# 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:

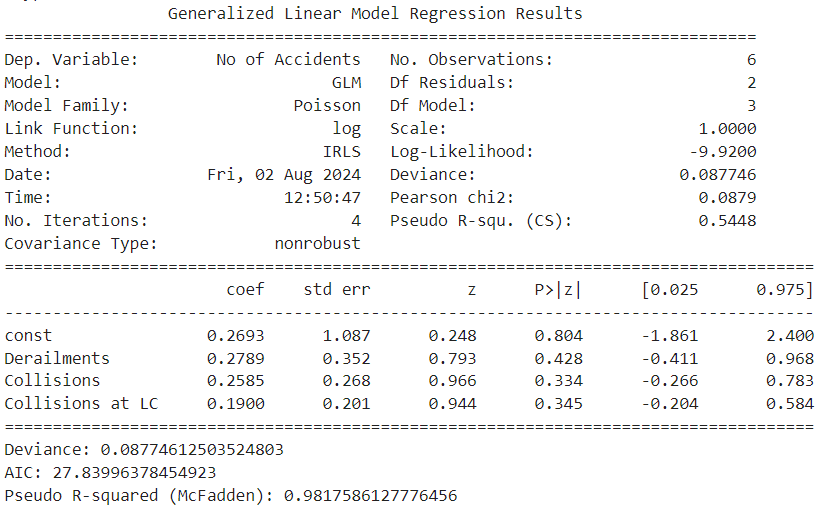
Depending on the evaluation results, refine the model if necessary. This might involve adjusting predictor variables or link functions or exploring other distributions that give the best fitting and accuracy of the model.

## 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.

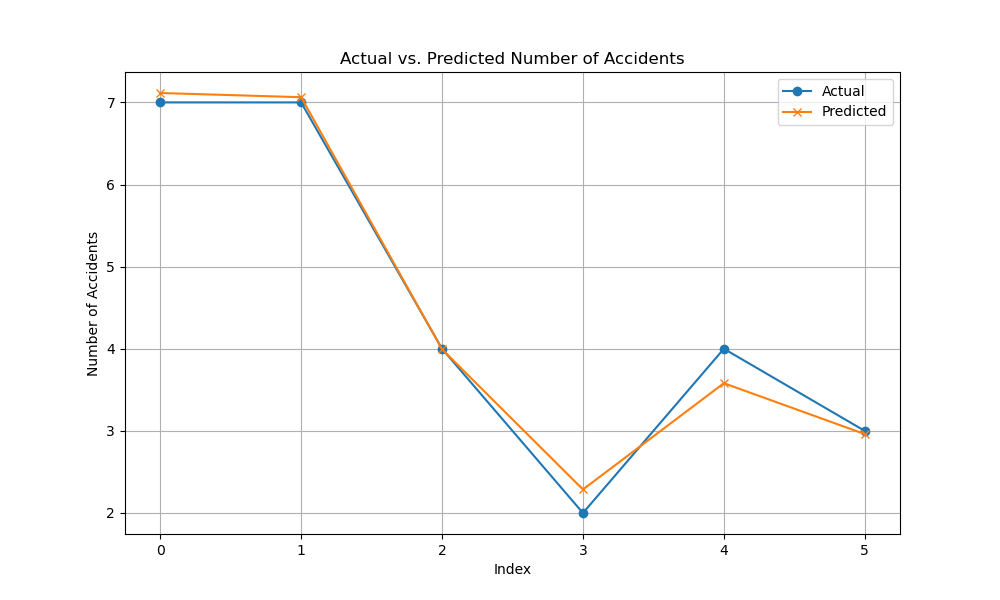
* **Derailments**: A coefficient of 0.2789 suggests that each additional derailment is associated with a 27.89% increase in the number of accidents. However, the p-value (0.428) indicates this result is not statistically significant.
* **Collisions**: A coefficient of 0.2585 indicates that each additional 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

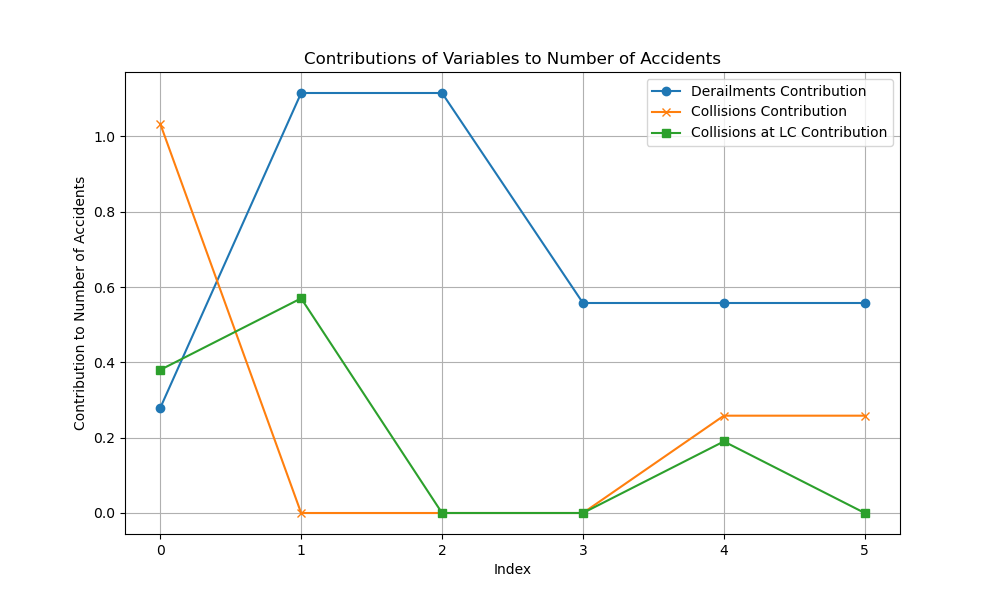


**Description:** The first graph displays the actual versus predicted number of accidents over the years. The actual values (represented by blue circles) and the predicted values from the GLM (represented by orange crosses) are plotted for each year from 2016 to 2021.

**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

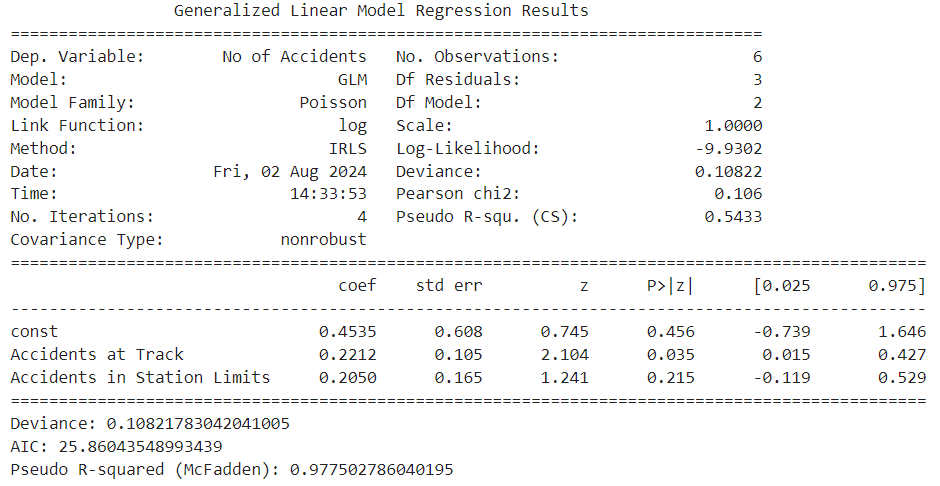


**Description:** The second graph illustrates the contributions of each predictor variable (Derailments, Collisions, and Collisions at Level Crossings) to the number of accidents. Each variable's contribution is plotted as a line showing its impact across the years.

**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.

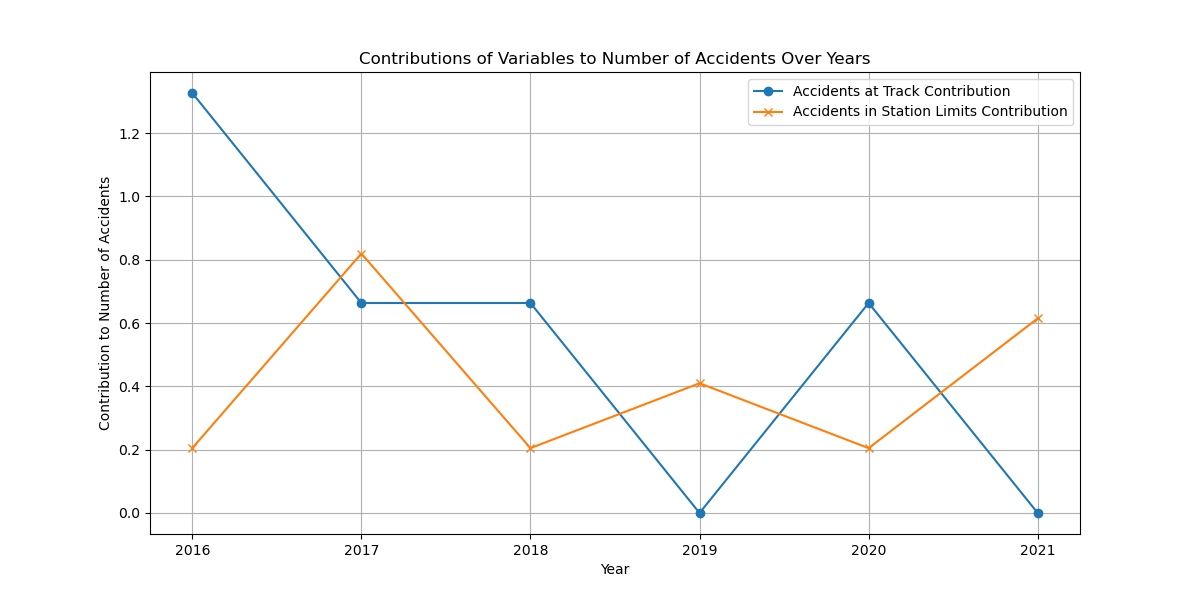
## AIC

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



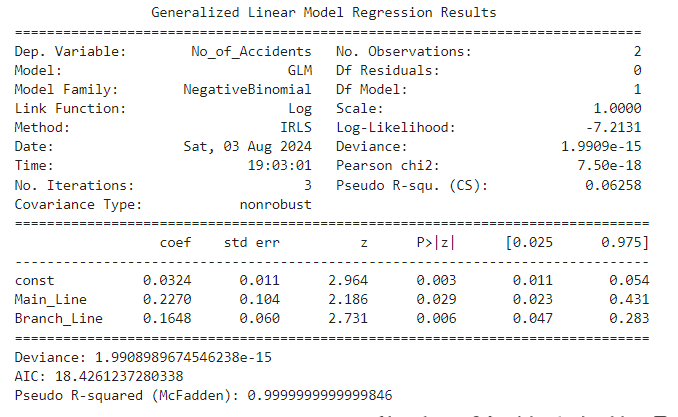
## Accidents at Track:

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

In this analysis, we used the Negative Binomial Generalized Linear Model (GLM) to model count data that may be overdispersed. The Negative Binomial model is especially useful when the count data's variance is larger than its mean, a situation which may arise with very sparse data or when count data are very variable in size.

