# 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:

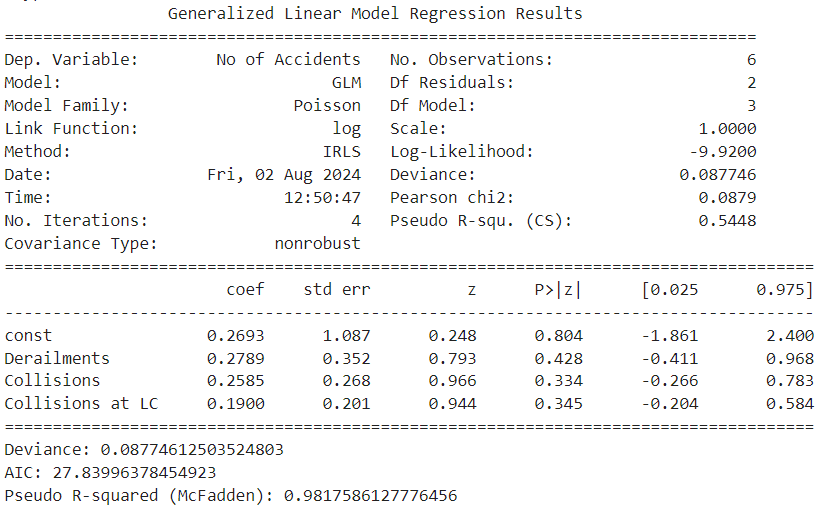
Depending on the evaluation results, refine the model if necessary. This might involve adjusting predictor variables or link functions or exploring other distributions that give the best fitting and accuracy of the model.

## 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.

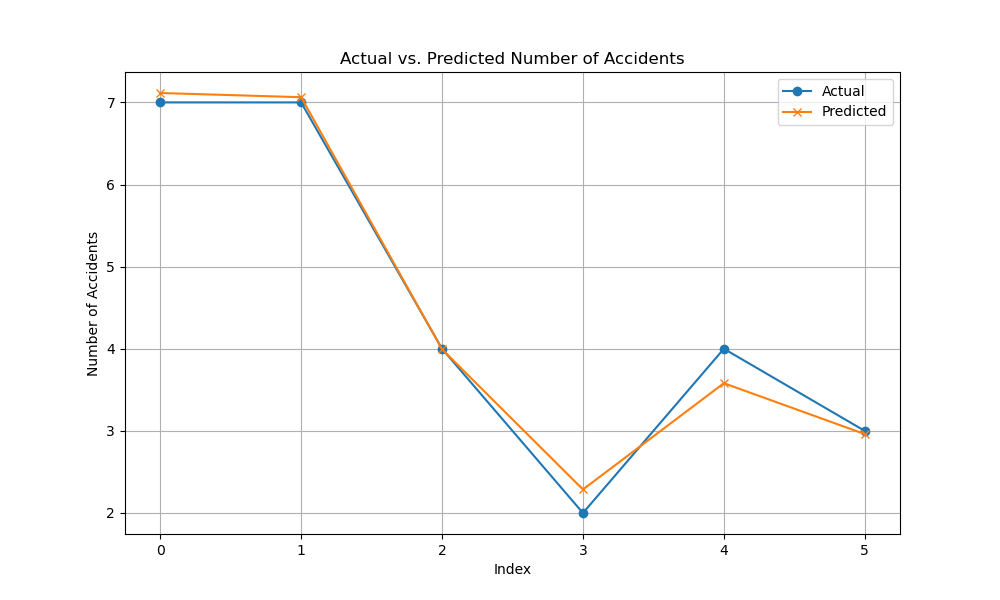
* **Derailments**: A coefficient of 0.2789 suggests that each additional derailment is associated with a 27.89% increase in the number of accidents. However, the p-value (0.428) indicates this result is not statistically significant.
* **Collisions**: A coefficient of 0.2585 indicates that each additional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

