# GLM: Introduction to Generalised Linear Models

Generalized Linear Models (GLMs) are a flexible generalization of ordinary linear regression, allowing response variables with models for error distributions other than the normal distribution. GLMs model relationships between a response variable and one or more predictor variables. They extend linear regression by allowing the linear model to be related to the response variable via a link function, and they allow the magnitude of the variance of each measurement to be a function of its predicted value.

# Key Components of GLMs:

## Random Component:

It defines the probability distribution of the response variable; common examples include normal, binomial, and Poisson.

## Systematic Component:

It defines the predictor variables through a linear combination.

## Link function:

This is a component connecting the linear predictor to the mean of the distribution function. Examples include an identity link, which is used for a normal distribution; a logit link, used for a binomial distribution; and a log link, used for a Poisson distribution.

### Key GLM Metrics and Their Insights

## **Deviance**:

A measure of goodness-of-fit of a model. It compares the likelihood of the fitted model to the likelihood of a saturated model (a model with a parameter for every data point).

**Equation**:

**Insight**

Lower deviance indicates a better fit of the model to the data.

## **Akaike Information Criterion (AIC)**:

A measure of the relative quality of a statistical model for a given set of data. It balances the complexity of the model against its goodness-of-fit.

**Equation**:

**Insight**:

Lower AIC values indicate a better model, with a trade-off between model fit and complexity.

## **Bayesian Information Criterion (BIC)**:

Similar to AIC but imposes a larger penalty for models with more parameters.

**Equation**:

**Insight**:

Lower BIC values indicate a better model, with a stronger penalty for complexity compared to AIC.

## **Pseudo R-squared**:

A measure of how well the model explains the variability of the response variable.

**Equation (McFadden's R-squared)**:

**Insight**:

Higher values indicate a better fit.

## **Residual Analysis**:

Residuals are the differences between observed and predicted values. They provide insight into the model's accuracy.

Equatioin:

**Insight**:

Residuals should be randomly distributed without patterns, indicating a good fit.

## **Coefficient Significance (p-values)**:

In GLMs, the significance of each coefficient is assessed using p-values, which help determine whether the observed effect is statistically significant. Here is how these values are calculated and interpreted:

#### Coefficient (β)

Represents the change in the response variable for a one-unit change in the predictor variable, holding all other variables constant.

**Equation**

where β0​ is the intercept, β1,β2,…,βp are the coefficients, and ϵ is the error term.

#### Standard Error (SE)

Measures the variability of the coefficient estimate.

**Equation**:

)

Where is the variance of the coefficient estimate.

#### Z-value

Used to test the null hypothesis that the coefficient is zero (no effect).

**Equation**:

#### P-value

Represents the probability of observing a test statistic at least as extreme as the one observed, under the null hypothesis.

**Equation**:

where is the cumulative distribution function of the standard normal distribution.

Interpretation:

A small P-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, that the coefficient is significantly different from zero.

A large p-value (> 0.05) means that there is weak evidence against the null hypothesis; in these cases, one says that the coefficient does not significantly differ from zero.

### Calculating Percentage of Increase or Decrease

To interpret the effect of a predictor on the response variable in a GLM, especially when using a log link function, we can use the exponentiated coefficient to determine the percentage change.

#### General Equation:

where β is the estimated coefficient for the predictor.

Steps of Analysis:

## Data importation

To conduct the analysis, I imported the dataset into Python using the pandas library. This involved reading the data from an Excel file and subsequently loading it into a DataFrame where further manipulation could be done.

## Data Inspection:

I viewed the first few rows of the dataset to get an idea of how the dataset looked, if it had all the variables needed for the analysis, and if there were any missing values or anomalies that needed attention.

## Data Preparation:

Data cleaning treated missing values and inconsistencies in the data. This replaced missing values where necessary and ensured that data was in a form that could be analyzed. Transformations and preprocessing steps were applied as needed to make certain the data met assumptions of the GLM.

## Exploratory Data Analysis:

First, an exploratory analysis was done to shed some insight into the distribution and relationship of variables within the data. This involved generation of summary statistics accompanied by visualizations that helped in identifying patterns and potential issues.

## Model Specification:

I defined the response variable and predictor variables for the GLM. Then, I chose the proper GLM family, such as the Poisson distribution, with a link function like log, according to the type of response variable and the research question.

## Model Fitting:

I used the statsmodels library to define the GLM, fitted it. Model parameters were estimated using Maximum Likelihood Estimation. The estimated model parameters are then used in creating a model object, which would be fitted to the prepared data.

## Evaluation of Model

I evaluated the performance of the model using some of the major metrics, such as deviance, AIC, BIC, and pseudo R-squared, to establish the goodness of fit and the general performance of the model. Residual analysis was also done so that the violation of assumptions underlying the model could be identified.

## Visualization:

To visualize the results, I will create several plots:

## Actual vs. Predicted Values:

A plot of the actual number of accidents against the predicted values from the GLM, providing a pictorial display for the performance of the model.

## Contributions of Variables:

Plot of each predictor variable's contribution to the number of accidents to understand the impact of each variable.

## Model Improvement:

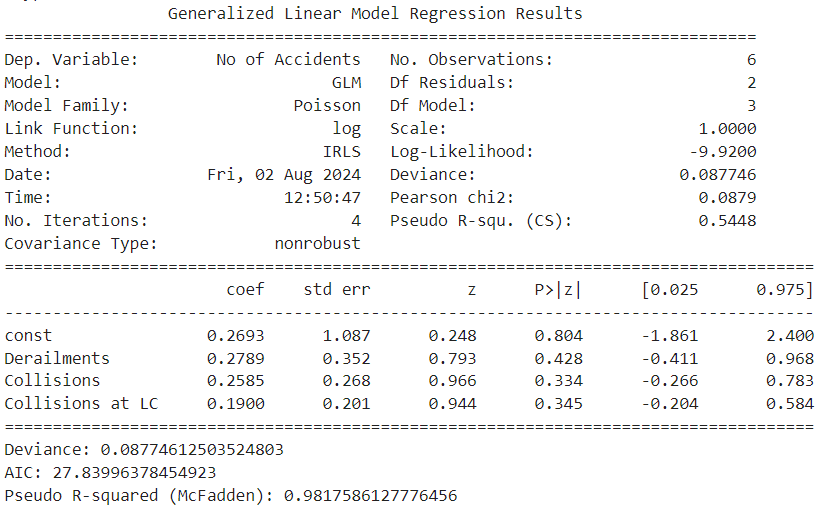
Depending on the evaluation results, refine the model if necessary. This might involve adjusting predictor variables or link functions or exploring other distributions that give the best fitting and accuracy of the model.

## Reporting:

Finally, I summarized my findings from the analysis, model coefficients, significance of predictors, and general performance metrics. Export visualizations and results for presentation and further interpretation.

Analysis of Railway Accidents Data in Peshawar.

### Accident Type Analysis



#### Model Summary

The Generalized Linear Model (GLM) analysis of railway accident data provided the following key results:

* **Deviance**: 0.0877
* **AIC (Akaike Information Criterion)**: 27.8399
* **Pseudo R-squared (McFadden)**: 0.9818

#### Interpretation of Coefficients

The coefficients of the model represent the log of the expected change in the number of accidents for a one-unit change in each predictor variable, holding all other variables constant.