## Deviance:

The value of deviance is very small (1.99×10−15), indicating that this model fits the data almost perfectly.

## AIC:

According to the Akaike Information Criterion, 18.43 is. The lower the AIC value, the better the fit; however in this case, the very small deviance does suggest the model fits the data very well.

## Pseudo R-squared (McFadden):

The pseudo R-squared comes out to be very nearly 1, indicating the almost perfect fit of the model. This is because of the very low deviance, which tells that the model explains almost all the variance in the data.

## Interpretation of Coefficients:

## Intercepts:

The coefficient on the intercept is 0.0324, with a standard error of 0.011. This coefficient has a p-value of 0.003, hence significantly different from zero, indicating nonzero baseline levels of accidents.

## Main Line:

The coefficient here is 0.2270, with standard error 0.104. This is statistically significant (p-value = 0.029), so a unit increase in accidents on the Main Line is associated with a 25.5% increase in the total number of accidents(e0.2270−1≈0.255).

## Branch Line:

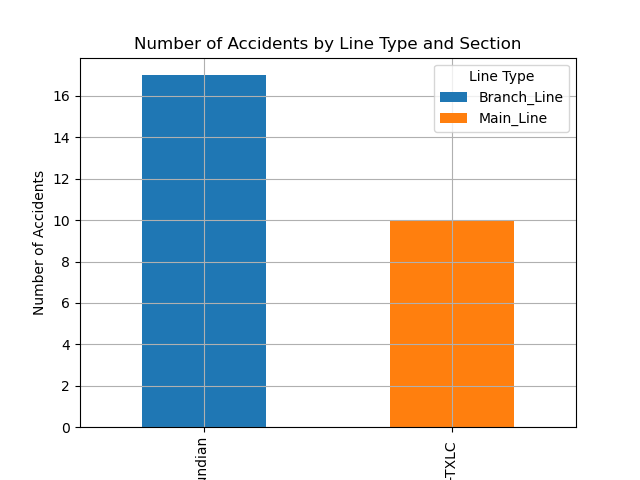
0.1648 with a standard error of 0.06. This proved to be statistically significant as well because the p-value came out to be 0.006, so the interpretation goes that a one-unit increase in accidents for the branch line is associated with a 17.9% increase in the total number of accidents since ???? 0.1648 −1 ≈ 0.179. Graph Interpretation:

Reason for using Negative Binomial Model:  
One will use the Negative Binomial model since it has the capability of handling overdispersion. Overdispersion describes the situation when the count data variance is greater than the mean. This happens quite often in real data, where simple Poisson regression might sometimes fail.

In our case, with few observations and perhaps overdispersion, the Negative Binomial model will fit more reliably than Poisson regression. Such a method of analysis is the key to appropriate modeling and interpretation of count data with high variability for valid statistical inferences to be drawn about the impacts of Main Line and Branch Line accidents on total accidents.

## 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.

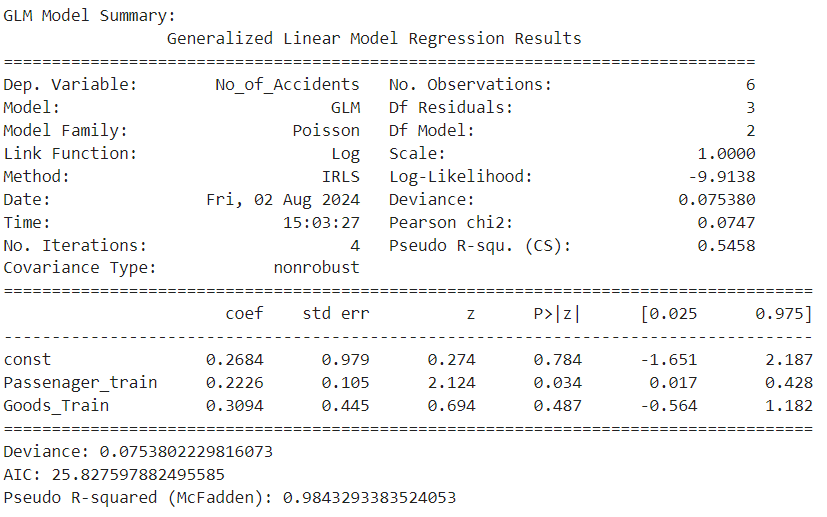
## Jhand -Kundian Section

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.

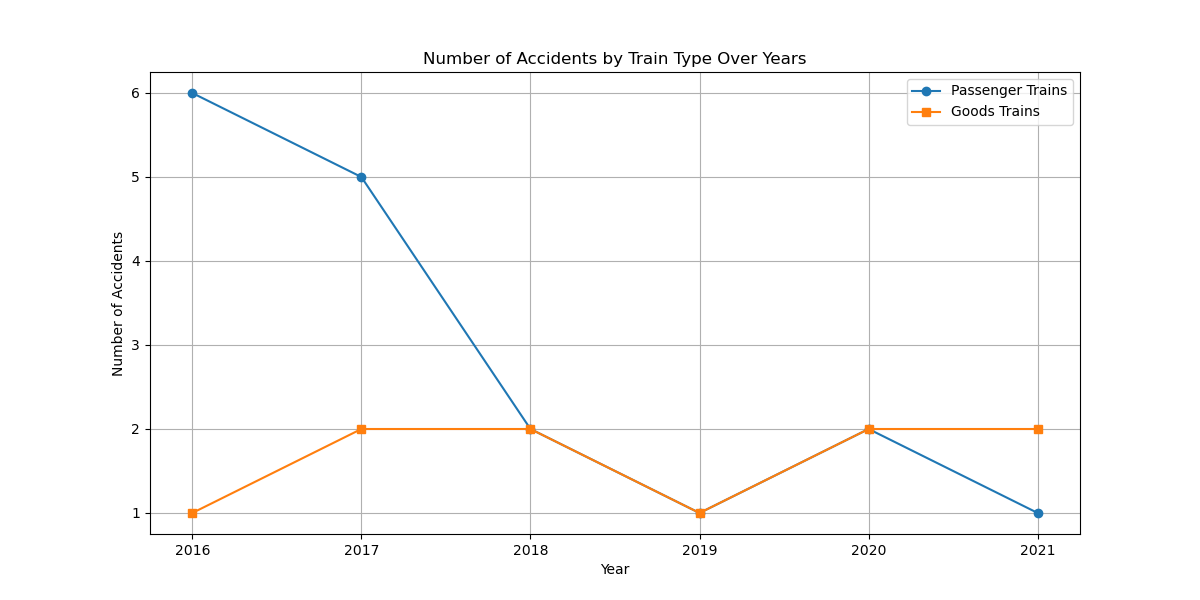
## AIC:

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

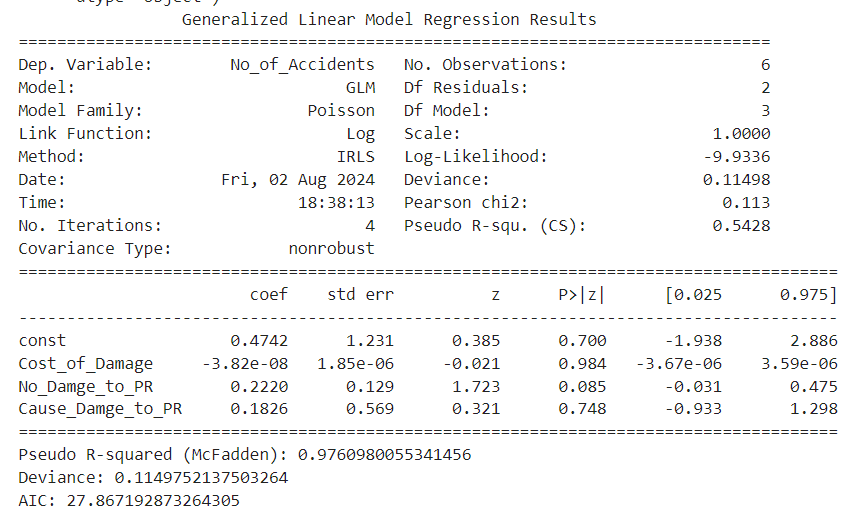
he trend in the number of accidents related to passenger trains generally shows a decreasing trend over the years, with a visible drop from 6 in 2016 to 1 in 2021. This trend represents a major reduction in mishaps related to passenger trains.