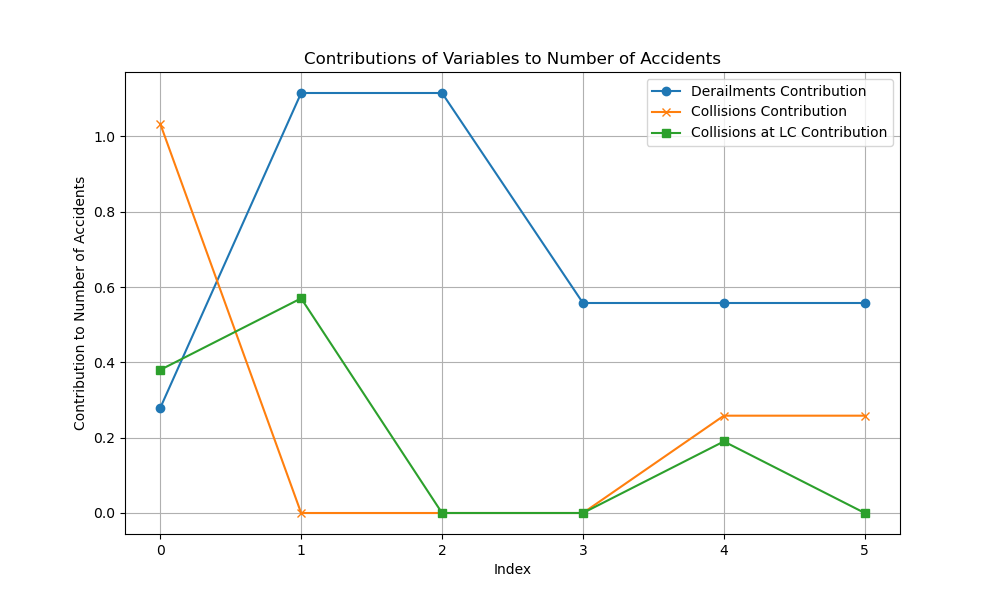


**Description:** The first graph displays the actual versus predicted number of accidents over the years. The actual values (represented by blue circles) and the predicted values from the GLM (represented by orange crosses) are plotted for each year from 2016 to 2021.

**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

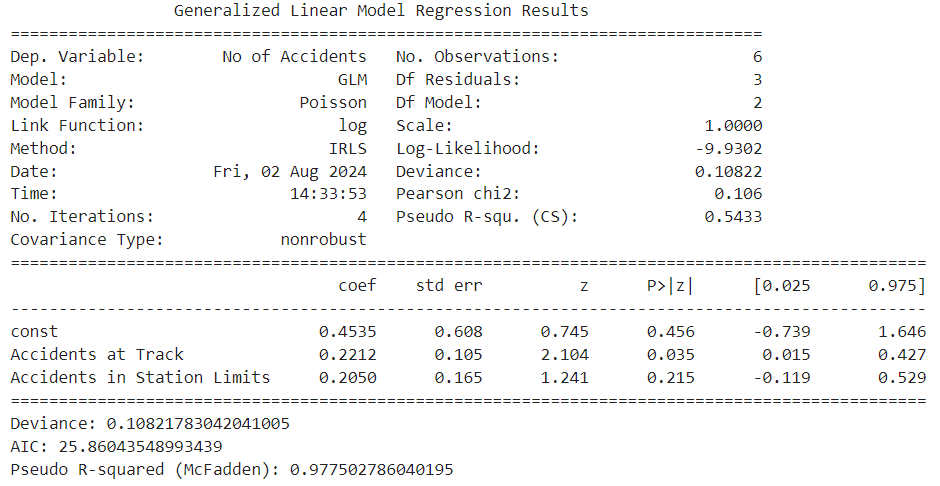


**Description:** The second graph illustrates the contributions of each predictor variable (Derailments, Collisions, and Collisions at Level Crossings) to the number of accidents. Each variable's contribution is plotted as a line showing its impact across the years.

