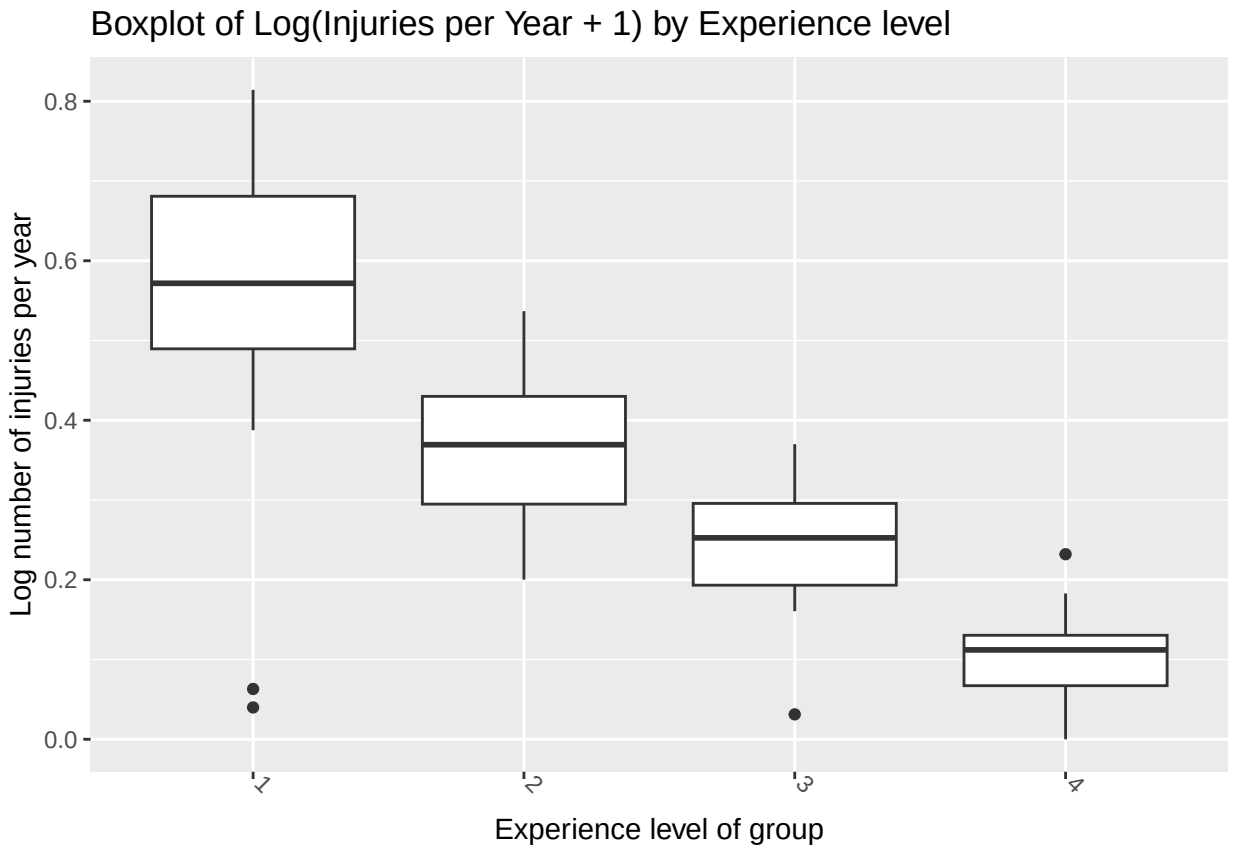
For more meaningful results in visualization, I will use the pre-prepared variable Injuries_per_year, specifically, the log(Injuries_per_year + 1).

Below are the plots that could be useful for our exploratory analysis.