## 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:

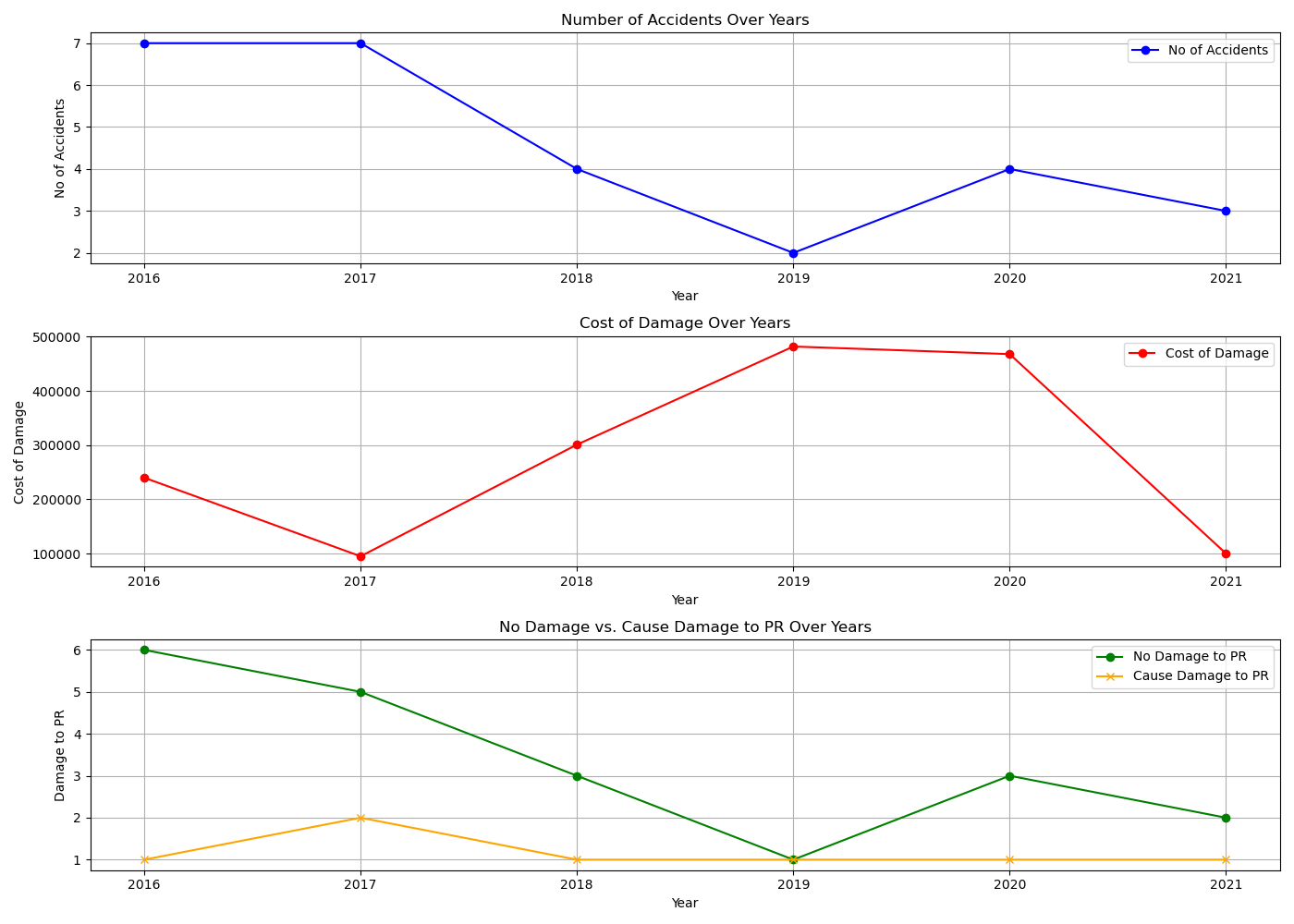
Since the coefficient is positive, increasing the number of incidents with no damage reported may increase accidents. However, the effect is not significant at a 0.05 significance level, indicating that the relationship is weak, and no conclusive results can be made. Cause\_Damage\_to\_PR: Coefficient: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

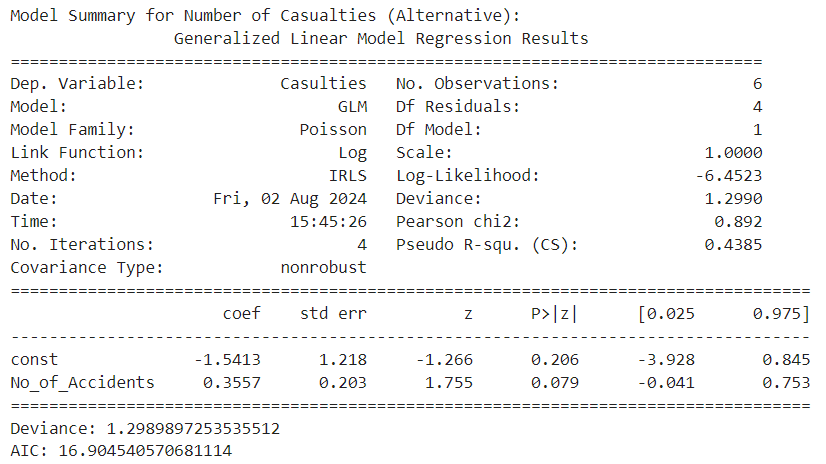
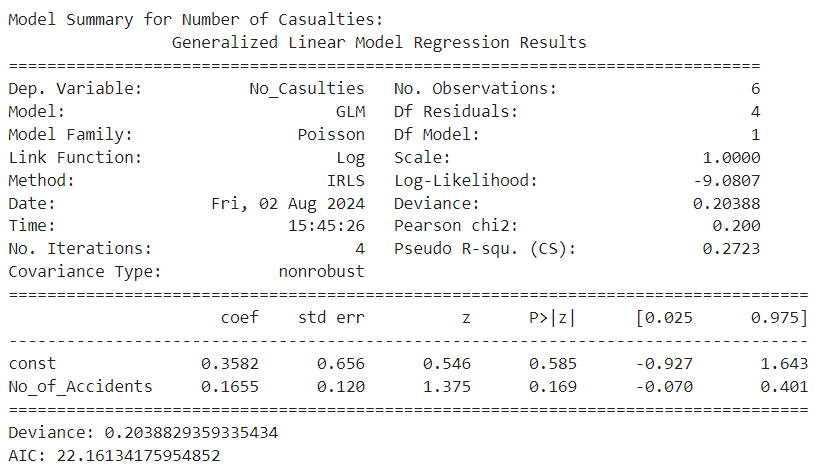


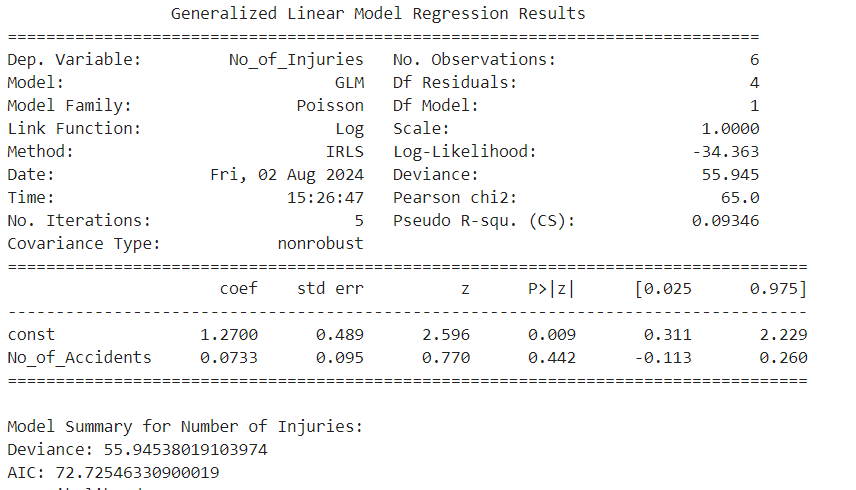
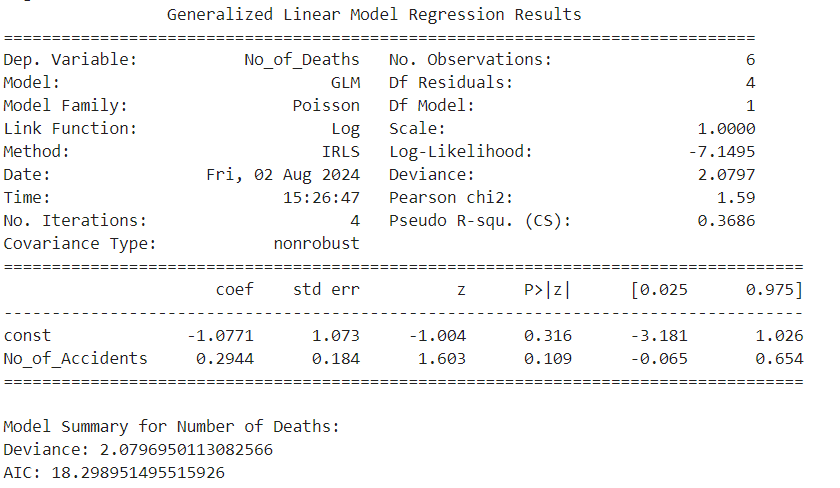
The first plot depicts the number of accidents per year. From this data, one can perceive the fact that the number of accidents is variable with no trend in it. This variability of accidents seems to vary from 2-7 every year, proving that some other external factors might be responsible for accidents,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

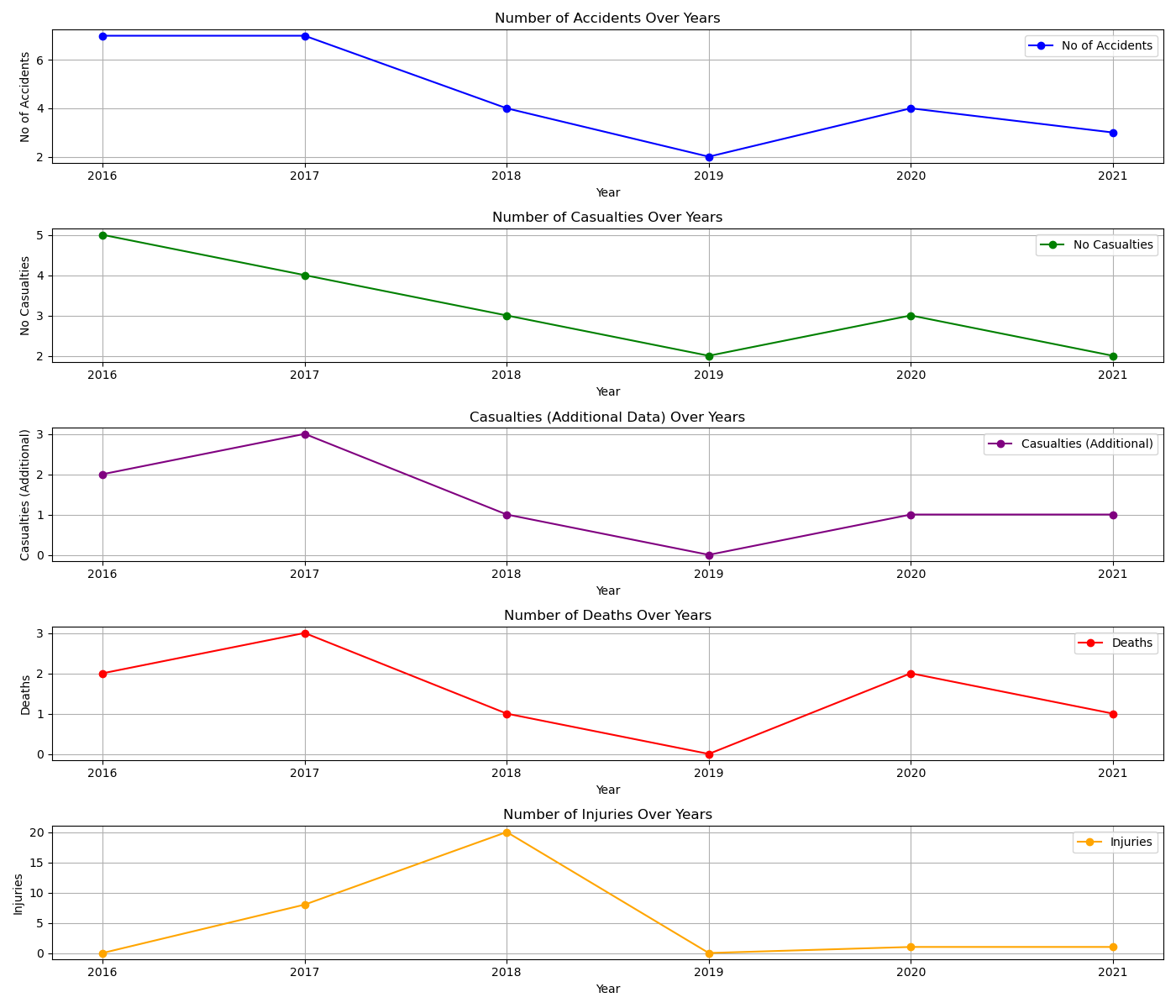
## Deviance 55.945

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.