**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.

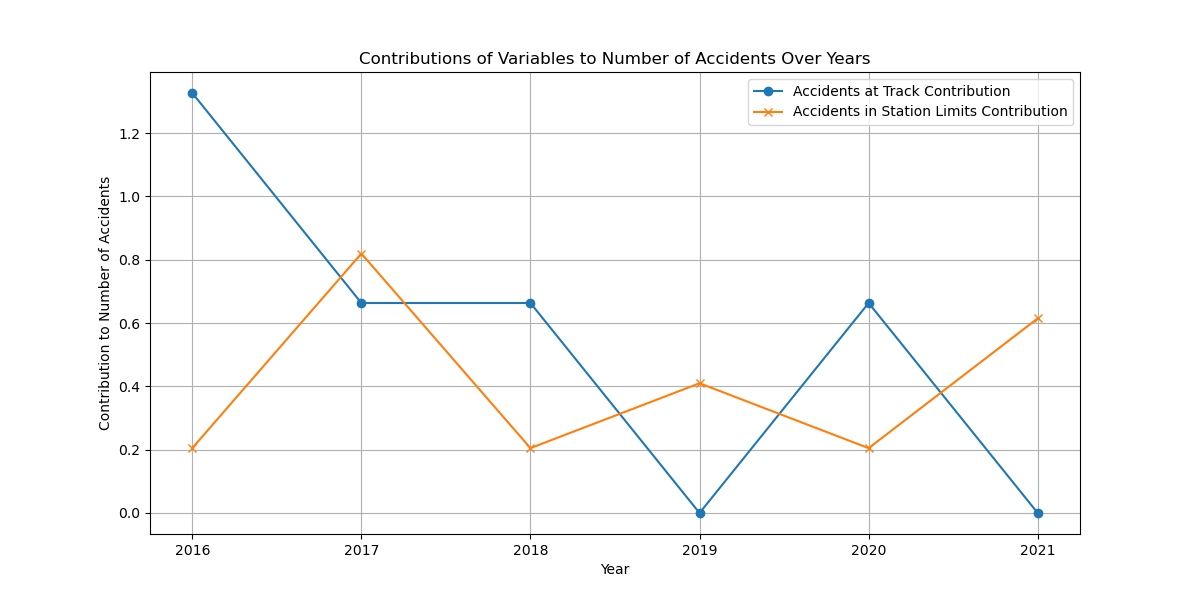
## AIC

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



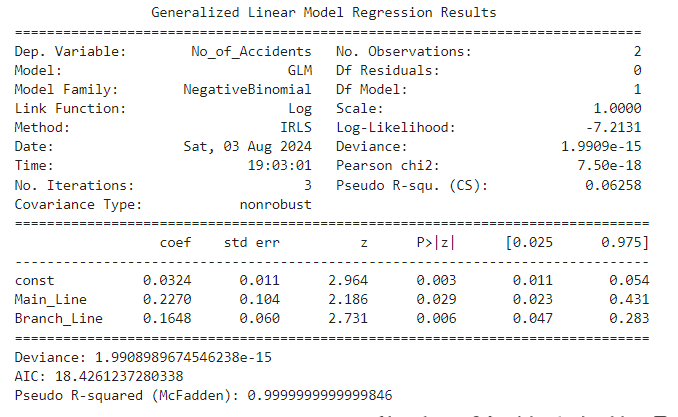
## Accidents at Track:

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:

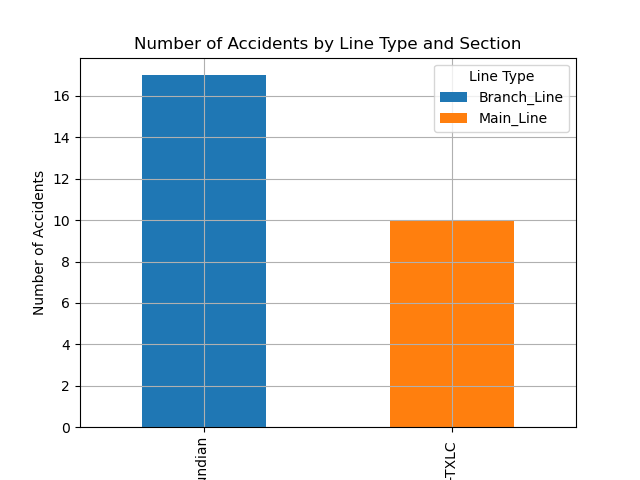
0.1648 with a standard error of 0.06. This proved to be statistically significant as well because the p-value came out to be 0.006, so the interpretation goes that a one-unit increase in accidents for the branch line is associated with a 17.9% increase in the total number of accidents since ???? 0.1648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