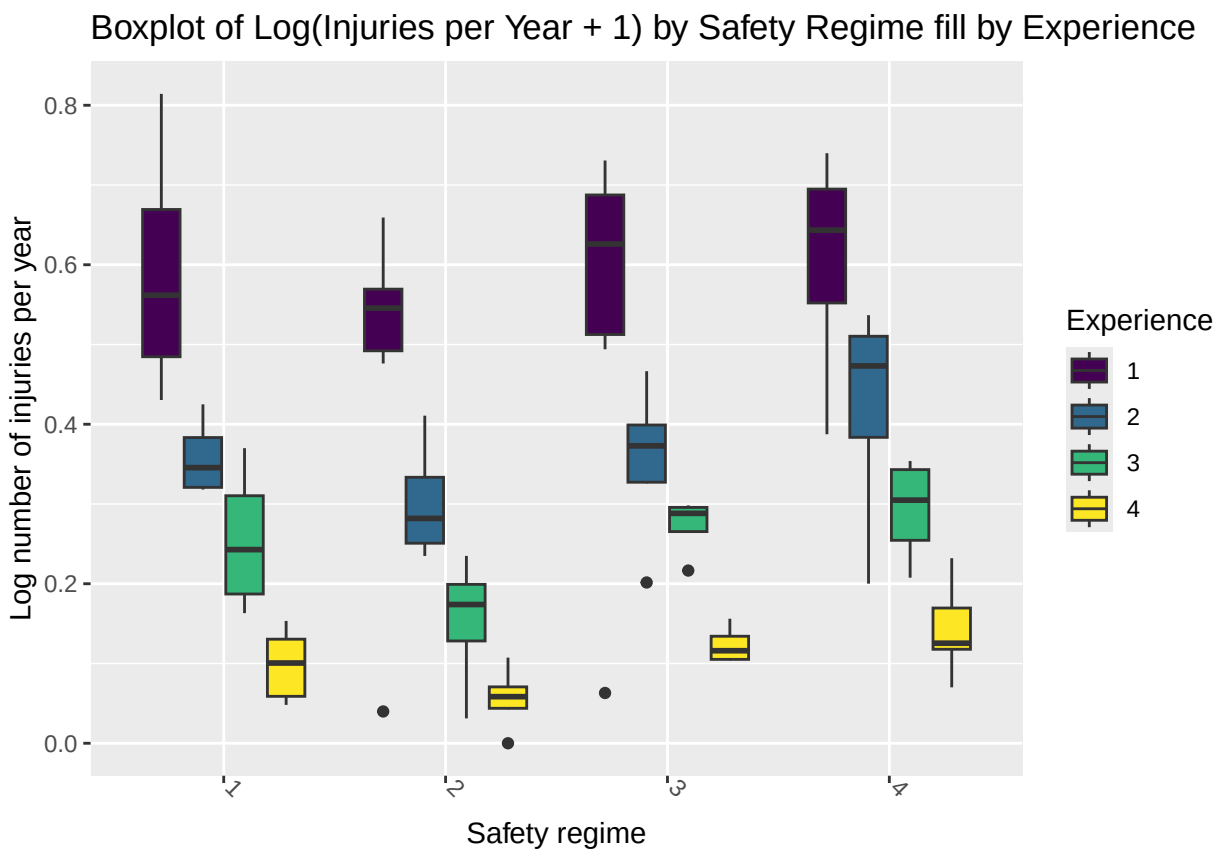
```r
p1 <- ggplot(data = injury_data,
             mapping = aes(
               x = Safety,
               y = log(Injuries_per_year + 1)
             )) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = -45, vjust = 0, hjust = 0)) +
  labs(x = "Safety regime", y = "Log number of injuries per year",
       title = "Boxplot of Log(Injuries per Year + 1) by Safety Regime")
p1
```



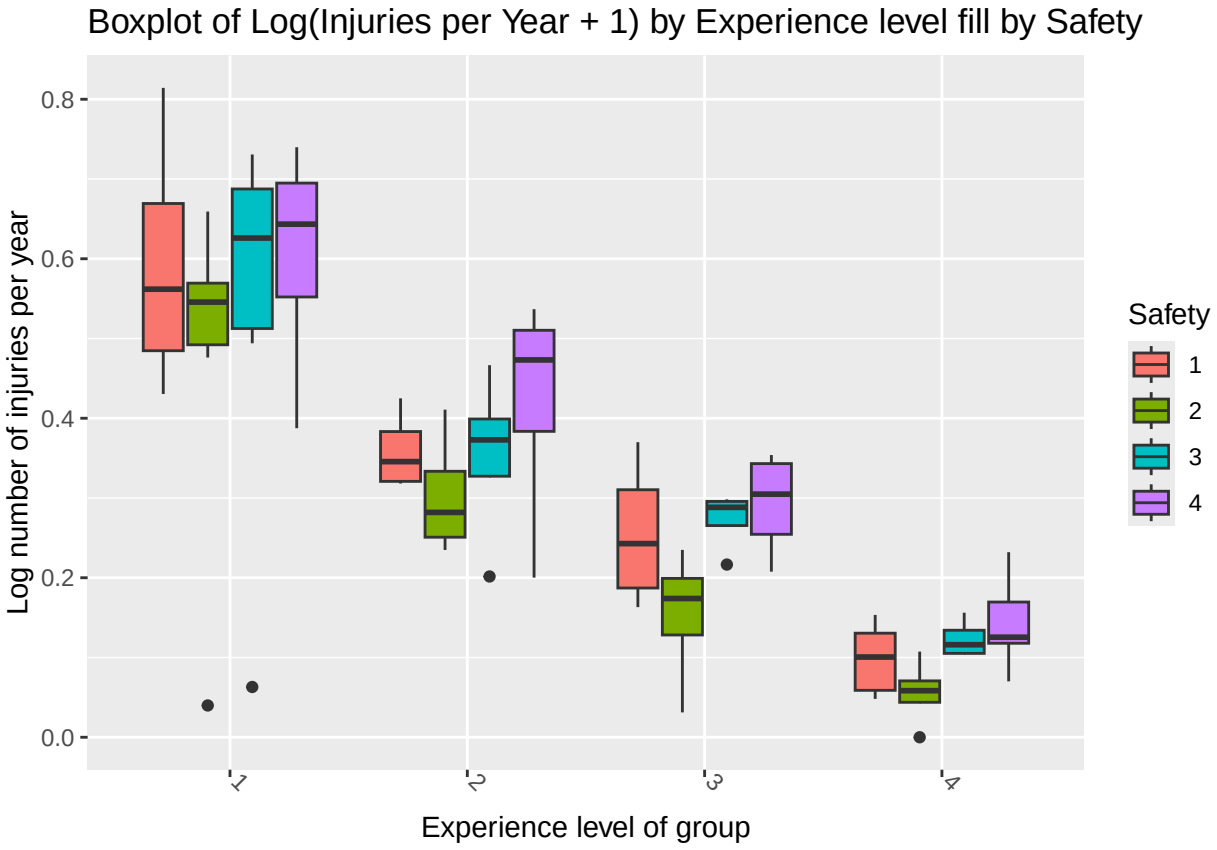

```
p2 <- ggplot(data = injury_data,
             mapping = aes(
               x = Experience,
               y = log(Injuries_per_year + 1)
             )) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = -45, vjust = 0, hjust = 0)) +
  labs(x = "Experience level of group", y = "Log number of injuries per year",
       title = "Boxplot of Log(Injuries per Year + 1) by Experience level")
p2
```




```
p3 <- ggplot(data = injury_data,
             mapping = aes(
               x = Safety,
               y = log(Injuries_per_year + 1),
               fill = Experience
             )) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = -45, vjust = 0, hjust = 0)) +
  labs(x = "Safety regime", y = "Log number of injuries per year",
       title = "Boxplot of Log(Injuries per Year + 1) by Safety Regime fill by Experience")
p3
```