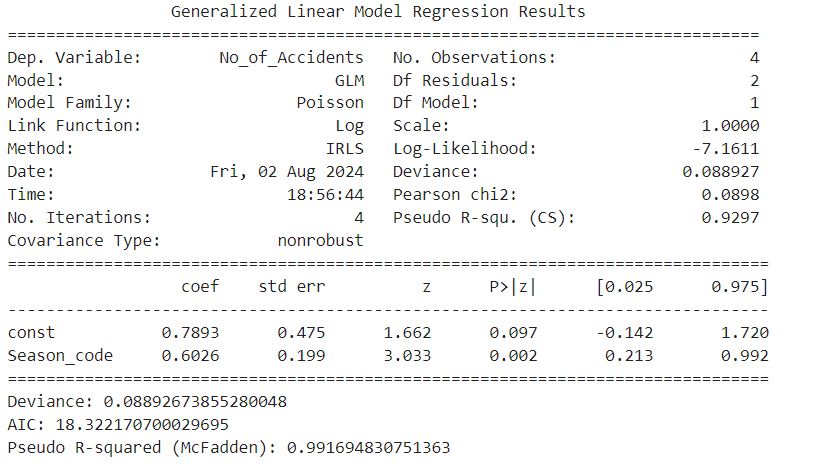
Number of Casualties Over Years: Casualties change greatly, peaking in 2017 and 2018, and then falling.

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:

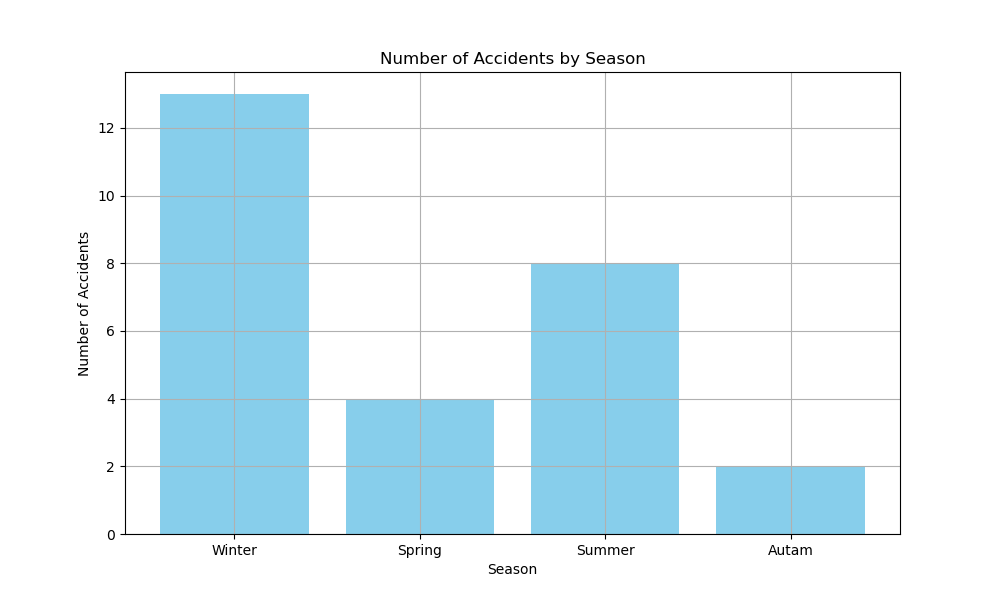
The coefficient for the intercept is 0.7893, with standard error 0.475; this is the expected log count of accidents when all other variables are zero. Its p-value of 0.097 suggests that it is not statistically significant at the 5% level but might be of further interest.

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.

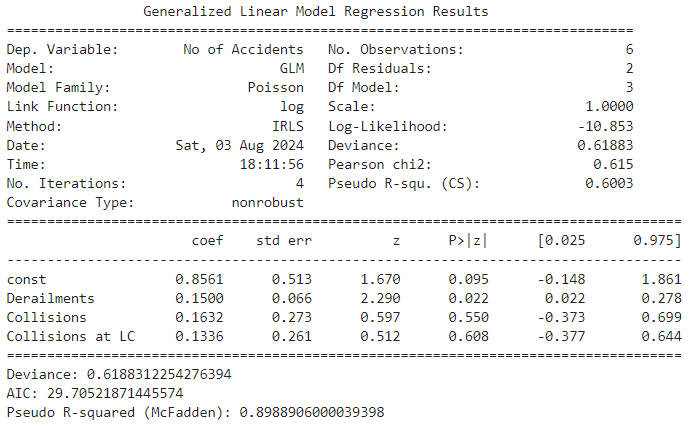
## Autumn:

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.

Analysis of Quetta Railway Accidents

# Accident Types Analysis



## Model Summary

The GLM analysis of railway accident data gave the following key results for the major variables:

## Deviance: 061883

## AIC: 29.7052

## Pseudo R-squared: 0.6003

## Interpretation of Coefficients

The coefficients in the model are interpreted as log expected change in accidents for a one-unit change in each predictor variable, net of the others.

## Derailments:

With a coefficient of 0.1500, this means that for each additional derailment, it increases the number of accidents by 15.00%. The p-value of 0.022 shows this result to be statistically significant.

## Collisions:

A coefficient of 0.1632 means that with one additional collision, there will be a 16.32% increase in the number of accidents. But the result is not statistically significant since the p-value is 0.550.

## LC Collisions:

A 0.1336 coefficient shows that for every additional collision at a level crossing, there will be 13.36% more accidents. The p-value of 0.608 shows this is not statistically significant.

## Fit metrics

## Deviance:

The model's deviance of 0.61883 is a fit metric—small deviance goes with good updipation.

## AIC:

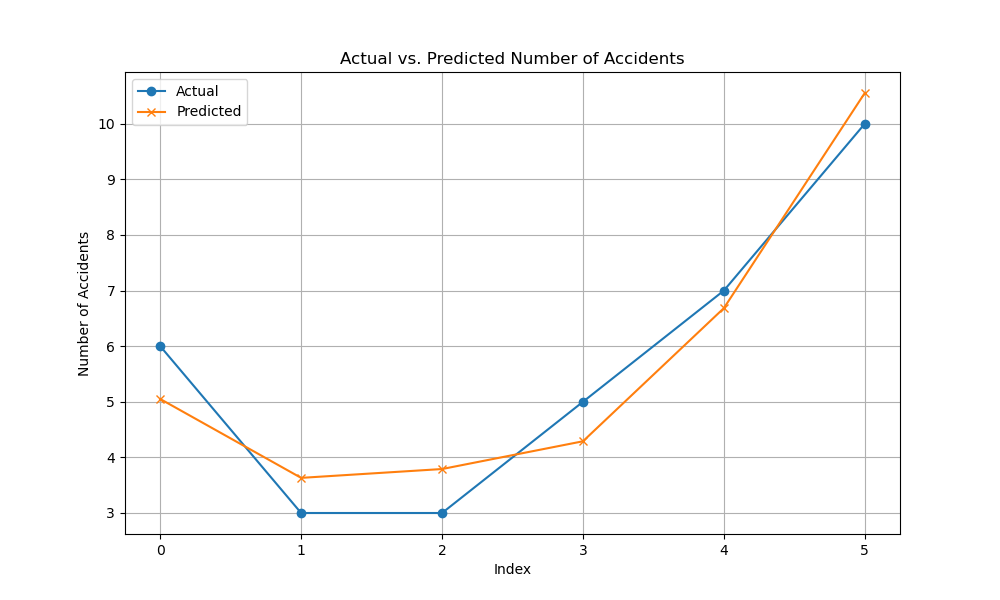
The AIC value is 29.7052 and it is useful for model comparison; the lower its value, the better is the model.

## Pseudo R-squared (McFadden):

With the value of 0.6003, this pseudo R-squared indicates that the model explains about 60.03% of the variance in the number of accidents, which demonstrates a medium model fit.

## Discussion of Graph Results

Graph 1: Observed vs. Fitted Number of Accidents



## Description:

The first graph indicates the actual vs. predicted number of accidents across the years. The actual values are given by the blue circles, and the predicted ones by the GLM are shown by the orange crosses for each year from 2016 to 2020.

## Interpretation

## Alignment:

The scatter graph of the predicted values against the actual values depicts a close alignment in some years but less in others. This indicates that the GLM model has some moderate performance in predicting the number of accidents based on the given predictors.

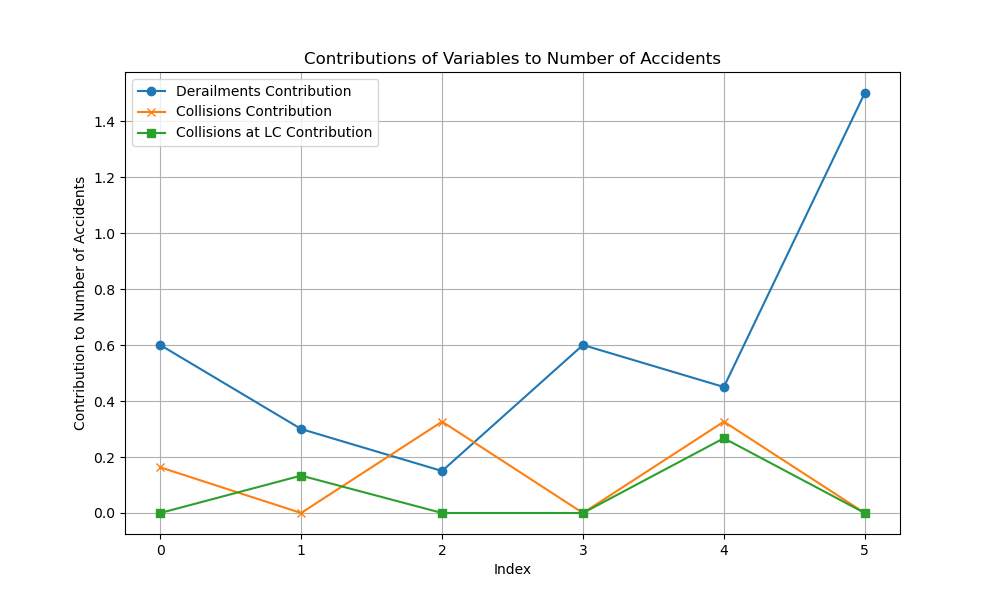
## Trend:

Both the actual and fitted lines follow similar trends across the years, suggesting that the model captures some variation in the number of accidents.