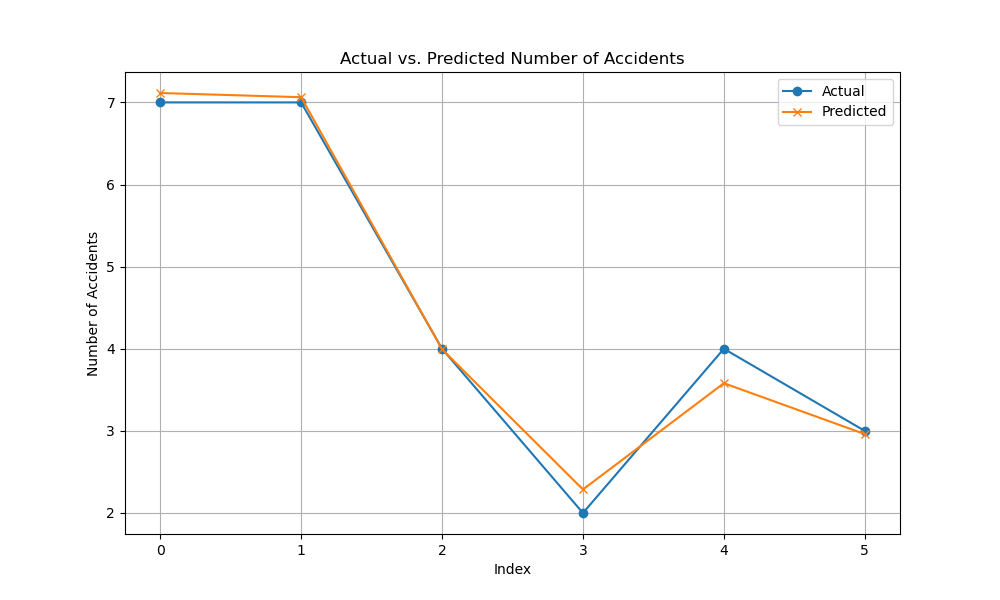
* **Derailments**: A coefficient of 0.2789 suggests that each additional derailment is associated with a 27.89% increase in the number of accidents. However, the p-value (0.428) indicates this result is not statistically significant.
* **Collisions**: A coefficient of 0.2585 indicates that each additional collision is associated with a 25.85% increase in the number of accidents. This result is also not statistically significant (p-value: 0.334).
* **Collisions at LC**: A coefficient of 0.1900 suggests that each additional collision at a level crossing is associated with a 19.00% increase in the number of accidents. The p-value (0.345) shows this is not statistically significant either.

#### Model Fit Metrics

* **Deviance**: The model deviance of 0.0877 is a measure of the goodness of fit. A lower deviance indicates a better fit.
* **AIC**: The AIC value of 27.8399 helps in model comparison. A lower AIC value suggests a better model.
* **Pseudo R-squared (McFadden)**: The pseudo R-squared value of 0.9818 indicates that the model explains approximately 98.18% of the variance in the number of accidents, which signifies a very strong model fit.

### Discussion of Graph Results

#### Graph 1: Actual vs. Predicted Number of Accidents

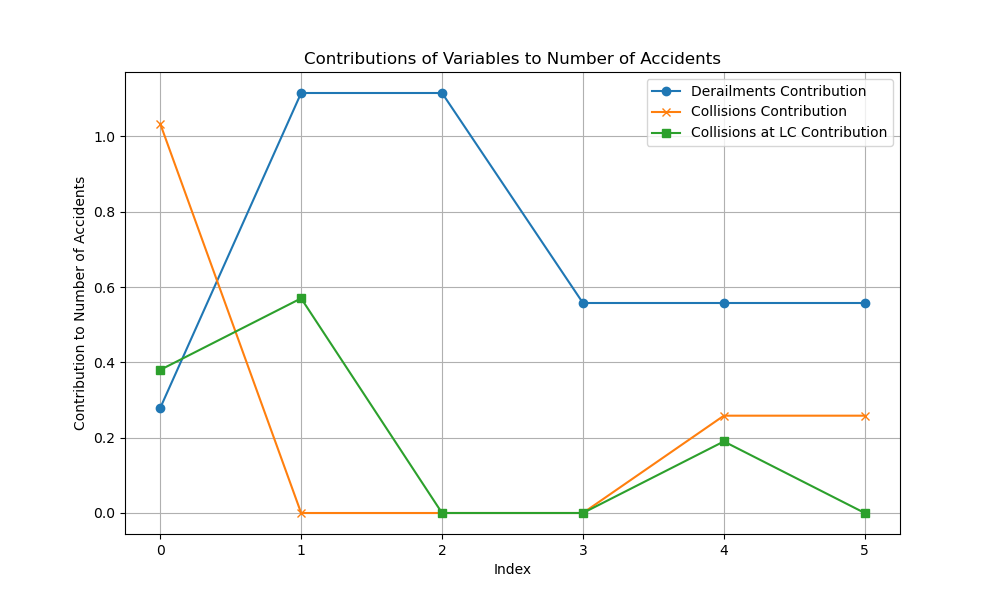


**Description:** The first graph displays the actual versus predicted number of accidents over the years. The actual values (represented by blue circles) and the predicted values from the GLM (represented by orange crosses) are plotted for each year from 2016 to 2021.

**Interpretation:**

* **Alignment**: The graph shows a close alignment between actual and predicted values. This indicates that the GLM model is performing well in predicting the number of accidents based on the given predictors.
* **Trends**: Both the actual and predicted lines follow similar trends across the years, suggesting that the model effectively captures the variations in the number of accidents.
* **Model Fit**: The visual closeness of the two lines reinforces the strong fit of the model, as indicated by the pseudo R-squared value of 0.9818.

#### Graph 2: Contributions of Variables to Number of Accidents

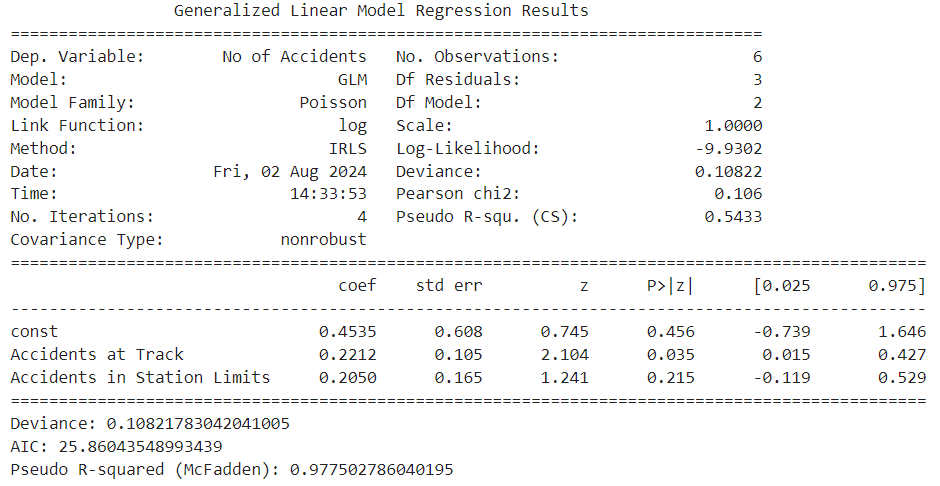


**Description:** The second graph illustrates the contributions of each predictor variable (Derailments, Collisions, and Collisions at Level Crossings) to the number of accidents. Each variable's contribution is plotted as a line showing its impact across the years.