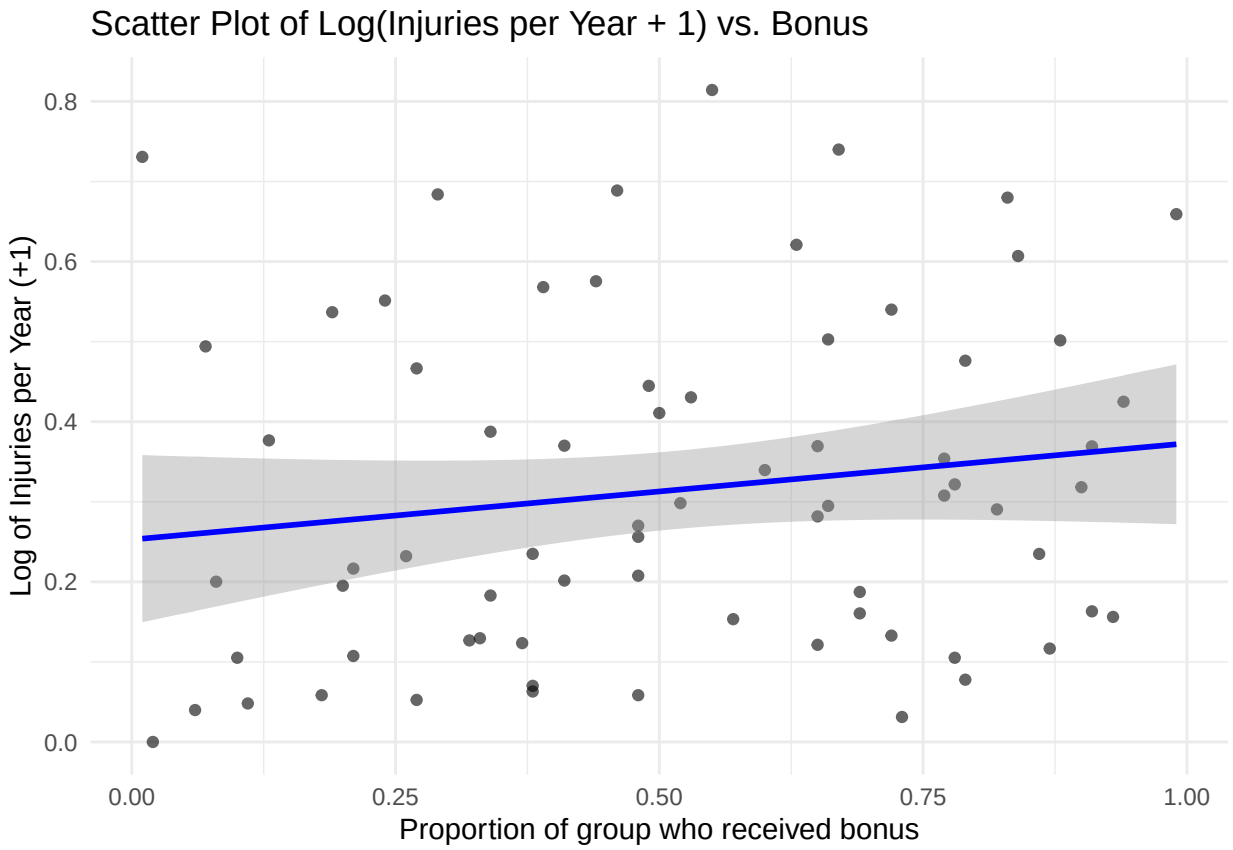



```
p7 <- ggplot(data = injury_data,
             mapping = aes(
               x = Experience,
               y = log(Injuries_per_year + 1),
               fill = Safety
             )) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = -45, vjust = 0, hjust = 0)) +
  labs(x = "Experience level of group", y = "Log number of injuries per year",
       title = "Boxplot of Log(Injuries per Year + 1) by Experience level fill by Safety")
p7
```