## Model Fit:

The graphical closeness of the two lines in most of the years reinforces that the model has a medium fit, as indicated by the pseudo R-squared value of 0.6003.

Graph 2: Variables' Contributions to Number of Accidents



## Description:

The second graph is that describing what each predictor variable—Derailments, Collisions, and Collisions at Level Crossings—contributes to the number of accidents. The contribution of each variable is plotted as a line showing its influence across the years.

## Interpretation:

## Derailments:

From the blue line, it is easy to deduce that the contribution of derailments in explaining the number of accidents is quite noticeable, with overwhelming effects in some years.

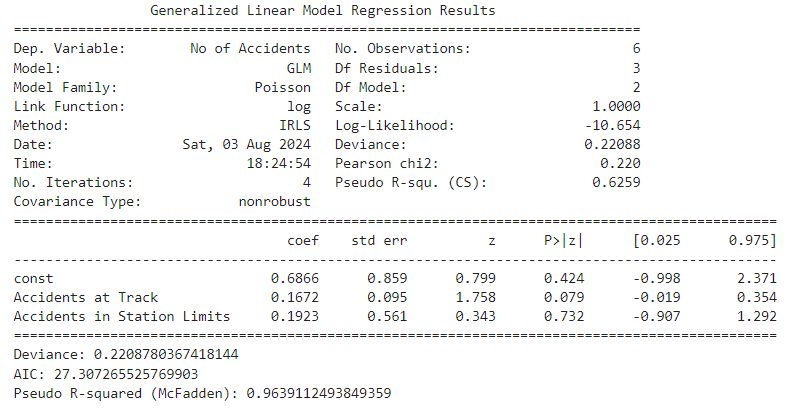
## Collisions:

The contribution of collisions, as depicted by the orange line, is a little, but not sharp and it changes over the years.

## Collisions at Level Crossings:

By comparison, the contribution coming from level crossings, reflected by the green line, stands relatively low.

# Accident Location Analysis



## Summary of the Fit of the Model

## Deviance is 0.22088

## AIC Akaike Information Criterion is 27.3073

## Pseudo R-squared McFadden is 0.6259.

## Interpretation of Coefficients

Model coefficients show what the log of the expected value of increasing one unit of each of the predictor variable increases the likelihood of, keeping other.

## Accidents at Track:

Coefficient of 0.1672, which can be interpreted to mean one more accident on the track will increase the overall number of accidents by 16.72%. P-value equals 0.079, a result that is very close to being statistically significant.

## Accidents in Station Limits:

For a unit increase in accidents at station limits, we find a coefficient estimate of 0.1923. Therefore, the number of accidents increases by 19.23% with an additional accident at stations. The p-value of this test is 0.732, so the null hypothesis is not rejected.

## Model Fit Metrics

## Deviance:

The deviance of the model, based on the model under calculation, is 0.22088. It is an indicator of goodness of fitness with other models: if the deviance is smaller, then it indicates a better fit.

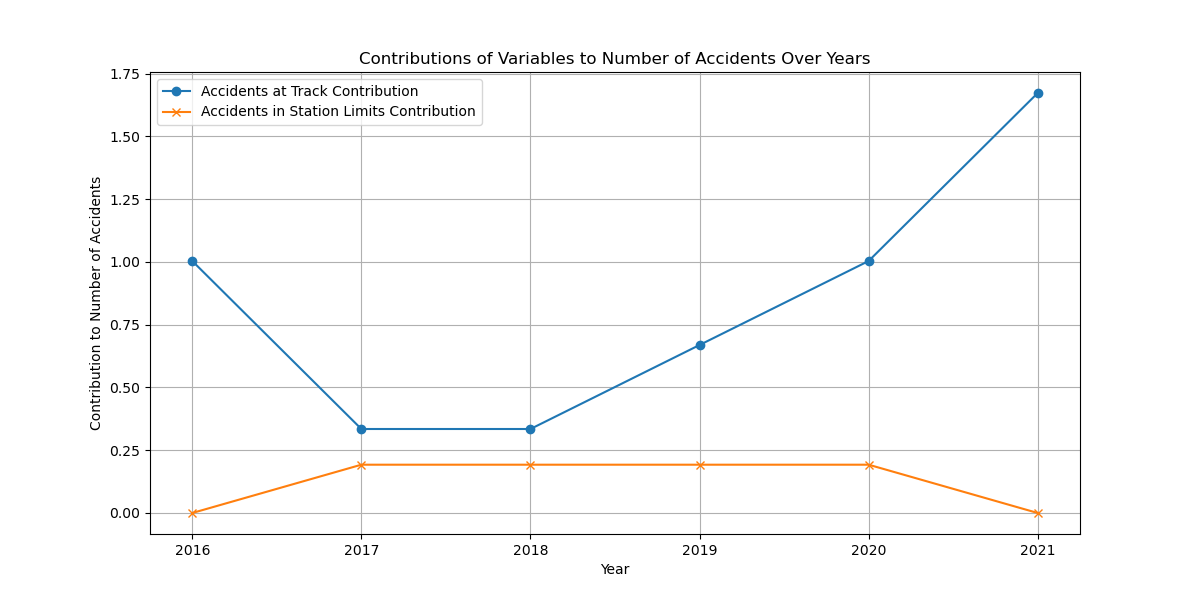
## 27.3073 is the AIC,

It helps in model comparison. The lower the value of AIC, the better is the model since it suggests a better package of information.

## Values of Pseudo R-squared (McFadden):

Values suggest that 0.6259 means that the model explains around 62.59% of the variance in the number of accidents, more or less moderate.

## Discussion of Graph Results

Graph : Variables' Contribution to No. of Accidents 

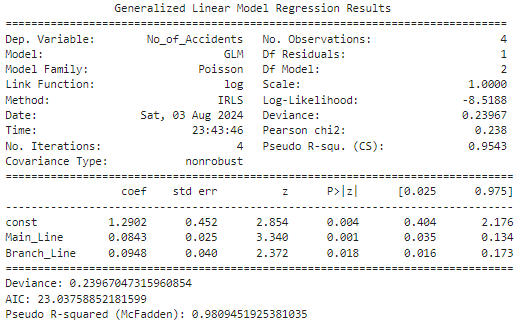
This is a graph that shows the contributions from the predictor variables of Accidents at Track and Accidents in Station Limits through the years to the number of accidents.

This blue line is contributing significantly in 2016 and, respectively, 2020 and means it has placed as a major variable which had contributed to the total accident number those years. While 2017 and 2018, the increment went down, which can be due to a lesser number of accidents at the track during those years. In 2019, the increment again is before a rising curve of the actual accidents at the track.

The orange line says accidents in the station limits, which are a smaller impact and more stable. The contribution starts for the year 2017 and goes flat until 2020; this is like the actual data, only one accident within station limits for the same years. In all, it is a stable trend showing that while it tends to contribute respectively much to the total of accidents, the accidents in station are of much less consequence than the really big accidents at track.

In general, it shows that accidents at tracks make more substantial and variable contributions to the total amount of accidents, while station limits add up to having smaller but very consistent contributions.

# Accident on Line Type Analysis



## Model Summary Coefficients

## Intercept const:

## coefficient 1.2902

### standard error 0.452;

The evidence that the overall intercept is statistically significant.

## Main Line

The coefficient is 0.0843, the standard error is 0.025, so the z-value is 3.340 with a p-value of 0.001. This would indicate a positive relationship between Main Line accidents and total accidents that is statistically significant.

## Branch Line:

The coefficient is 0.0948, having a standard error of 0.04, which gives a z-value of 2.372 with a p-value of 0.018. This indicates that there is a positive relationship statistically significant to Branch Line accidents against the total number of accidents.

## Interpretation of Coefficients

## intercept:

This corresponds to the baseline log count of accidents when counts of accidents on the main line and branch line are both zero, and it has a positive and significant coefficient, indicating a large baseline level of accidents.

## Main Line:

The positive coefficient of 0.0843 means that, all other things held constant, for every additional accident on the main line, the log of the expected number of total accidents increases by 0.0843 units. The relationship is statistically significant with a p-value of 0.001.

## Branch Line:

The positive coefficient of 0.0948 shows that for each additional accident that occurs on the Branch Line, the log of expected number of accidents of all types increases by 0.0948 units if main line accidents are held constant. The relation is statistically significant with a p-value of 0.018.

## Metrics for Goodness of Fit

### Deviance:

The deviance of 0.23967 indicates that the model fits well to the data. Generally, the smaller the deviance, the better the fit.

## AIC:

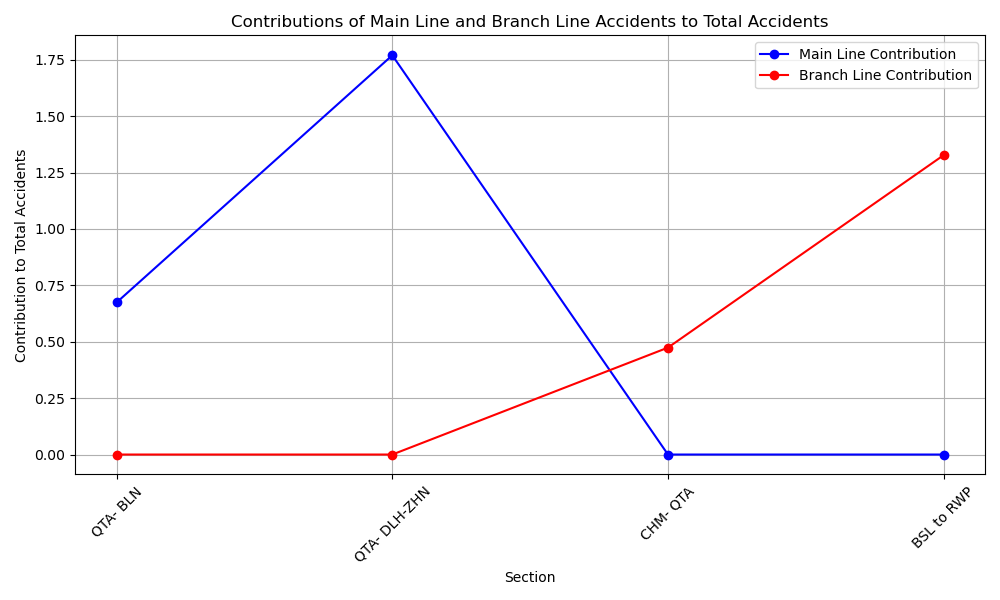
Akaike information criterion is 23.0376. AIC is a measure of how well a model fits the data; AIC values close to zero are indicative of very good fits, and smaller is always better.