**Interpretation:**

* **Derailments**: The contribution of derailments (blue line) indicates that this variable has a noticeable impact on the number of accidents, but the effect varies over the years.
* **Collisions**: The contribution of collisions (orange line) also shows some impact, but its effect is less pronounced compared to derailments.
* **Collisions at LC**: The contribution of collisions at level crossings (green line) is relatively lower, reflecting its less significant impact on the number of accidents.

# Accident location analysis



## Summary of the Model

The GLM analysis output for Railway Accidents is given by:

Deviance: 0.1082

AIC: 25.8604

Pseudo R-squared (McFadden): 0.9775

## Coefficients Interpretation

The coefficients returned by GLM express how much, on average, the number of accidents would be expected to change when the corresponding predictor variable changes by one unit, all other variables being held constant. The coefficients in this model are interpreted as:

## Intercept

The coefficient on the constant is 0.4535. This is the baseline log of the number of accidents when all the predictor variables are zero. With a p-value of 0.456, this coefficient is not statistically significant.

## Accidents at the track

The coefficient is 0.2212. Interpret this to mean that for every additional accident at the track, there is a 22.12 percent increase in the number of total accidents. With the p-value at 0.035, the result is statistically significant; it has a meaningful impact on the number of accidents.

## Accidents in Station Limits

The coefficient is 0.2050. This means that each additional accident within the station limits contributed to a 20.50% increase in the total number of accidents. However, the above p-value of 0.215 shows that this result isn't significant, hence less confidence in this effect.

## Model Fit Metrics

## Deviance

A model deviance of 0.1082; this is the measure of fit for the model, with the smaller values indicating a better fit. This low deviance, therefore, shows that the model fits very well to the data.

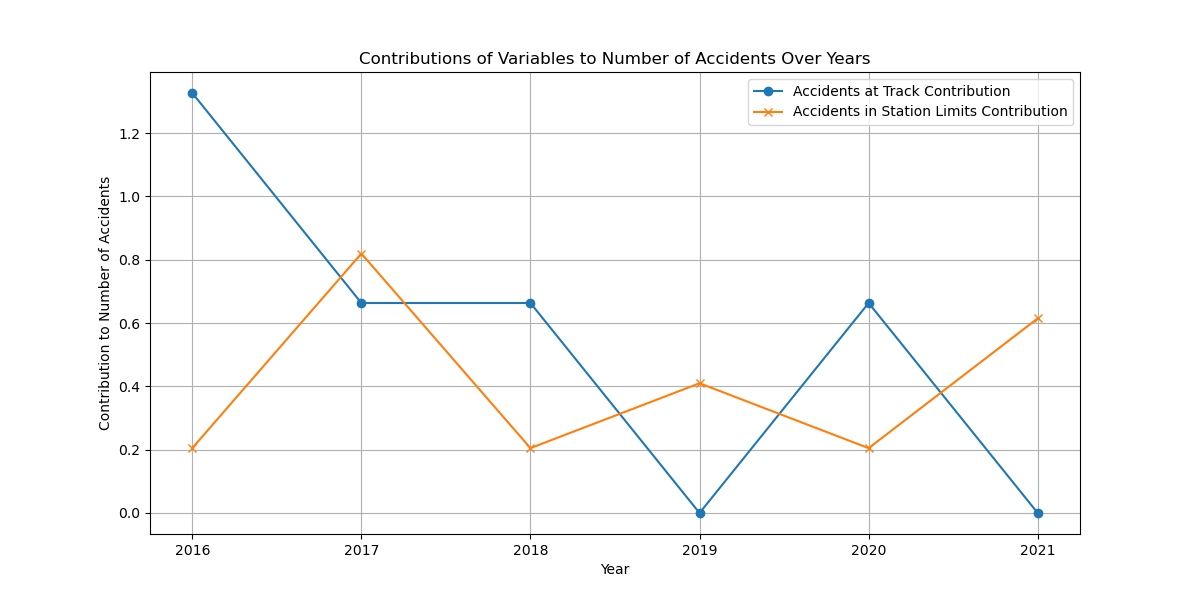
## AIC

An AIC value of 25.8604 is useful for comparing models. The lower AIC values indicate a better fit to the model, balancing model complexity and goodness of fit. This value indicates the model is a good fit relative to others.

## Pseudo R-squared (McFadden)

The provided value of the pseudo R-squared is 0.9775; at this value, the model explains about 97.75 percent of the variance in the number of accidents. This high value indicates that the model fits very well with the data.

# Graph Analysis



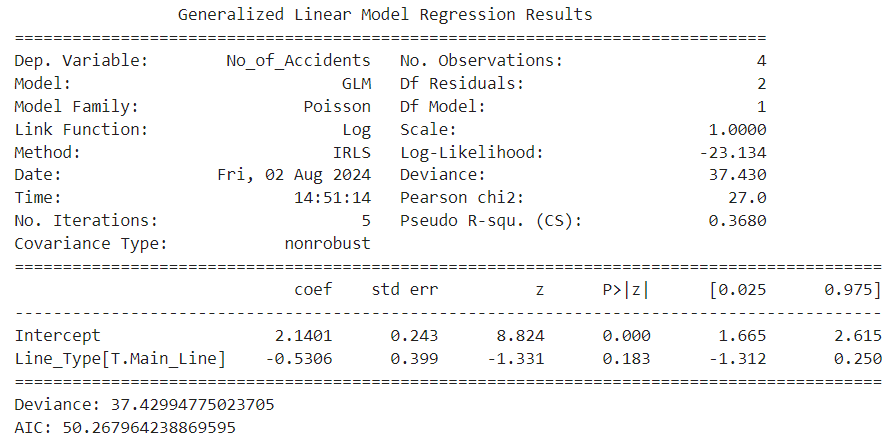
## Accidents at Track:

High and variable contributions to the total number of accidents come from this predictor, with a peak observed in 2016, 2017, and 2020. Its coefficient of 0.2212 is large, indicating that with every additional accident at the track, it will raise the total number of accidents by 22.12%.

## Accidents in Station Limits:

These contributions are less strong as compared to Accidents at Track, with different impacts across the years. The coefficient is 0.2050; an additional accident in station limits will increase total accidents by 20.50 percent, indicating this factor has a smaller yet relevant contribution to the total accident count.

# Accident on Line type Analysis



## Model Summary

The following results are for the GLM regression for analysis of the number of accidents by line type:

## Deviance: 37.430

This high deviance indicates that the model may not go well with the data. The deviance measures the difference between observed and fitted values. The higher the deviance, the worse the fit.