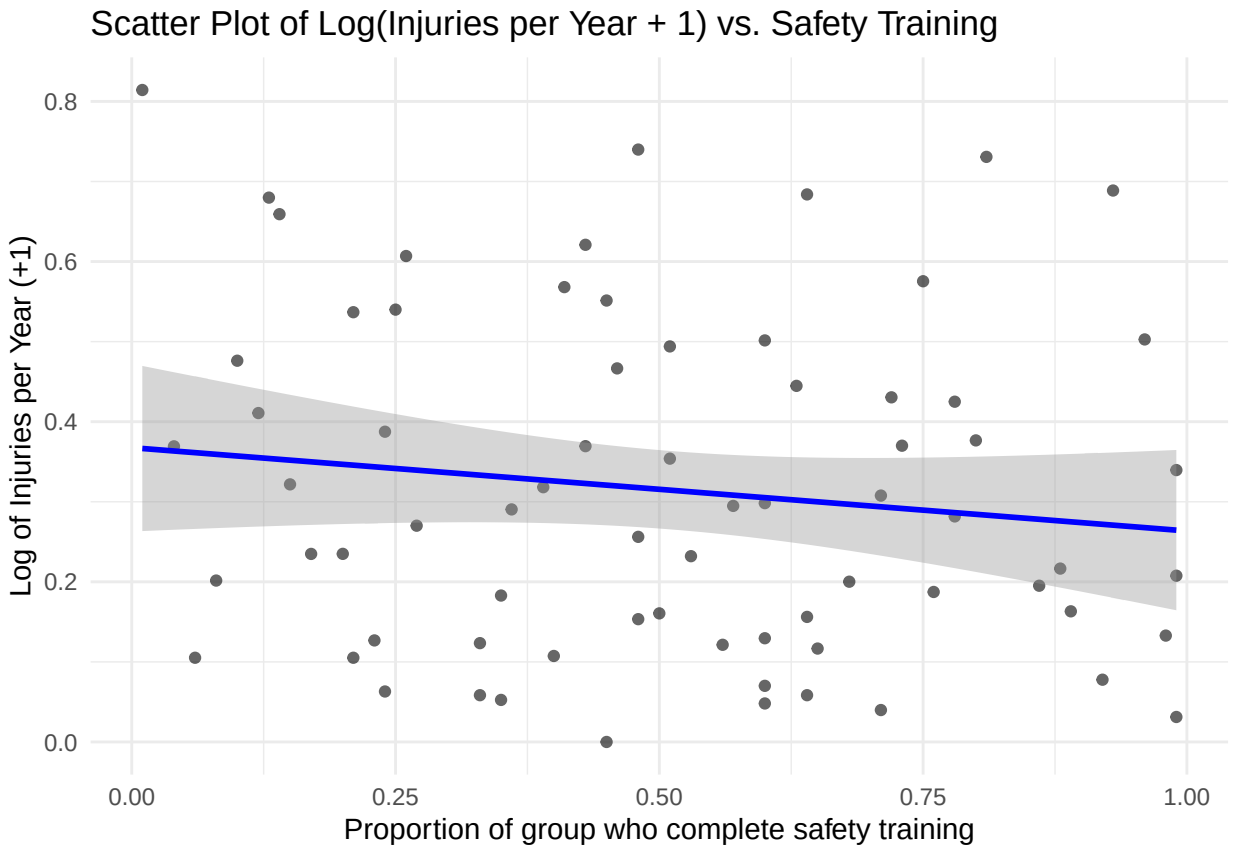

```
p4 <- ggplot(data = injury_data, aes(x = bonus, y = log(Injuries_per_year + 1))) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +  # Add a linear regression line
  labs(x = "Proportion of group who received bonus", y = "Log of Injuries per Year (+1)",
       title = "Scatter Plot of Log(Injuries per Year + 1) vs. Bonus") +
  theme_minimal()
p4
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



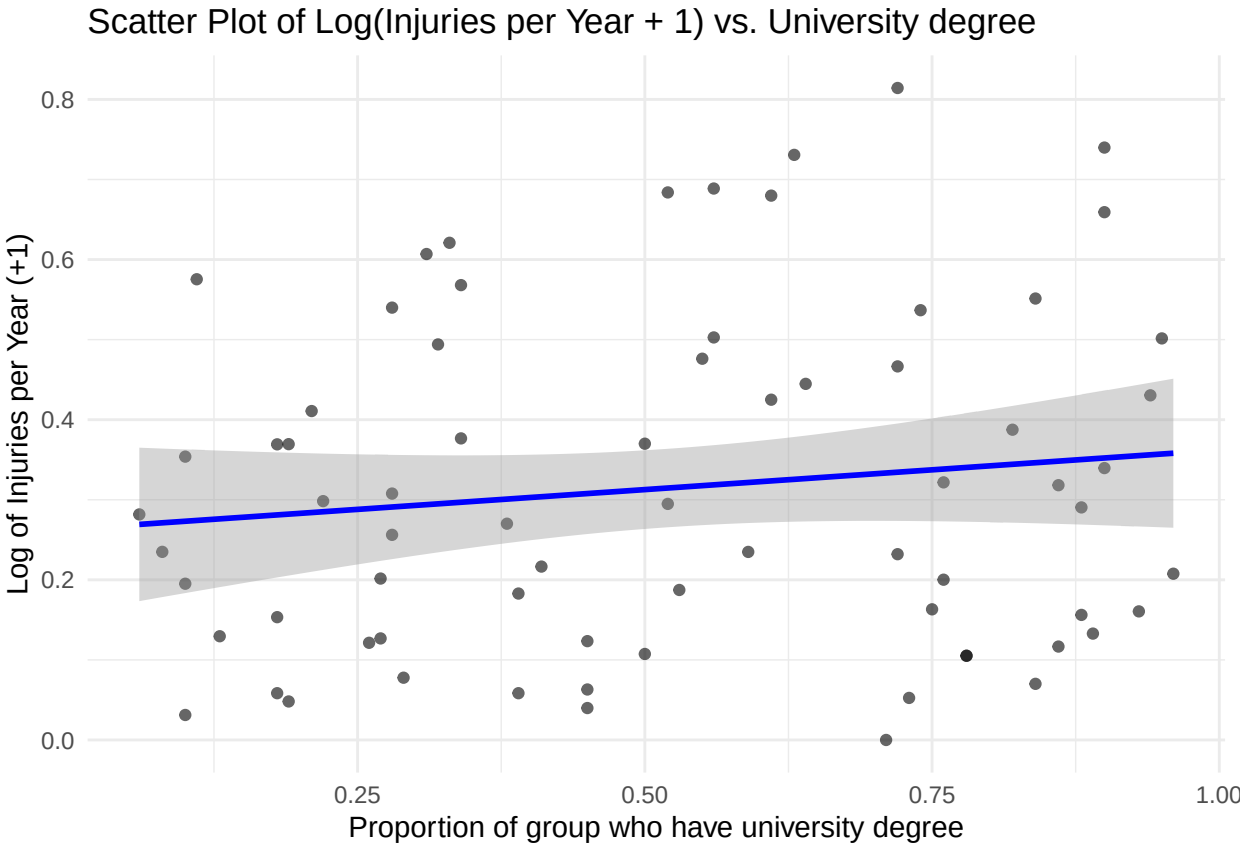