## Main Line:

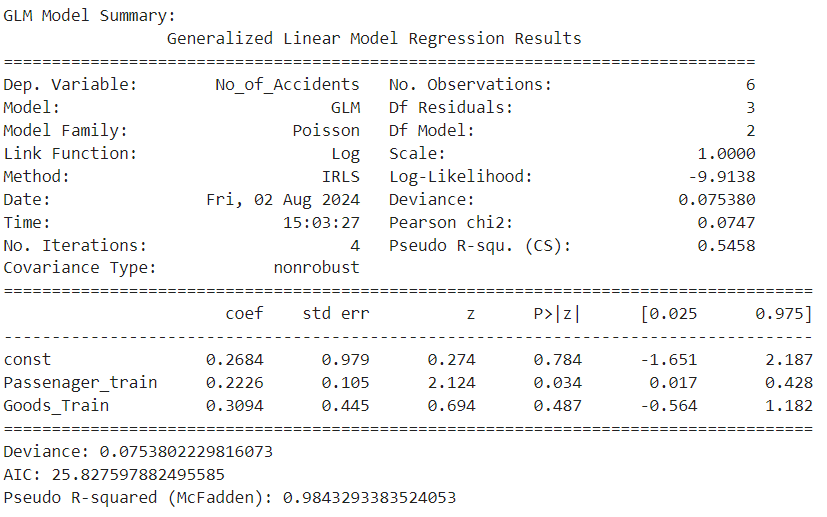
The main line has considerably more no. of accidents that are 10 Accidents in section PSC-TXLC and hence it reflects a high degree of concentration of accidents on the main line While in section Jhand-Kundian it has relatively low contribution of 0 Accidents

## 

## Branch Line.

While the branch line contribution is more in Jhand-Kundian section with number of accidents 17 which is comparatively high then main branch and having no contribution in accidents in section PSC-TXLC section.

# 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.

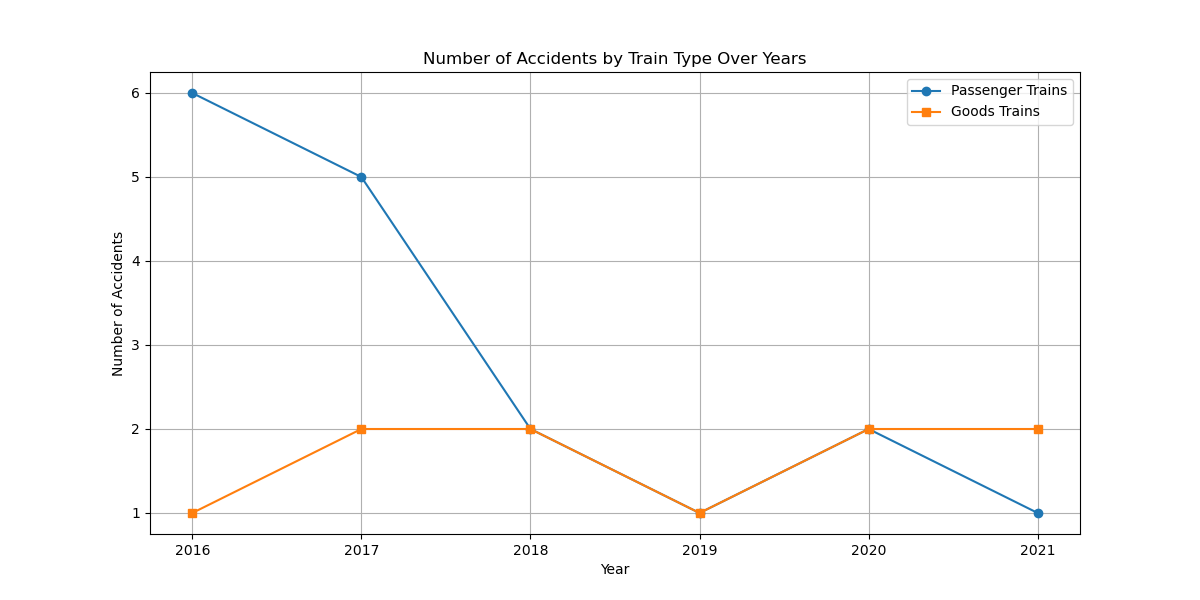
## AIC:

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

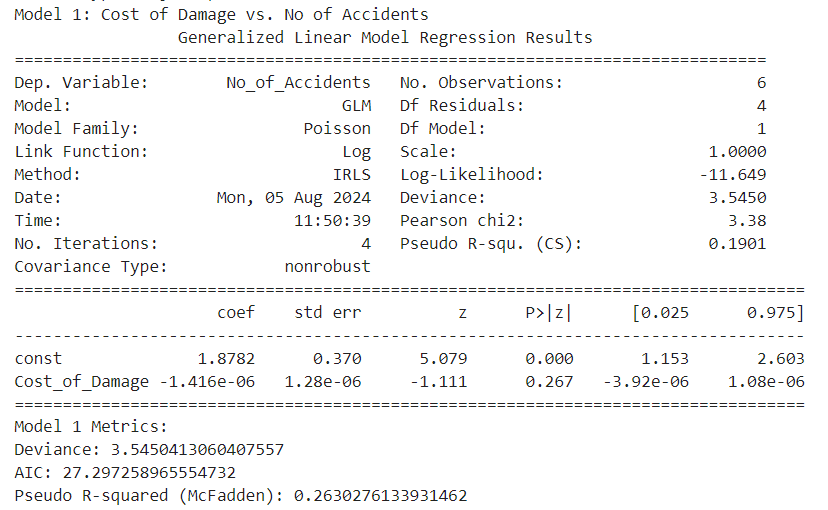
he trend in the number of accidents related to passenger trains generally shows a decreasing trend over the years, with a visible drop from 6 in 2016 to 1 in 2021. This trend represents a major reduction in mishaps related to passenger trains.

## 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 and damage/no damage to pr



## Model : Cost of Damage vs. Number of Accidents

The model below estimates the relationship between cost of damage and number of accidents using Poisson regression.

## Generalized Linear Model Regression Results:

## Number of Accidents

Deviance: 3.545

Pseudo R-squared (McFadden): 0.1901

## Coefficients:

Constant: 1.8782 (p < 0.001, significant)

## Damage Cost:-1.416e-06 (p = 0.267, not-significant)

## Metrics:

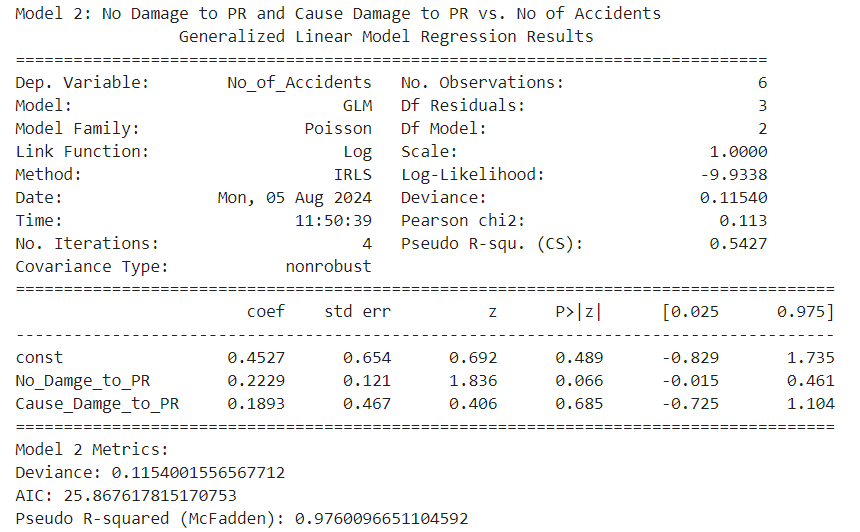
Deviance:3.545

AIC:27.297

pseudo R-squared: McFadden = 0.263

## Interpretation:

The coefficient for cost of damage is negative, suggesting some sort of inverse relationship between cost of damage and the number of accidents. However, it is not statistically significant, with a p-value of 0.267, which means that cost of damage does not turn out as a strong predictor of the number of accidents. The pseudo R-squared value of 0.263 indicates modest explanatory power of the model.