## Pseudo R-squared (McFadden):

Given the pseudo R-squared value of 0.9809, the variability of the number of accidents explained by the model can be considered as high as 98.1%, meaning it will be very strongly fitted.

## Graph Analysis

Contributions of Main Line and Branch Line Accidents to Total Accidents



## Main Line Contribution:

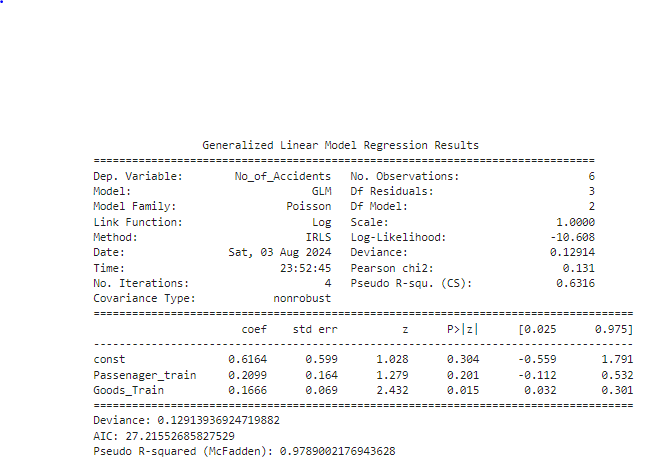
This contribution of the Main Line accidents is to the total number of accidents. Every point on the blue line is related to a section, and the contributions are computed by considering the number of Main Line accidents from the concerned section and the estimated coefficient from the Poisson model. In the case of sections QTA-BLN and QTA-DLH-ZHN, the contribution of the Main Line to the total number of accidents is very high.

## Branch Line Contribution:

The red line plots the contribution of the branch line accidents to the total accidents. The points on this line depict each section, and for every such point in a section, it graphs the calculated contribution based on the number of branch line accidents in that section with the estimated coefficient from the Poisson model. These contributions are high for the sections CHM-QTA and BSL to RWP.

It can be observed in the graph that the contribution of Main Line and Branch Line accidents varies across sections. Thus, those sections which have a higher contribution by the share of Main Line accidents, like QTA-BLN and QTA-DLH-ZHN, contribute much to the total accidents from Main Line accidents. Similarly, contributions that have a high share of Branch Line accidents, such as CHM-QTA and BSL to RWP, are more influencing in their total count of accidents.

# Accident of Train Types Analysis



## Model Summary

A Generalized Linear Model with a Poisson distribution and log-link function was fitted for the relationship between the type of train and the number of accidents, with variables: constant, Passenger Train accidents and Goods Train accidents.

## Coefficients

## Intercept:

The coefficient is 0.6164 with a standard error of 0.599 and gives a z-value of 1.028 with a p-value of 0.304, indicating this intercept is not statistically significant.

## Passenger Train:

The coefficient is 0.2099 with a standard error of 0.164, a t-value of 1.279, and a p-value of 0.201. This would suggest a non-significant relationship between the total number of accidents and those relating to passenger trains.

## Goods Train:

This has a coefficient of 0.1666 with a standard error of 0.069, making a z-value of 2.432 with a p-value of 0.015, hence statistically significant in establishing a positive relationship between the Goods Train accidents and the total number of accidents.

## Interpretation of Coefficients

## Intercept:

The intercept is the baseline log count of accidents in the absence of counts of both Passenger Train and Goods Train accidents. The coefficient is positive but statistically non-significant, and hence one cannot conclusively determine the baseline level of accidents from this model.

## Passenger Train:

The coefficient is positive, 0.2099, so for every additional accident that occurs to a Passenger Train, the log of the expected number of total accidents will increase by 0.2099 units, holding Goods Train accidents constant. The relationship is not statistically significant due to a p-value of 0.201.

## Goods Train:

The positive coefficient of 0.1666 indicates that for every additional accident involving a Goods Train, the log of the expected number of total accidents goes up by 0.1666 units, while holding constant the number of Passenger Train accidents. This relationship is statistically significant with a p-value of 0.015.

## Model Fit Metrics

## Deviance:

The deviance of 0.12914 suggests that the model fitted well with the data. For most cases, a lower deviance value will indicate a better fit.

## AIC:

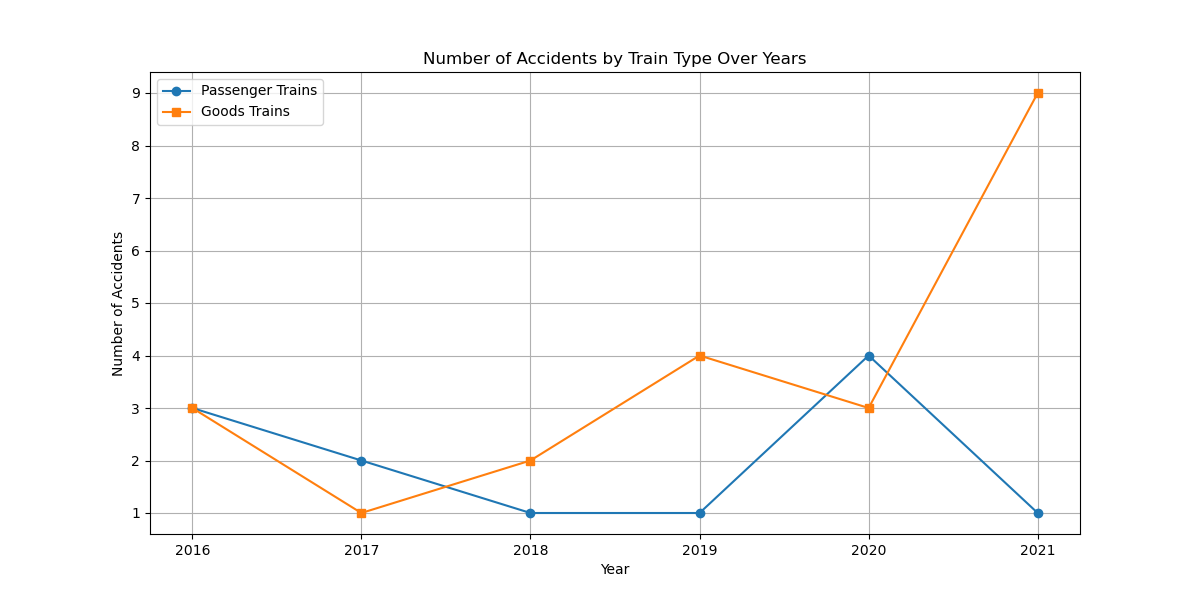
Akaike Information Criterion: The AIC value is 27.2155. It is used for model comparison. The model with a lower AIC value will have a better fit.

## Pseudo R-squared (McFadden):

The pseudo R-squared value is 0.9789, so the model would fit very strongly, explaining about 97.89 percent of the variability in the number of accidents.

## Graph Analysis

Contributions of Passenger Train and Goods Train Accidents to Total Accidents



This plot helps in understanding how much of the total count of accidents each type of accident has relatively contributed.

## Contribution of Passenger Train:

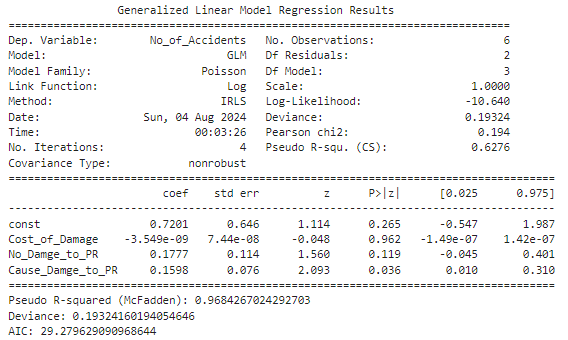
The blue line describes the contribution of Passenger Train accidents to the total number of accidents in the graph. Each point on the line corresponds to a different year and plots the contribution calculated by the number of accidents concerning the variable Passenger Train within that year and the estimated coefficient by the Poisson model. From the graph, contributions in some years, like 2016 and 2020, are high with respect to passenger train accidents, while in other years, such as 2019 and 2021, the contribution is low.

## Goods Train Contribution:

That contribution of Goods Train accidents to the total number of accidents is shown by the red line. Every point on this line refers to a year and plots the calculated contribution based on the number of Goods Train accidents for that particular year and the estimated coefficient from the Poisson model. The graph shows that Goods Train accidents have always contributed a fair share of the total accidents over the years, with a noticeable increase in 2021.

The graph shows the relative impact of Passenger Train and Goods Train accidents across different years. In this case, the steady contribution of Goods Train accidents indicates that these accidents are core to the overall number of railway accidents. In contrast, the contribution of Passenger Train accidents seems more variable, indicating that specific incidents or factors could have occurred in certain years that impacted these accidents.

# Cost of Damage Analysis



## Model Summary

The Generalized Linear Model with the Poisson distribution and a log link function was fitted to estimate the effect of various predictors on accident frequency. The variables included are Cost of Damage, No Damage to PR, and Cause Damage to PR.

## Coefficients

The coefficient is 0.7201, the standard error is 0.646, and the z value is 1.114 with a corresponding p-value of 0.265. Due to this large p-value, the intercept is not statistically significant.

## Damage Cost:

The coefficient is -3.549e-09 with standard error 7.44e-08, so that z = -0.048 with p-value 0.962. This would suggest that there is no statistically significant relationship between the cost of damage and the number of accidents.

## No PR Damage:

The coefficient is 0.1777 at standard error 0.114, producing a z-value of 1.560 with a p-value 0.119. This means a positive relationship between no damage to PR and the number of accidents; however, the relationship is not statistically significant.