```
p5 <- ggplot(data = injury_data, aes(x = training, y = log(Injuries_per_year + 1))) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +  # Add a linear regression line
  labs(x = "Proportion of group who complete safety training", y = "Log of Injuries per Year (+1)",
       title = "Scatter Plot of Log(Injuries per Year + 1) vs. Safety Training") +
  theme_minimal()
p5
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
p6 <- ggplot(data = injury_data, aes(x = university, y = log(Injuries_per_year + 1))) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +  # Add a linear regression line
  labs(x = "Proportion of group who have university degree", y = "Log of Injuries per Year (+1)",
       title = "Scatter Plot of Log(Injuries per Year + 1) vs. University degree") +
  theme_minimal()
p6
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Based on the exploratory plots, here are the insights:

1. **Influence of Safety Regimes on Injuries:** Among the four safety regimes, we can see that Regime 2 has the lowest median value for log-transformed number of injuries per year, while the other three regimes have similar median values that are slightly higher (see "Boxplot of Log(Injuries per Year + 1) by Safety Regime").
2. **Influence of Experience Level on Injuries:** The visualization analyzing the effect of experience level on the number of injuries shows a clear trend: log-transformed number of injuries per year decreases as the experience level increases. Specifically, Experience Level 4 corresponds to the lowest number of injuries (see "Boxplot of Log(Injuries per Year + 1) by Experience Level"). This trend is consistent across all safety regimes, indicating that within each safety group, experience level has a similar impact (see "Boxplot of Log(Injuries per Year + 1) by Safety Regime fill by Experience")
3. **Influence of Bonus, Safety Training, and University Degree on Injuries:** The scatter plots analyzing the influence of bonus, safety training, and university degree on the log-transformed number of injuries per year do not show a clear trend; the data is quite scattered. Adding a linear regression line helps to identify any potential trends, but the regression lines have very slight slopes, indicating a weak or negligible effect.

Modelling approach and justification

To address the questions, it is best to first create and select the most optimal model and then focus on specific questions (attributes).

Since our target variable, 'Injuries', is a count, I initially chose a Poisson GLM to model it, using 'safety', 'experience', 'bonus', 'training', and 'university' as covariates and 'Hours' as an offset variable.

Stepwise selection based on AIC was used to determine which variables should be included in the model, applying both backward and forward selection methods.

We will also compare the performance of a Negative Binomial model if needed.

Poisson Model

```
#Full model with all possible interactions for backwards selection:
full_interaction_model <- glm(data = injury_data,
    formula = Injuries ~ Safety + Experience +  bonus + training + university,
    offset = log(Hours),
    family = poisson(link = "log"))

#Model with no variables present for forwards selection:
null_model <- glm(data = injury_data,
    formula = Injuries ~ 1,
    offset = log(Hours),
    family = poisson(link = "log"))

#Perform backward and forward selection:
backward_sel_model <- stepAIC(
  full_interaction_model,direction = "backward",trace = 0)
forward_sel_model <- stepAIC(
  null_model,
  scope = formula(full_interaction_model),
  direction = "forward",
  trace = 0) ## trace = 0 prevents automatic output of step AIC function.
```

```
#Inspect models:
formula(backward_sel_model)
```

```
## Injuries ~ Safety + Experience + bonus + training + university
```

```
#Inspect models:
forward_sel_model$formula
```

```
## Injuries ~ Experience + Safety + training + bonus + university
```

```
#Inspect models:
AIC(backward_sel_model)
```

```
## [1] 1555.333
```

```
#Inspect models:
AIC(forward_sel_model)
```