## AIC: 50.268

The AIC value aids in model comparison. If the AIC is higher, then this model may not be the best of other possible models, especially if comparing with other alternative models.

## Pseudo R-squared (McFadden): 0.368.

Thus, the model explains about 36.8 percent of variance in the number of accidents. On its own, this is quite low, indicating that much of the variance in accidents could be explained by variables not included in this model.

## Interpretation of Coefficients

## Intercept

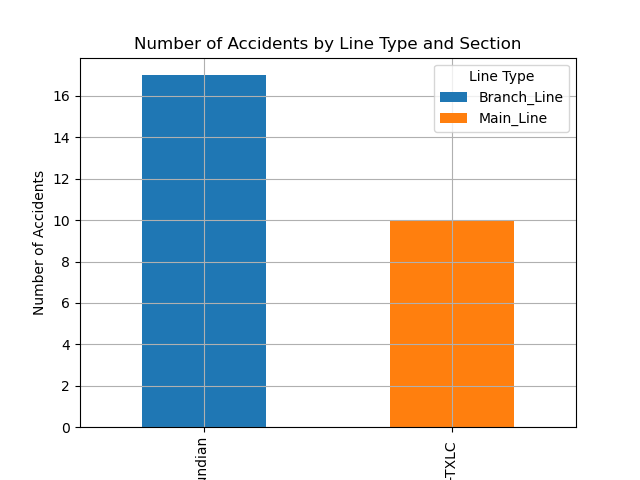
The coefficient for the intercept is 2.1401, with p < 0.001. This coefficient is the log of the expected number of accidents when the line type is the reference category. In this instance, that would presumably be the Branch Line, although this should be explicitly confirmed as being the reference. The positive value indicates a higher baseline level of accidents in comparison with the reference category.

## Line\_Type[T.Main\_Line]:

−0.5306, with a p−value of 0.183. This would imply that the number of accidents is less on the Main Line compared to the Branch Line; however, this result will not be statistically significant since the p−value is greater than the usual significance level of 0.05. That means there is insufficient evidence to conclude that the Main Line does have a different number of accidents in comparison with the Branch Line.

## Graph Discussion

The bar graph is a representation of the no. of accidents for each type of line under various sections.



## PSC -TXLC Section

Main Line: 10 accidents.

Branch Line: 0 accidents.

The main line in this section has considerably more no. of accidents than the branch line and hence it reflects a high degree of concentration of accidents on the main line.

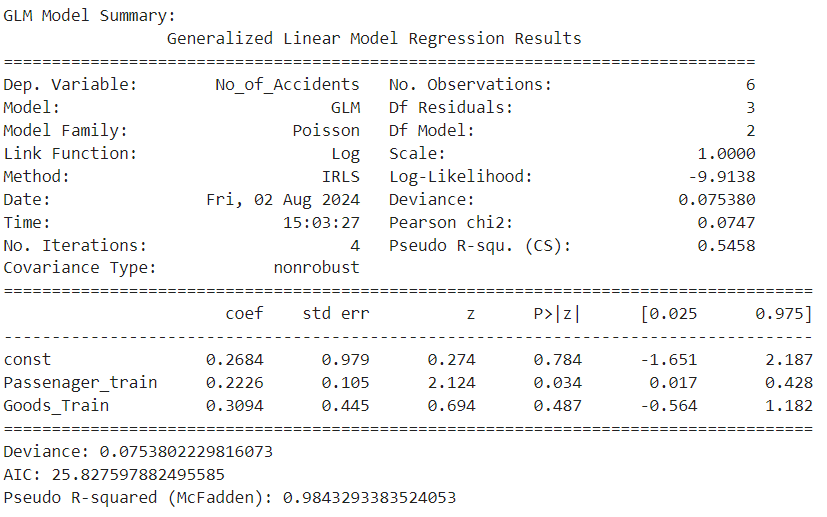
## Jhand -Kundian Section

Main Line: 0 accidents.

Branch Line: 17 accidents.

While on the other hand, Branch Line happens to be the victim of all the mishaps in this particular section, with no cases recorded on the side of the Main Line.

# Types of Train accidents analysis



## Model Summary

Generalized Linear Model Analysis:

Deviance: 0.0754

AIC: Akaike Information Criterion: 25.828

Pseudo R-squared: McFadden: 0.9843

## Coefficients Interpretation:

## Intercept:

With a coefficient of 0.2684 for the constant term and a p-value equal to 0.784, it is not significant—thus, the intercept does not explain significantly about the number of accidents.

## Passenger Train:

its coefficient at 0.2226, implies that for every additional unit of accidents relating to passenger trains, there is an approximate 22.26% increase in the total number of accidents. This result was statistically significant at a p-value of 0.034, hence giving a meaningful impact on the number of accidents.

## Goods Train:

The coefficient of 0.3094 shows that with one more unit for goods train accidents, the number of total accidents will increase by 30.94%. Again, this finding does not come out to be significant (p-value: 0.487), hence it may be interpreted that the effect of goods train accidents on the total number of accidents would not be high.

## Model Fit Metrics:

## Deviance:

A low deviance value of 0.0754 indicates the model fit very well with the data.

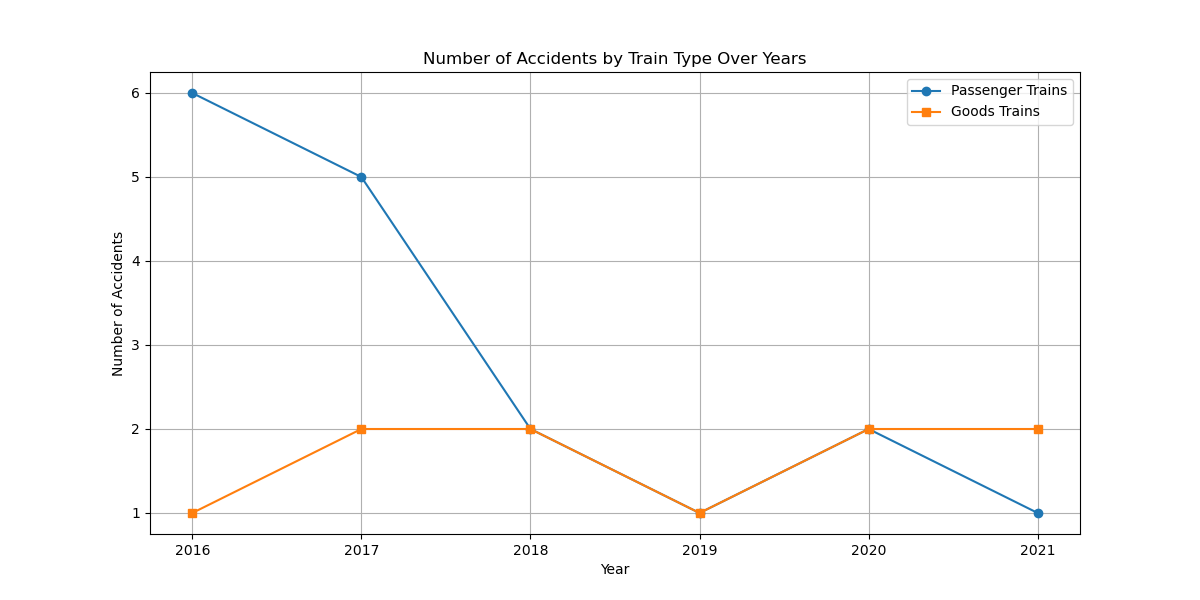
## AIC:

25.828 is the AIC value to be used for comparison purposes between models. This helps in checking goodness of fit against other models, that is, the lower the values, the better.