## Damage to PR:

Coefficient 0.1598; standard error 0.076; z-value 2.093; p-value 0.036: This means that there is a positive relationship, statistically significant, between Cause Damage to PR and the number of accidents.

Interpretation of Coefficients

## Intercept:

This is the log count of accidents at base when all other variables are zero. Its coefficient is positive but non-significant, so baseline accidents are not definitively captured by this model.

## Cost of Damage:

Given the negative coefficient, an increase in the cost of damage slightly reduces the expected number of accidents, although at a p-value of 0.962, it is not statistically significant.

## No Damage to PR:

The positive coefficient of 0.1777 means that as cases where there is no damage to PR increase by one, the log of the expected number of accidents increases by 0.1777 units, other things being equal. However, this relationship is not statistically significant.

## Damage to PR:

The coefficient is positive, at 0.1598, implying that for every additional case of damage to PR, the log of the expected number of accidents increases by 0.1598 units, holding all other factors constant. The relation is statistically significant with a p-value of 0.036.

## Fit Metrics

## Deviance:

Deviance of 0.19324 indicating the model fits well to data. The smaller, the better it usually is.

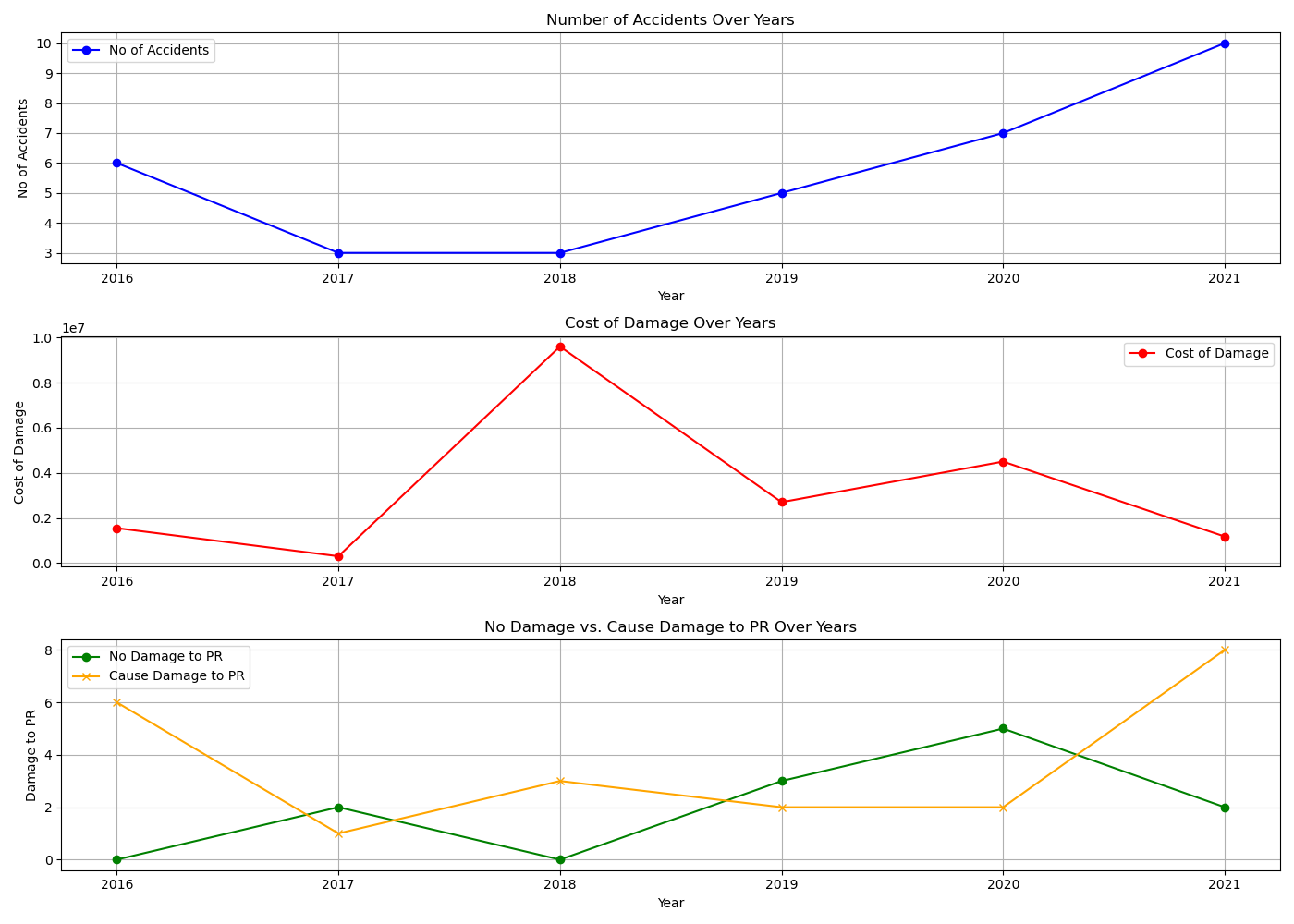
## AIC :

AIC is 29.2796. (Used for comparing models. AIC should be as small as possible.).

## Pseudo R-squared (McFadden):

A pseudo R-squared value of 0.9684 suggests that the model fits very strongly, explaining about 96.84% of the variability in the number of accidents.

## Graph Analysis



The graphs presented have been plotted from data to show trends for the number of accidents, cost of damage, and damages to PR during the period from 2016 to 2021.

## Graph 1: Number of Accidents Over Years

The first subplot indicates the number of accidents during these years. There is an evident spike in the number of accidents in 2021, which goes up to a maximum of 10. Other years show a very fluctuating trend where the count of accidents has varied between 3 and 7.

## Graph 2: Cost of Damage Over Years

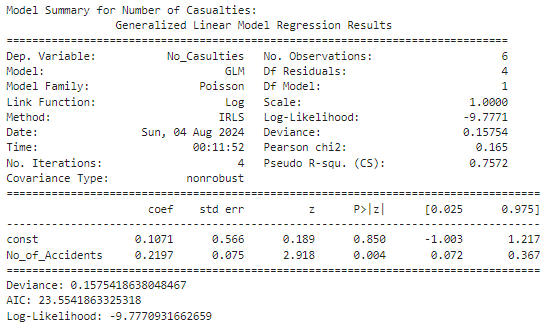
The cost of damage over the years is represented through the second subplot of the graph. In 2018, the cost of damage is the highest, accounting for 9,600,000. Other years portray varying costs that sometimes considerably change, hence proving that some of the years were financially impacted by accidents more than others.

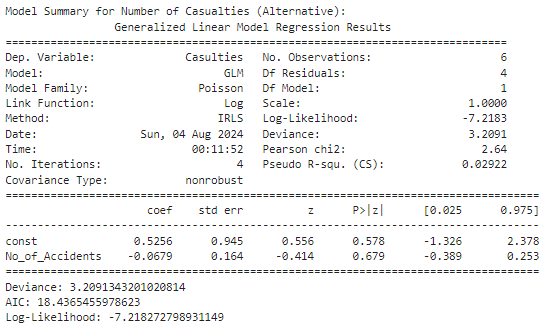
## Graph 3: No Damage vs. Cause Damage to PR Over Years

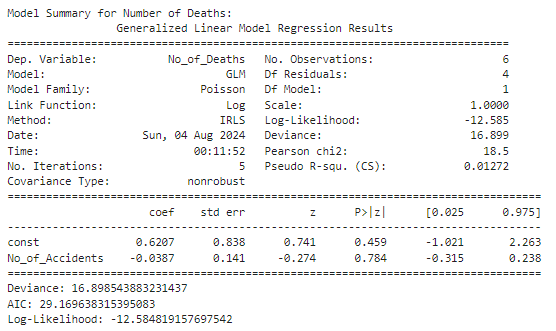
The third subplot indicates the number of incidents with PR damage versus those without. It can be noted from the graph that those incidents causing damage to PR have highly contributed to accidents, mostly in 2016 and 2021. A general trend is evident on the green and orange lines whereby a higher count of accidents corresponds to a number of incidents causing damage.

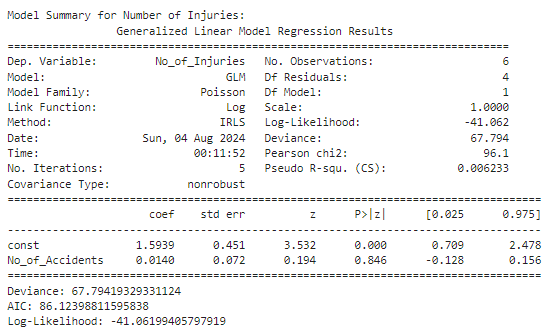
Indeed, cost of damage does not influence accident count substantially; however, cases of damage to PR show a positive relationship with accident frequency. This finding suggests that focusing on preventing damage to PR may actually be effective in reducing the overall railway accident count. The graphs further depict the variability in the count of accidents and financial impacts across the years, thereby bringing out an important message on the essence of constant surveillance with focused interventions to improve Railway Safety and reduce costs.

# Analysis of Accident casualties/injuries/deaths









## Summary Model of Number of Casualties

## Overview of the Model

This model that evaluates the number of accidents and the relationship it had on the number of casualties is statistically significant to prove an increase in the number of accidents increases the number of causalities correspondingly.

## Coefficients Interpretation

## Constant:

The intercept term, 0.1071, is the expected log count of casualties when there are no accidents. This term, however, is not significant statistically, p = 0.850, so it may not be relevant for this model.

## No\_of\_Accidents:

The coefficient for No\_of\_Accidents, 0.2197, is significant, p = 0.004, which means that for each additional accident added, the log-count of casualties will increase by 0.2197.

## Metrics on Goodness of Fit

## Deviance:

0.1575 -The very low value shows good fit.

## AIC:

23.55 -To compare models, with lower values indicating better fit.

## Log-Likelihood:

-9.777 - A measure of how well the model fits with higher values (less negative) being better.

## Summary for Number Casualties Alternative

## Overview of the Model

The alternative model assessing the relationship between the number of accidents and the other measure for casualties, Casulties, suggests there is no significant relationship.

## Interpretation of Coefficients

## Constant:

The intercept term, 0.5256, is not statistically significant with a p-value of 0.578.

## No\_of\_Accidents:

The coefficient for No\_of\_Accidents −0.0679 is not statistically significant with a p-value of 0.679, which suggests that there is no meaningful relationship with this alternative measure of casualties.

## Assessing the Model Fit

## Deviance:

3.209, which is higher than in the previous model, so it is a poorer fit.

## AIC:

18.44, which again can be used for comparing with other models.