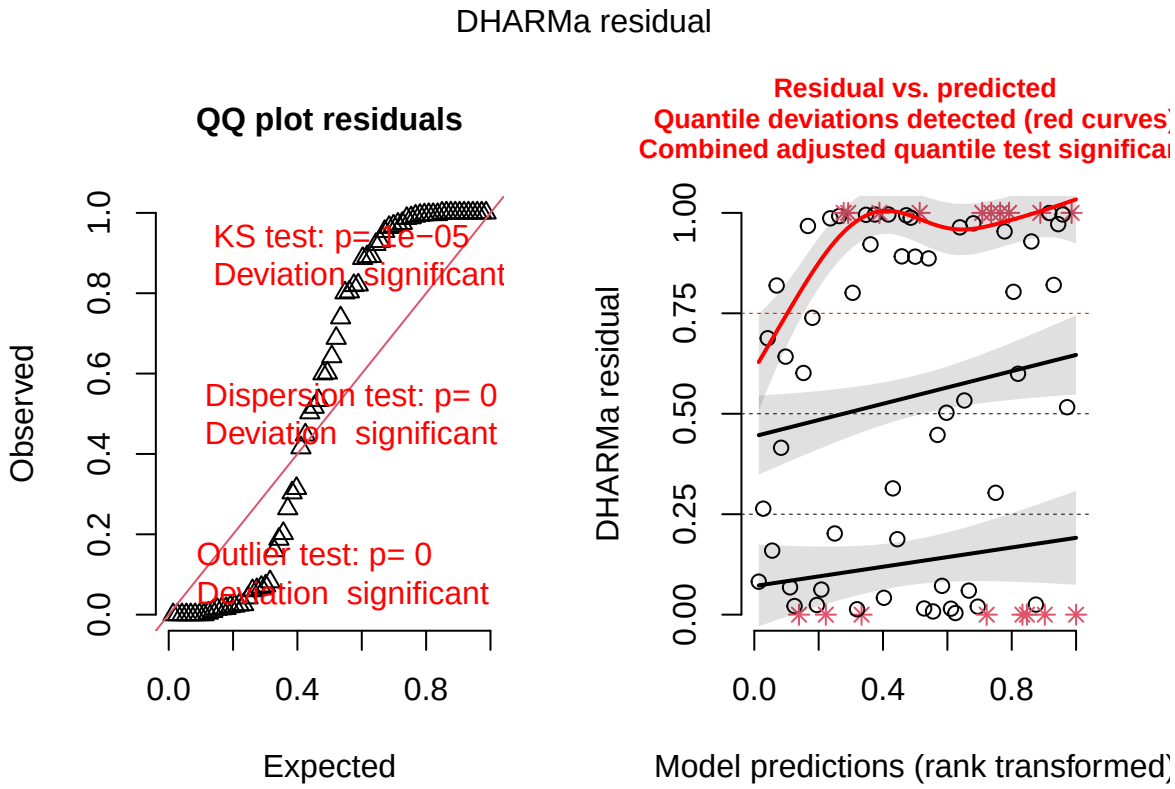
```
## [1] 1555.333
```

Backward and forward selection arrive at the exact same model, which includes all variables. We can now inspect the residual plots using the simulation method provided by the DHARMa package.

Check residuals:

```
#Simulate residuals from the model:
poisson_residuals = simulateResiduals(backward_sel_model)

#Plot observed quantile versus expected quantile to assess distribution fit, and predicted value versus
plot(poisson_residuals)
```



Based on the QQ plot of residuals, we can see significant deviation, with a curve instead of a line. All three tests (KS, Dispersion, and Outlier tests) show significant deviations. On the residuals vs. predicted plot, there is a significant deviation from the expected line (red curves). The combined adjusted quantile test is also significant. This indicates that the Poisson model is not well-fitted, and we should consider other models, such as the Negative Binomial model.

Check dispersion:

```
disp_result <- dispersiontest(backward_sel_model)
print(disp_result)
```

```
##
##  Overdispersion test
##
## data:  backward_sel_model
## z = 2.7261, p-value = 0.003204
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
```

```
## dispersion
##   13.32541
```

The p-value for the test of dispersion is highly significant (z=2.73, p-value = 0.003204), indicating that the data is more variable than expected under the Poisson GLM. Given this and that the Poisson regression model is not well-fitted (based on plots above). Let's compare it with a Negative Binomial model.

Negative-Binomial model

For the Negative Binomial model, as with the Poisson GLM model, I will use 'Injuries' as the target variable; 'safety', 'experience', 'bonus', 'training', and 'university' as covariates; and 'Hours' as an offset variable.

Stepwise selection based on AIC was applied to determine which variables should be included in the model, using both backward and forward selection methods.

```
# Full model with all possible interactions for negative binomial regression
NB_full_model <- glm.nb(
  Injuries ~ Safety + Experience + bonus + training + university + offset(I(log(Hours))),
  data = injury_data,
  link = "log"
)

# Null model with only an intercept for forward selection
NB_null_model <- glm.nb(
  Injuries ~ 1 + offset(I(log(Hours))),
  data = injury_data,
  link = "log"
)

#Perform backward and forward selection:
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",tra
```

```
#Inspect models:
formula(NB_backward_sel_model)
```

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

```
#Inspect models:
formula(NB_forward_sel_model)
```

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

```
#Inspect models:
AIC(NB_backward_sel_model)
```

```
## [1] 651.8713
```