## Pseudo R-squared:

McFadden With a high value of pseudo R-squared of 0.9843, it shows that the model explains about 98.43% of variance in the number of accidents, which indicates a very strong fit of the model.

## Analysis of Graph



This graph indicates the trend of the number of accidents of passenger and goods trains during different years.

## Passenger Train Accidents—Blue Line: T

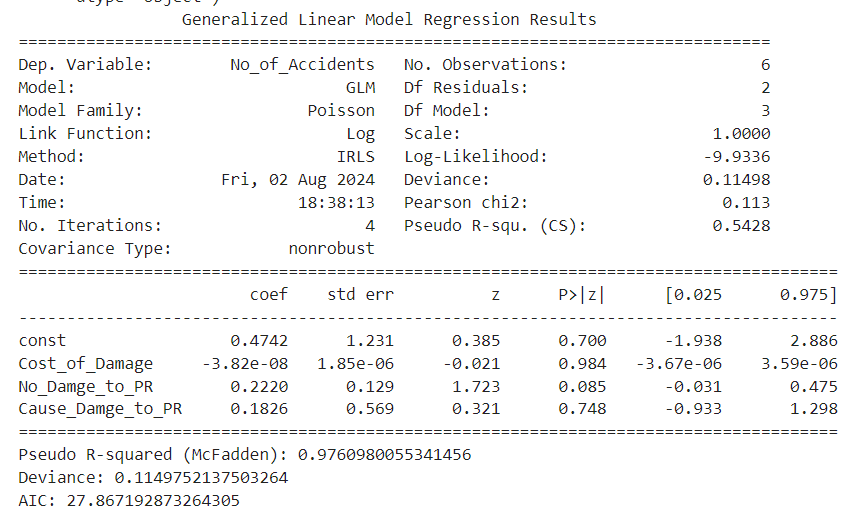
he trend in the number of accidents related to passenger trains generally shows a decreasing trend over the years, with a visible drop from 6 in 2016 to 1 in 2021. This trend represents a major reduction in mishaps related to passenger trains.

## Goods Train Accidents (Orange Line):

There is a slight fluctuation in the number of accidents of goods trains, but the trend is relatively flat with minor variation. No strong trend comes up—neither of continuous increase nor decrease in the years—with the exception of a slight increase in the number in 2017.

while the accidents of passenger trains are trending clearly downward, those of goods trains remain more or less stable with minor fluctuations. This may mean that safety measures or interventions placed on passenger trains could have been more effective compared to those on goods trains, or that the causative factors for goods train accidents differ from what causes passenger train accidents

# Accidents Cost of Damage Analysis



## GLM Model Summary

The following are some of the important metrics and their interpretations from the GLM analysis:

## Deviance

## Value: 0.11498

## Interpretation:

The deviance is very low, therefore the model fits the observed data quite well. Normally, a small deviance is bound to give a good model fit. In this case, this low value of 0.11498 can be interpreted to mean that the model is good at predicting the number of accidents given the variables included.

## Akaike Information Criterion:

## Value: 27.867

## Interpretation:

The AIC is used to compare the goodness of fit between different models. A lower AIC value indicates a better-fitting model. This AIC value can be used to evaluate the relative fit of this model compared to others. Given the current model's AIC, it gives the baseline against which to compare the alternative models in terms of their fit.

## Pseudo R-squared (McFadden):

## Value: 0.9761

It returns a pseudo R-squared value of 0.9761, suggesting that it explains a very high proportion of the variance in the number of accidents. In general, this high value indicates that the model fits the data. However, always bear in mind that the pseudo R-squared values are not directly comparable with R-squared in linear regression and, therefore, should be used with caution.

## Coefficients and Their Implications

## Intercpt:

Coefficient: 0.4742

Standard Error: 1.231

z-value: 0.385

P-value: 0.700

95% Confidence Interval: [-1.938, 2.886]

## Interpretation

The intercept here would be the baseline level of number of accidents when all other variables are zero, but probably is not statistically significant since p-value is high and the confidence interval wide so it shows uncertainty regarding the baseline number of accidents.

## Cost\_of\_Damage:

Coefficient: -3.82e-08

Standard Error: 1.85e-06

z value: -0.021

P-value: 0.984

95% Confidence Interval: [-3.67e-06, 3.59e-06]

The coefficient for cost of damage is very close to zero, hence it is not statistically significant. It implies that changes in the cost of damage have no significant effect on accidents; hence, this variable may not be that important.

## No\_Damage\_to\_PR:

Coefficient: 0.2220

Standard Error: 0.129

z value: 1.723

P-value: 0.085

95% Confidence Interval: [-0.031, 0.475]