## Log-Likelihood:

-7.218, less negative than the previous model but not indicative of a good fit.

## Model Summary for Number of Deaths

## Model Overview

This model measures the relationship between the number of accidents and the number of deaths; it returns no significant relationship statistically.

## Coefficients Interpretation

## Constant:

The intercept term, 0.6207, is not significant statistically with p = 0.459.

## No\_of\_Accidents:

The coefficient, for the number of accidents, is -0.0387, which is not significant because p = 0.784. This means there is no significant effect on the number of deaths.

## Model Fit Metrics

## Deviance:

16.899. Since this value is very high, it is a poor fit.

## AIC:

29.17. This value is higher than the other models; thus, this model has a poorer fit.

## Log-Likelihood:

-12.585. This value is lower compared to the other models; hence, this is also a poor fit.

## Model Summary for Number of Injuries

## Model Overview

A model looking at the relationship between number of accidents and number of injuries did not show any statistically significant relationship.

## Coefficients Interpretation

## Constant:

The intercept term, 1.5939, with a p < 0.001, is statistically significant and represents the expected log count of injuries when the number of accidents is zero.

## No\_of\_Accidents:

As the coefficient is 0.0140, it is not significant since p = 0.846; therefore there is no significant effect on the number of injuries.

## Model Fit Metrics

## Deviance:

It is 67.794, hence a poor fit as this value is very high.

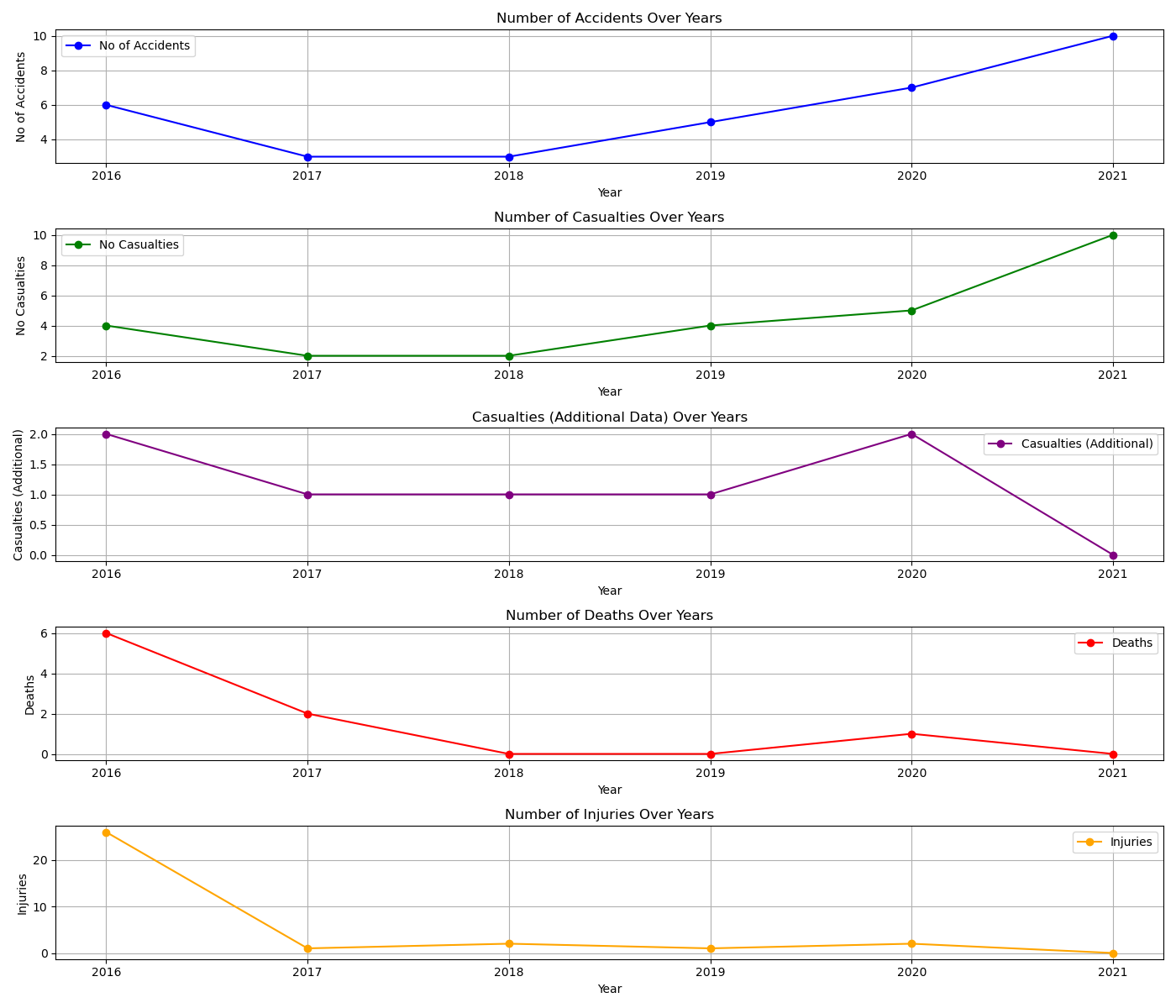
## AIC:

It is 86.12, highest amongst all the models, hence is the poorest fit.

## Log-Likelihood:

It is -41.062, lowest among all, hence poorest

## Graph Analysis



The plot shows the trend in accidents over the years. An upward trend can be seen in general, but there is a distinct peak in 2021.

Number of Casualties Over Years

This plot conveys how the number of casualties varied over the years. As can be seen, the number of casualties peaked in 2021, matching the peak for accidents.

Casualties (Additional Data) Over Years

This plot gives the alternate measure for casualties, which remains relatively low compared to the main measure.

Deaths Over Years

The count of deaths is always low, even going up to completely zero in some years.

Number of Injuries Over Years

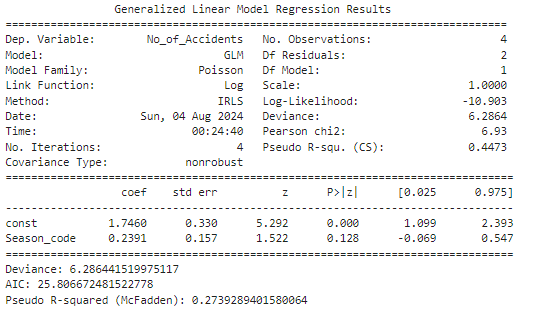
This plot shows a decreasing trend of the number of injuries over the years, with the count peaking in 2016 and almost nil in the previous years.

## Discussion

Results show that the number of accidents significantly relates to the number of casualties but not to deaths or injuries. The relationship with this alternative measure for casualties is also insignificant. This result indicates that efforts should be concentrated on reducing the number of accidents if casualties are going to be reduced, while some other factors may need to be brought into play concerning the actual deaths and injuries.

It is supported by graphical analysis, showing trends and variations over the years for different severity measures. Peak accidents and casualties in 2021 are hence a critical point that requires further investigation.

# Seasons Analysis of Accidents



## Explanation and Analysis

## Logic for Using Season Codes Instead of Season Names

Categorical variables should be encoded into numerical values as predictors in regression models, especially in GLMs. That is because encoding transforms the categories into a format usable by the model. The use of season names as is cannot work since statistical models require numerical values as inputs for computation.

Encoding is a process that transforms categorical variables into numerical values. In this case, category codes were used, where each season has a different integer code. For instance,

winter is 0,

spring is 1,

summer is 2,

autumn is 3.

Such encoding makes it easier to include categorical data in the model, which has groups that are well differentiated and stay apart. The results are presented below.

## Model Summary of Number of Accidents by Season

## Model Overview

It models the relationship between the season and the number of accidents. Since it is count data for the dependent variable, the model used is Poisson regression.

## Interpretation of Coefficients

Constant

This is the intercept term, 1.7460, which is the expected log count of accidents for the base category, Winter, which is coded 0. This term is statistically significant (p < 0.001), indicating that this base level does have a meaningful contribution toward the number of accidents.

## Season\_code:

The coefficient for the season code is 0.2391 and is not statistically significant because p = 0.128, so changes in the season code do not significantly impact the number of accidents.

## Model Fit Metrics

## Deviance:

6.286, The lower the better the fit.

## AIC:

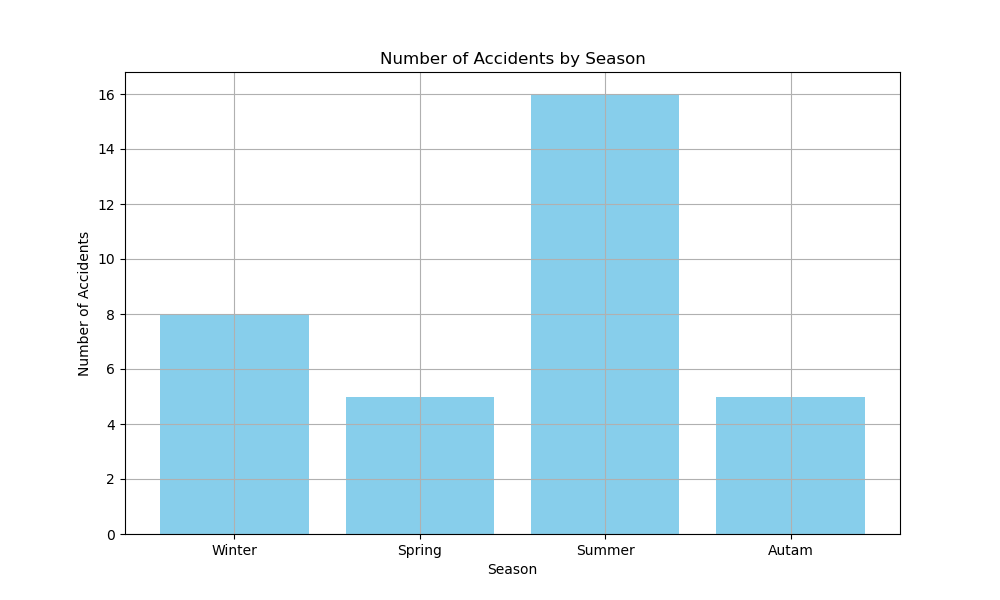
25.81, The lower the better the fit; used for model comparison.

## Pseudo R-squared (McFadden):

0.2739, a measure of the fit of the model where larger values approach 1 and indicate better explanatory power.

## Graph Analysis

Number of Accidents by Season



Bar plot of number of accidents by season. The plot tells:

Summer tops with the highest number of accidents (16).

Winter is next, with 8 accidents.

Spring and Autumn, in that order, have the least number of accidents, with 5 cases each.