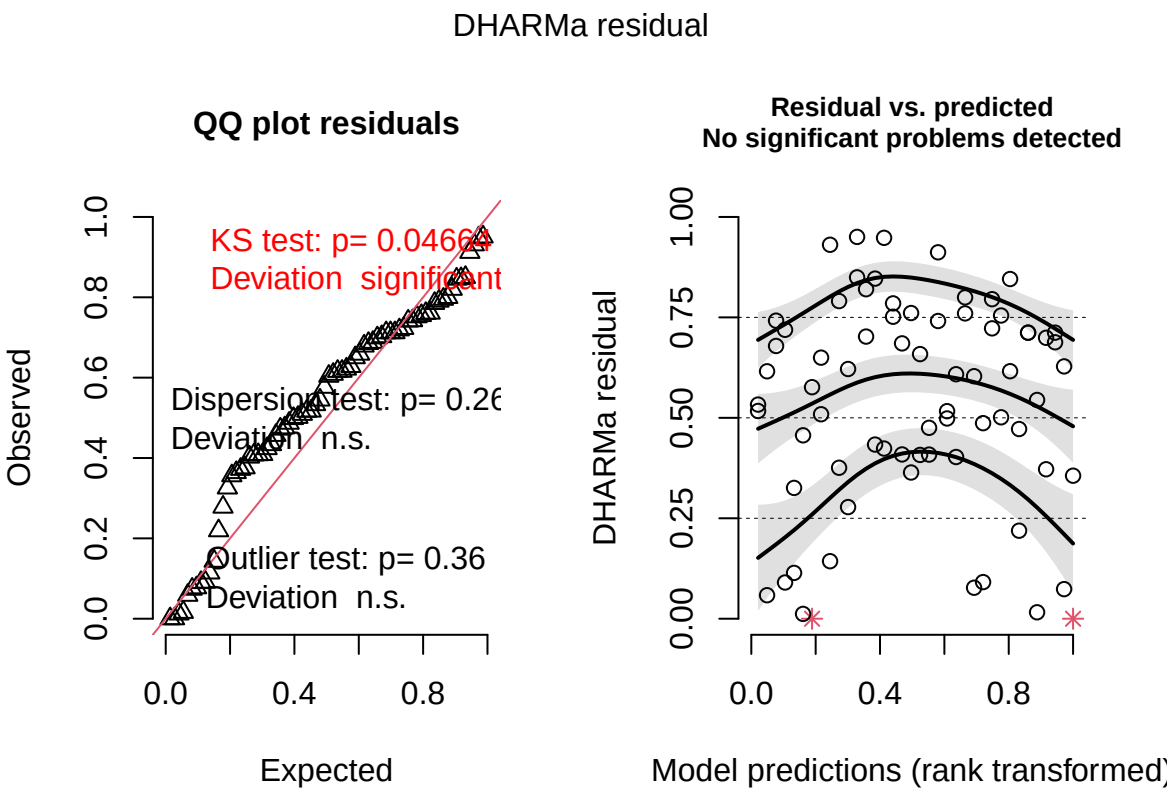
```
#Inspect models:
AIC(NB_forward_sel_model)
```

```
## [1] 651.8713
```

Backward and forward selection arrive at the exact same model, which only includes Experience and Safety as variables. We can now inspect the residual plots using the simulation method provided by the DHARMa package.

Check residuals:

```
NB_residuals = simulateResiduals(NB_backward_sel_model)
plot(NB_residuals)
```



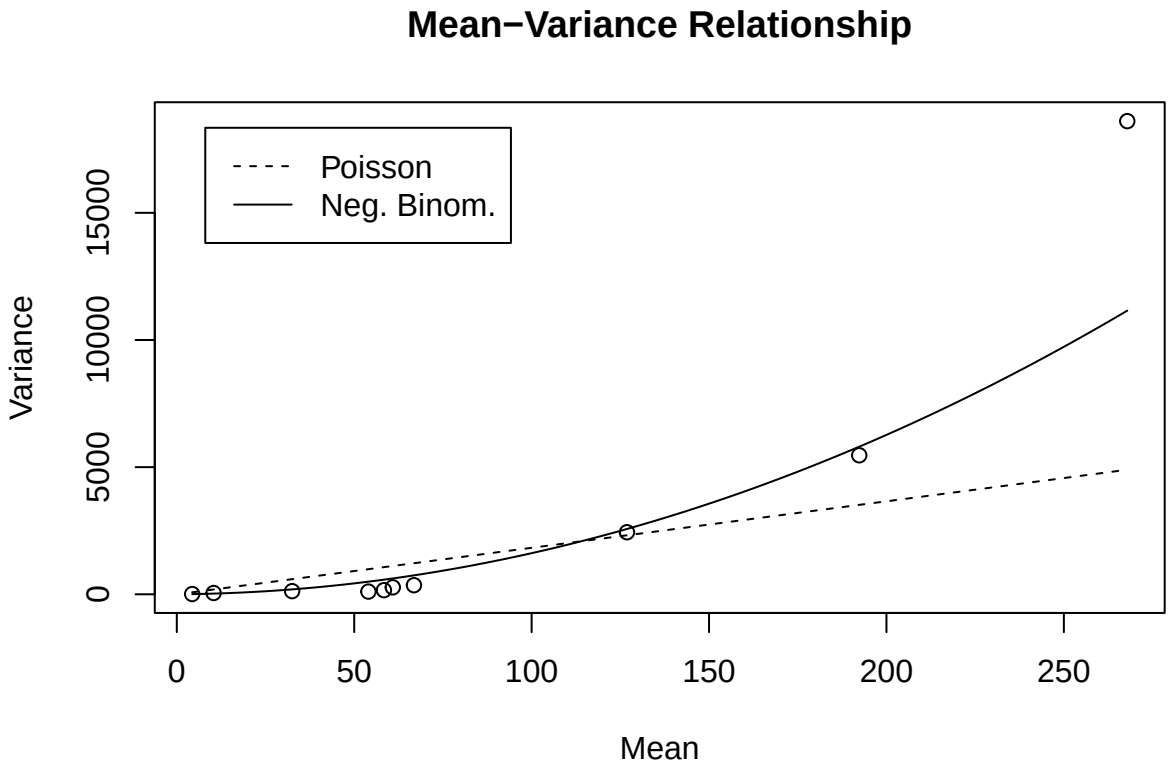
Based on the QQ plot of residuals for the Negative Binomial model, we see a much closer alignment with the expected line, indicating improved model fit compared to the Poisson model. While the KS test still shows a slight deviation ($p = 0.04664$), the Dispersion ($p = 0.264$) and Outlier ($p = 0.38$) tests are not significant, suggesting no issues with dispersion or outliers. The residuals vs. predicted plot also does not indicate any significant problems. Overall, these results indicate that the Negative Binomial model provides a better fit for the data than the Poisson model.