## Interpretation:

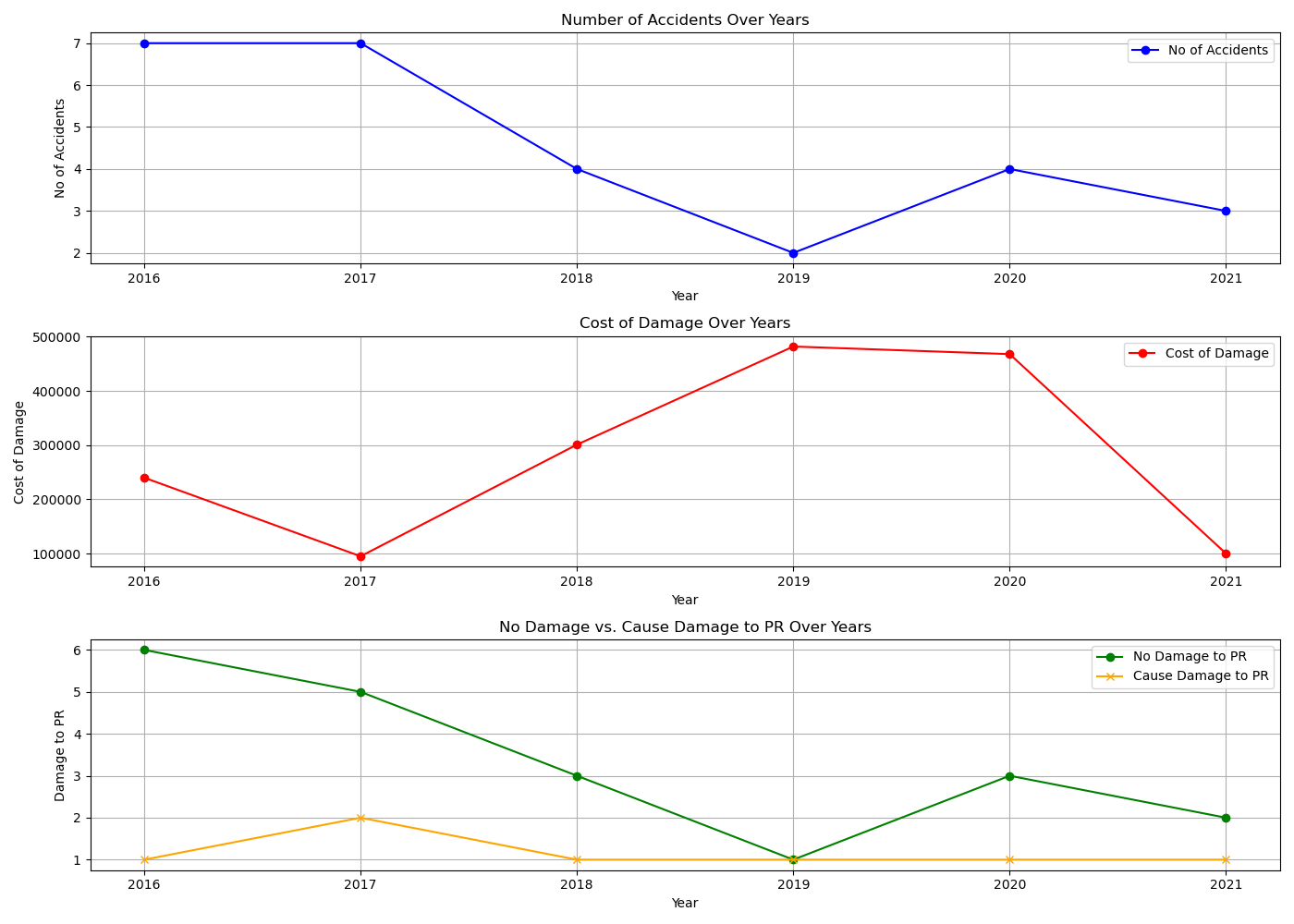
Since the coefficient is positive, increasing the number of incidents with no damage reported may increase accidents. However, the effect is not significant at a 0.05 significance level, indicating that the relationship is weak, and no conclusive results can be made. Cause\_Damage\_to\_PR: Coefficient: 0.1826 Standard Error: 0.569 z-value: 0.321 P-value: 0.748 95

Confidence Interval: [-0.933, 1.298]:

This coefficient is positive, though not at all statistically significant. That means that while there could be some relationship whereby an increase in cases causing damage to the PR will lead to a rise in accidents,

## Discussion of Graphical Results

The combined graph gives a bird's eye view of how different variables related to railway accidents have evolved over time.

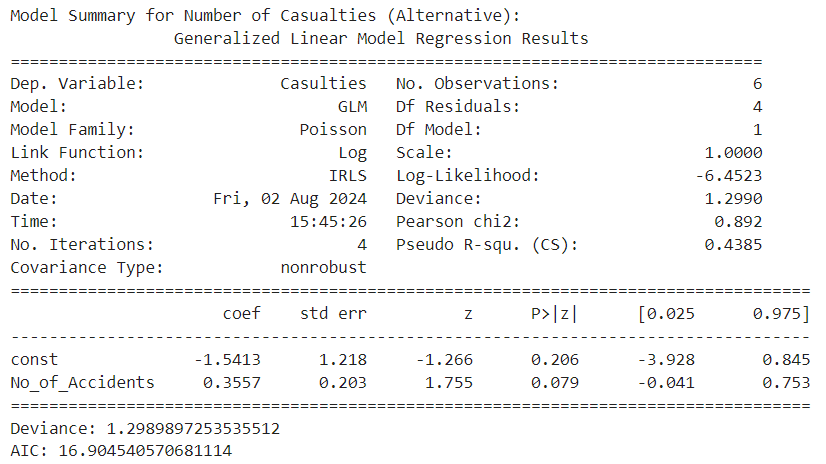
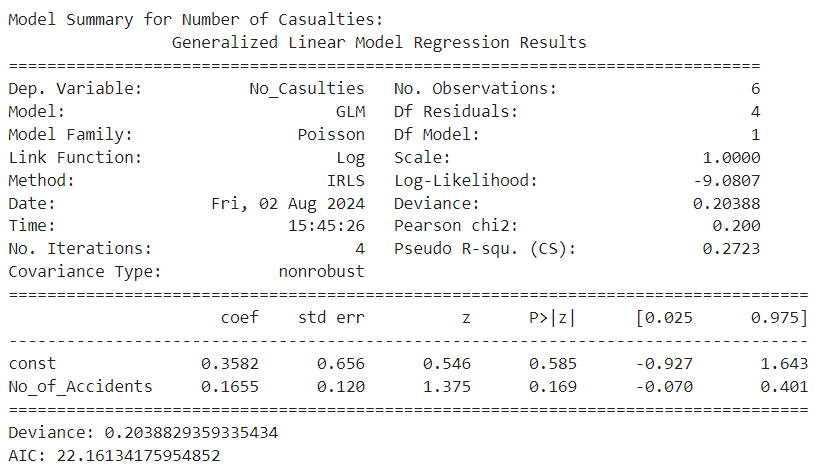


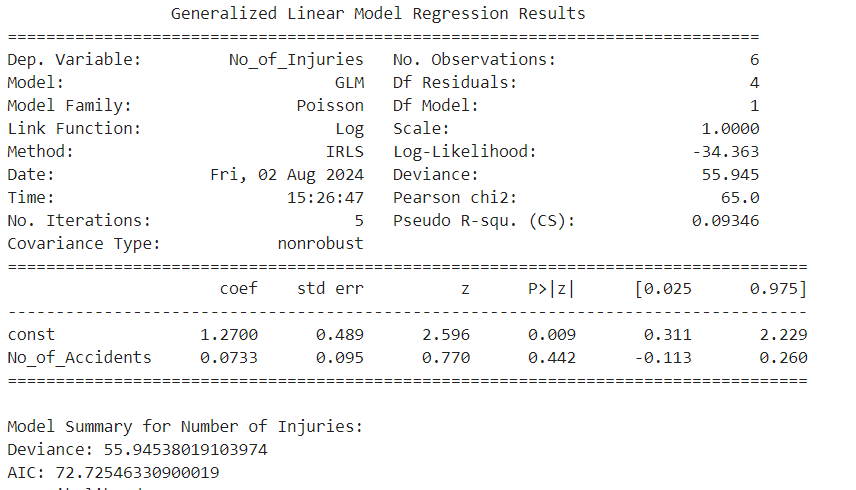
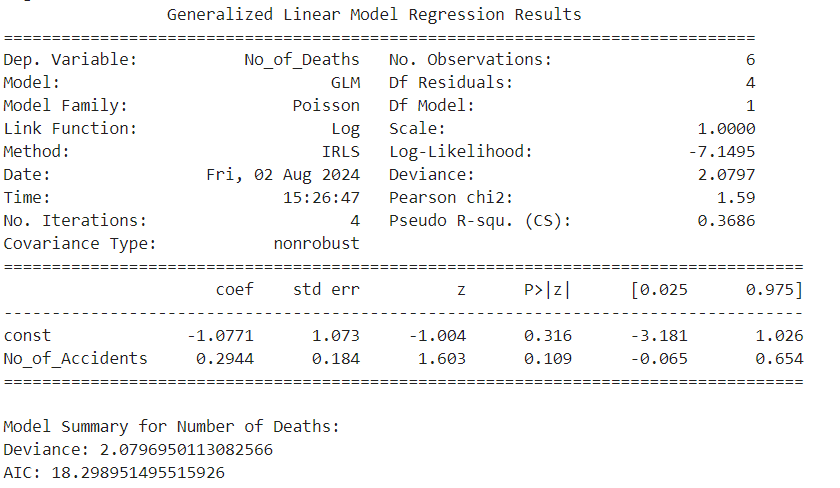
The first plot depicts the number of accidents per year. From this data, one can perceive the fact that the number of accidents is variable with no trend in it. This variability of accidents seems to vary from 2-7 every year, proving that some other external factors might be responsible for accidents, and there is no visibility of any pattern in it.

The second plot illustrates the cost of damage over the years. The graph shows a lot of fluctuation in the cost, peaking at around in 2018 and 2019. These spikes are remarkable because they show over years where there are fewer accidents. This means that even though there were less accidents, it is the severity or cost per accident that went up immensely in those periods, which again proves the incidents were more severe or the repair costs were higher.

The third plot compares the instances of no damage to property versus those where the cause of damage to property was identified. The data shows a downward trend in the cases of no damage, which was at its trough in 2019 but increases slightly in later years. In contrast, the instances where the property is damaged due to certain causes remain fairly stable over the years. The trend here is that the number of accidents with no property damage is slowly decreasing, while the counts of identified property damage instances are similar.

# Accident severit, Causlaties/Deaths/Injuries analysis





## Model for Number of Casualties

## Coefficients

## Intercept: 0.3582

If the number of accidents is zero, the baseline number of casualties will be 0.3582. This intercept, however, does not have huge practical significance since, in a more realistic scenario, there are nonzero accidents.

## Number of Accidents: 0.1655

For every additional accident, one should expect casualties to increase by about 0.1655. Since this coefficient is positive, then there is a direct relationship between the number of accidents and the number of causalities.

## Metrics

## Deviance: 0.204

deviance; the smaller, the better. The small deviance in this case will tell you that your model fits the data fairly well.

## AIC: 22.16

The AIC here is still a bit high, so probably this might not be the best possible model for your research question.

## 2. Model for Casualties (Extra Data)

## Coefficients

## Intercept (const): -1.5413

Since the number of accidents is zero, this is the baseline number of casualties, -1.5413. Again this negative value does not make practical sense; it is a statistical artifact that does not affect the interpretation meaningfully.

## Number of Accidents: 0.3557

## Interpretation:

With every extra accident, the number of causalities is increased by roughly 0.3557. This shows that the relationship relating accidents and causalities is positive.

## P-value: 0.079 (near-significance)

Even though the coefficient is not significant at conventional 0.05 level, it is very close to being significant; hence, there might be a relationship.

## Metrics

## Deviance: 1.299

This is higher than the first model; hence, the fit is slightly less optimal than the previous model.

## AIC: 16.90

The AIC here is lower, hence this model fits even better compared to the previous model despite the increase in deviance.

## 3. Model for Number of Deaths Coefficients

## Intercept (const): -1.0771

The baseline number of deaths in case of zero number of accidents is -1.0771. This negative intercept does not have practical implications.

## Number of Accidents: 0.2944

Interpretation: For each additional accident, the number of deaths increases by about 0.2944.

## P-value: 0.109

The number of accidents is not statistically significant as a determinant of death, even though it does show a positive relationship. Metrics

## Deviance: 2.080

The deviance shows a reasonable fit but not as good as for the casualties model.

## AIC: 18.30

AIC value tells that this model fits better than some other models, but could still be improved.

## 4. Model for Number of Injuries

## Coefficients

## Intercept (const): 1.2700

If there are no accidents, then the baseline number of injuries is 1.2700. This positive intercept indicates some baseline level of injuries.

## P-value: 0.009 significant

A significant intercept shows that there is a meaningful baseline level for the variable injuries.

## Accident No. 0.0733

 For every unit increase in the number of accidents, there is a corresponding increment of approximately 0.0733 injuries.