Support for the negative binomial model was also indicated by AIC which was 1555.333 for the Poisson regression model found and 651.8713 for the negative binomial model found (substantially lower).

Let's compare our models, exploring the mean-variance relationship for the Poisson and Negative Binomial models.

```
# Get estimate of phi_hat
res <- backward_sel_model$df.residual
phi_hat <- deviance(backward_sel_model)/res
```

```r
# Plot mean-variance relationship
xb <- predict(NB_backward_sel_model)
g <- cut(xb, breaks=quantile(xb,seq(0,100,10)/100))
m <- tapply(injury_data$Injuries, g, mean)
v <- tapply(injury_data$Injuries, g, var)
plot(m, v, xlab="Mean", ylab="Variance",
main="Mean-Variance Relationship")
x <- seq(min(m), max(m), length.out = 100)
#x <- seq(2.22,6.34,0.02)
lines(x, x*phi_hat, lty="dashed")
lines(x, x*(1+x/NB_backward_sel_model$theta)) # VAR[Y] = mu + mu^2/theta
legend("topleft", lty=c("dashed","solid"),
legend=c("Poisson","Neg. Binom."), inset=0.05)
```



This shows that the data is better explained by the negative binomial model, compared to the over-dispersed Poisson model.

Modelling results

Based on the analysis above, a negative binomial GLM was selected. Let's take a look at the summary of this model.

```r
summary(NB_backward_sel_model)
```

```
##
## Call:
## glm.nb(formula = Injuries ~ Safety + Experience + offset(I(log(Hours))),
```

```
##     data = injury_data, init.theta = 6.59200209, link = "log")
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.77729    0.10554 -83.167   <2e-16 ***
## Safety2      -0.25880    0.15530  -1.666   0.0956 .
## Safety3       0.04599    0.14198   0.324   0.7460
## Safety4       0.21189    0.14333   1.478   0.1393
## Experience.L -1.37419    0.10430 -13.176   <2e-16 ***
## Experience.Q -0.15312    0.10212  -1.499   0.1338
## Experience.C -0.12371    0.10240  -1.208   0.2270
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(6.592) family taken to be 1)
##
##     Null deviance: 245.953  on 71  degrees of freedom
## Residual deviance:  76.155  on 65  degrees of freedom
## AIC: 651.87
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  6.59
##           Std. Err.:  1.30
##
##  2 x log-likelihood:  -635.871
```

```
summary(NB_backward_sel_model)$coefficients[,4]
```

```
##  (Intercept)      Safety2      Safety3      Safety4 Experience.L Experience.Q
## 0.000000e+00 9.561909e-02 7.459857e-01 1.393169e-01 1.209544e-39 1.337568e-01
## Experience.C
## 2.270247e-01
```

Interpretation

The coefficient estimates from the negative binomial GLM show a significant monotonic decrease in the log-expected number of injuries as experience level increases from 1 to 4 (Experience.L: Estimate = -1.37419, p-value $<$ 2e-16).

Additionally, Safety regime 2 is associated with a decrease in the log-expected number of injuries compared to Safety regime 1, but this effect is marginally significant (Estimate = -0.25880, p-value = 0.096, slightly above the 0.05 threshold). Other safety regimes show an increase in the log-expected number of injuries, but these effects are not significant.

Recommendations and conclusions

Based on the current analysis of workplace injury data and the chosen negative binomial GLM, I provide answers and recommendations to the questions raised.

1. Of the various safety regimes in place across the company, which one would be recommended to become the international standard for our company, based solely on injury prevention performance?

   Based on the current data, the analysis showed that Safety Regime 2 results in fewer injuries than the other safety regimes, but this effect is marginally significant (p-value = 0.096, slightly above the 0.05 threshold). Therefore, based on the current data, it cannot be clearly highlighted and recommended as the international standard for our company. However, since this effect is close to being significant, it is recommended to accumulate or analyze a larger dataset for further analysis.

2. It has been suggested by senior management that industry experience is more important than the safety regime when it comes to preventing injuries. The idea is that a policy should be developed that is directly related to lowering employee turnover will reduce injury rates. Do the available data support this assertion?*

   Based on the analysis, we can significantly support the thesis that industry experience is more important than the safety regime in preventing injuries. This is because experience has a strong and statistically significant linear effect on reducing the number of injuries (p-value $<$ 2e-16), while the effects of the safety regimes are not statistically significant. The coefficient for experience (-1.37419) is also larger than the coefficients for Safety2, Safety3, and Safety4, indicating that experience has a greater impact on reducing the number of injuries. Thus, the data support the potential effectiveness of the idea that lowering employee turnover will reduce injury rates.

3. If there is any relationship between:

- Injuries and the annual bonuses a proportion of employees receive.
- Injuries and whether staff have received any formal external qualifications, e.g., external safety training or a university degree.

   Since the variables reflecting the presence of a bonus, safety training, and a university degree were not included in our final model, this indicates that these factors do not have an impact on reducing or increasing the number of injuries.

My conclusions are based solely on the current analysis and data. It is recommended to continue collecting more data across different time periods and locations, or to analyze additional variables that might influence injury rates (for example, location, age, occupation and others).