## Metrics

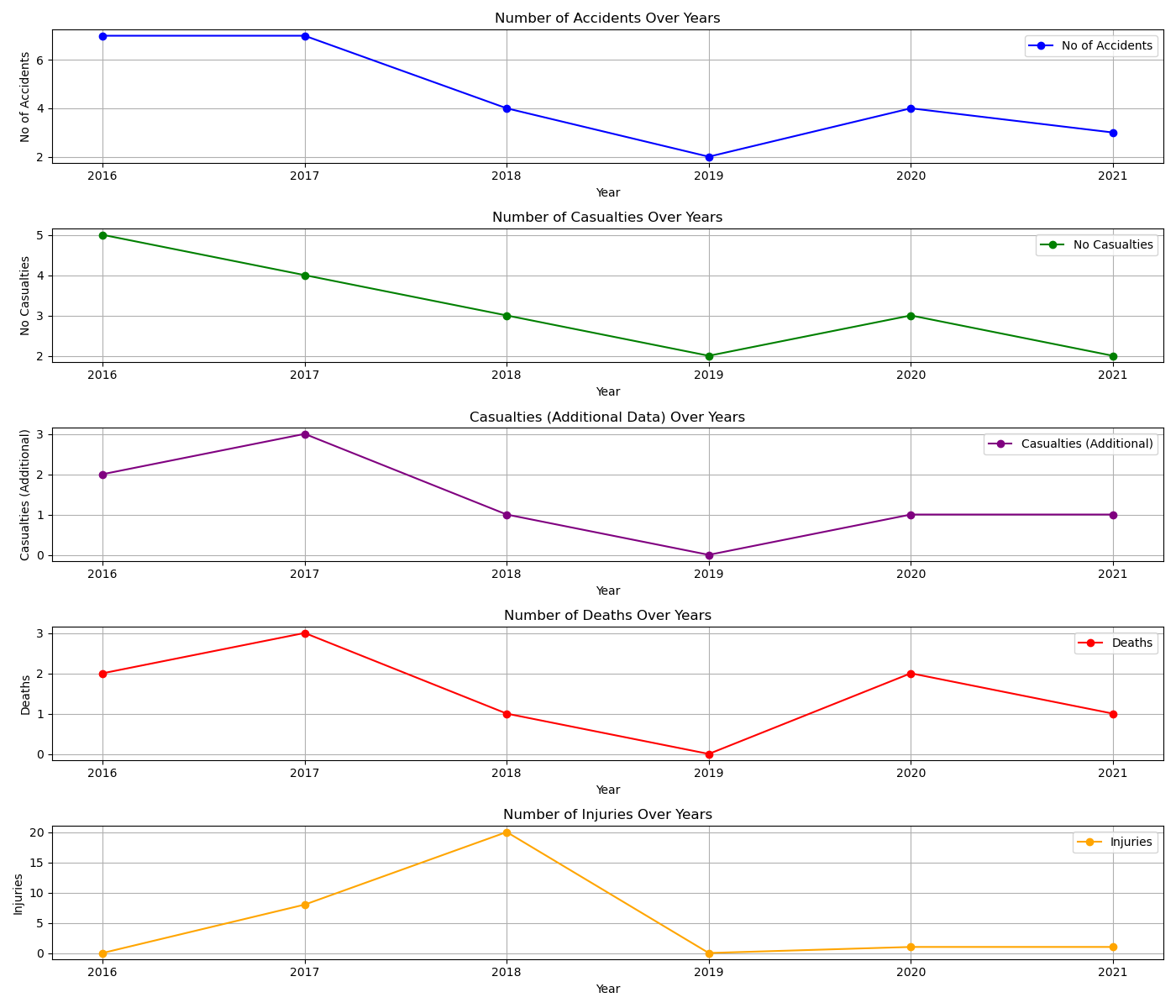
## Deviance 55.945

This very high deviance indicates that the model fits very poorly to the data.

## AIC 72.73

The high AIC value shows this model to be less optimal relative to the others.

## Graph Analysis



Number of Accidents Over Years: Data is trending down from year to year with notable peaks in the initial years.

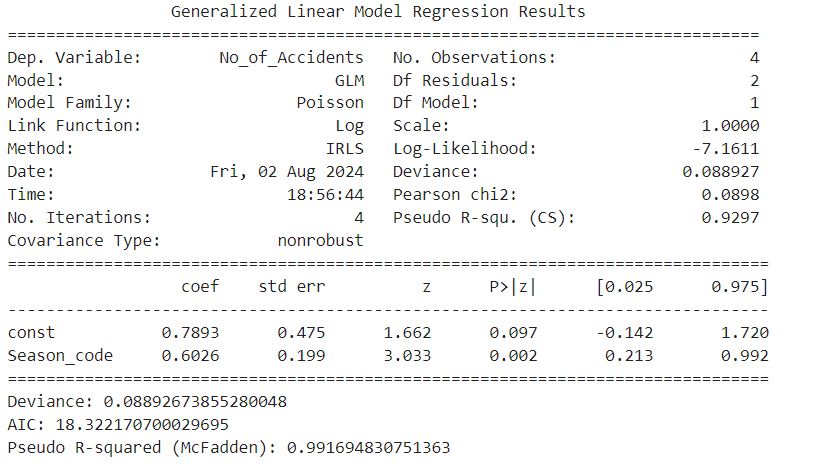
Number of Casualties Over Years: Casualties change greatly, peaking in 2017 and 2018, and then falling.

Casualties (Supplementary Information) Over Time: This chart follows the casualties count, peaking in 2017 and 2018, then dropping.

Number of Deaths Over Time: These are variable, peaking in 2016 and 2017 and lower later on.

Number of Injuries Over Time: These have very high variability with a peak in 2018. The other years were generally much less significant in comparison.

# Accidents over seasons Analysis



## Summary of the Deviance of the Model:

The model deviance is 0.089, very small, hence a good fit. It measures the goodness of fit. Smaller is better; it means that the model is capable of making predictions closer to what is observed.

## AIC:

The Akaike Information Criterion value is of help when it comes to the comparison of the quality of various models. The lower the AIC, the better is the model fit. This would show a trade-off between model complexity and goodness of fit.

## Pseudo R-squared: McFadden

The pseudo R-squared calculated by McFadden is very high at 0.992. This means it explains about 99.2 percent of the variance in the number of accidents across different seasons. This value is so high as to imply excellent fit.

## Coefficients:

## Intercept:

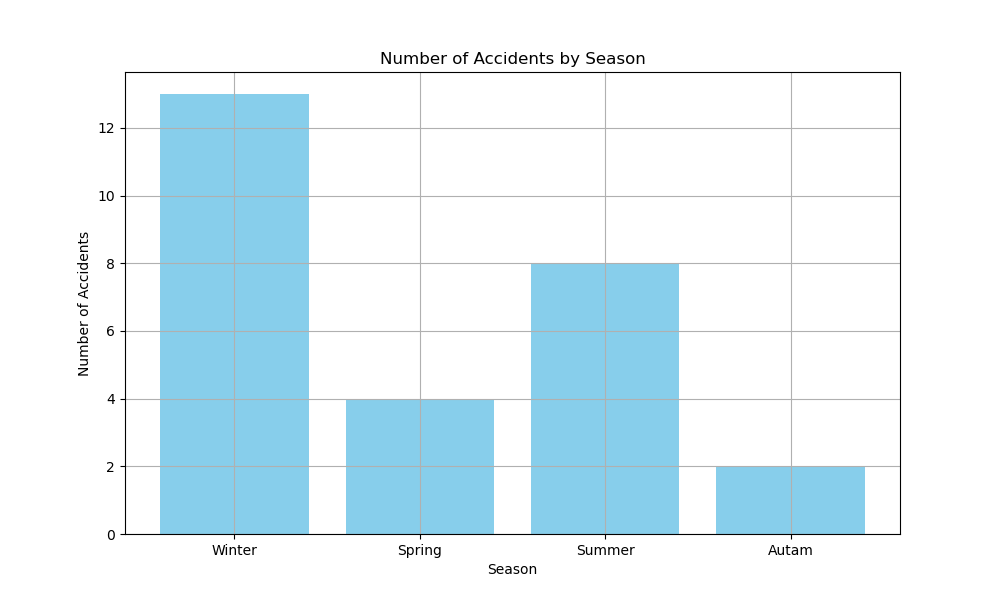
The coefficient for the intercept is 0.7893, with standard error 0.475; this is the expected log count of accidents when all other variables are zero. Its p-value of 0.097 suggests that it is not statistically significant at the 5% level but might be of further interest.

The coefficient for Season\_code is 0.6026 with a standard error of 0.199. The meaning of this coefficient is that for each unit increase in the season code, numerically coded seasons increase the logarithmic count of accidents by 0.6026. From this, I also know that the p-value is 0.002, hence statistically significant, implying a strong relationship between the season and the number of accidents.

## Why Use Season\_Code:

Many variables, most commonly categorical data, typically contain seasonal information. In order to integrate this into an analysis in statistical modeling, each category would be recoded into numerical codes, such as Season\_Code. This will allow the model to use each season as a different category but still give the ability to estimate the effect of each season on the number of accidents. The seasonal codes are just arbitrary numerical representations for each season, used by the model to make estimates of the effect of each season on the number of accidents.

The graph indicates the number of accidents according to different seasons:



## Winter:

It has the highest number of accidents at 13.

## Spring:

It has a lower number with 4 accidents.

## Summer:

It records 8 accidents.

## Autumn:

It has the lowest count with 2 accidents.

This graph enhances the findings from the GLM by giving a visual idea of how the number of accidents changes with seasons. It shows that winter is the season when there are the most accidents, thus agreeing with the positive coefficient for Season\_code. It helps illustrate with a real-world pattern behind the numerical results for the GLM, showing the practical differences in accident rates across the seasons.