## Model : No Damage to PR and Cause Damage to PR vs. Number of Accidents

The second model investigates how 'No Damage to PR' and 'Cause Damage to PR' affect the number of accidents with a Poisson regression.

## Dependent Variable: Number of Accidents

Deviance: 0.115

Pearson Chi-Square: 0.113

## Coefficients:

Constant (Intercept): 0.4527 (p = 0.489, not significant)

## No Damage to PR: 0.2229 (p = 0.066, marginally significant)

## Cause Damage to PR: 0.1893 (p = 0.685, not significant)

## Model Metrics:

Deviance: 0.115

AIC: 25.868

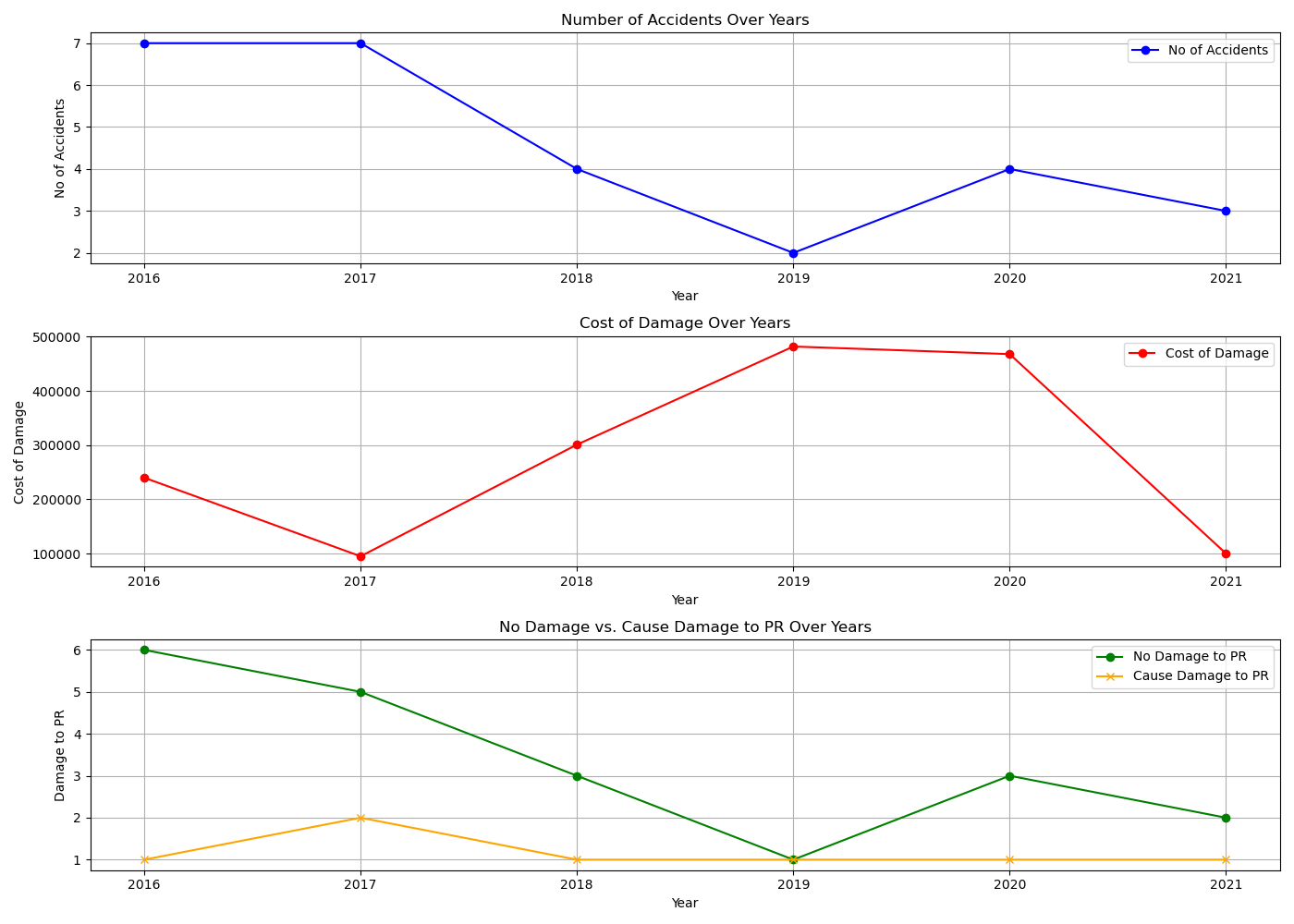
Pseudo R-squared (McFadden): 0.976

## interpretation:

The coefficient for 'No Damage to PR' is positive and marginally significant, so it may indicate a positive relationship with the number of accidents when there is no damage to PR. The coefficient for 'Cause Damage to PR' is also positive but not statistically significant, indicating a weak relationship to the number of accidents. The pseudo R-squared 0.976 indicates that this model explains a large proportion of variance in the number of accidents, hence it will be quite strong.

## 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.

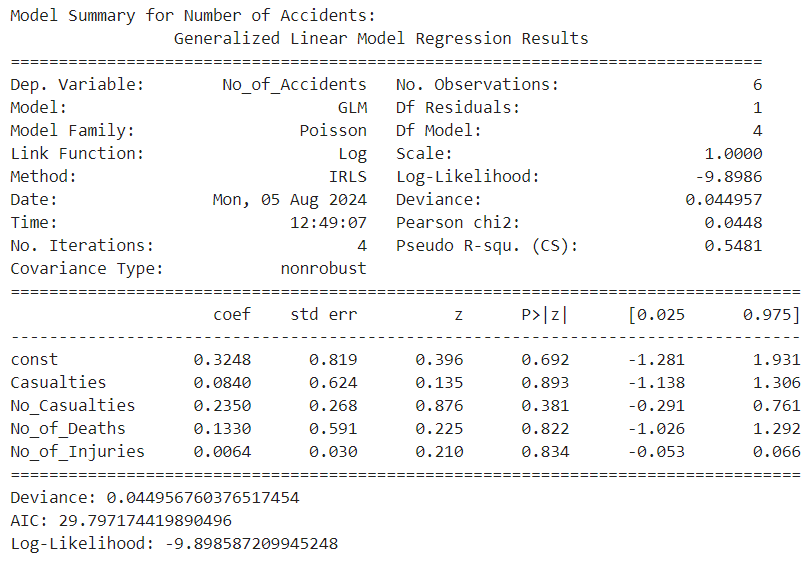


The first plot depicts the number of accidents per year. From this data, one can perceive the fact that the number of accidents is variable with no trend in it. This variability of accidents seems to vary from 2-7 every year, proving that some other external factors might be responsible for accidents,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



## Model Summary

The result is interpreted using a Generalized Linear Model with the accidents as dependent variables and casualties-no casualties, death, and injuries as predictor variables.

## Intercept

This is the baseline or average estimate of the number of accidents when all predictor variables are equal to zero. With all of the predictor variables at zero, the estimate for the number of accidents is around 0.32, which is not statistically significant (p = 0.692), so the constant alone does not significantly predict the number of accidents.

## Casualties:

The coefficient of casualties is 0.0840, which is not significant at p = 0.893. This means that casualties do not impact the number of accidents to any appreciable degree within this model.

## No\_Casualties:

The coefficient is 0.2350 for no casualties, but again, this result is not statistically significant with p = 0.381. While this variable returns a positive sign of association with the number of accidents, this effect does not stand significant in the context of the model.

## No\_of\_Deaths:

The coefficient on deaths is 0.1330, not statistically significant (p = 0.822); it means that the number of deaths does not substantially affect the number of accidents in this model.

## No\_of\_Injuries:

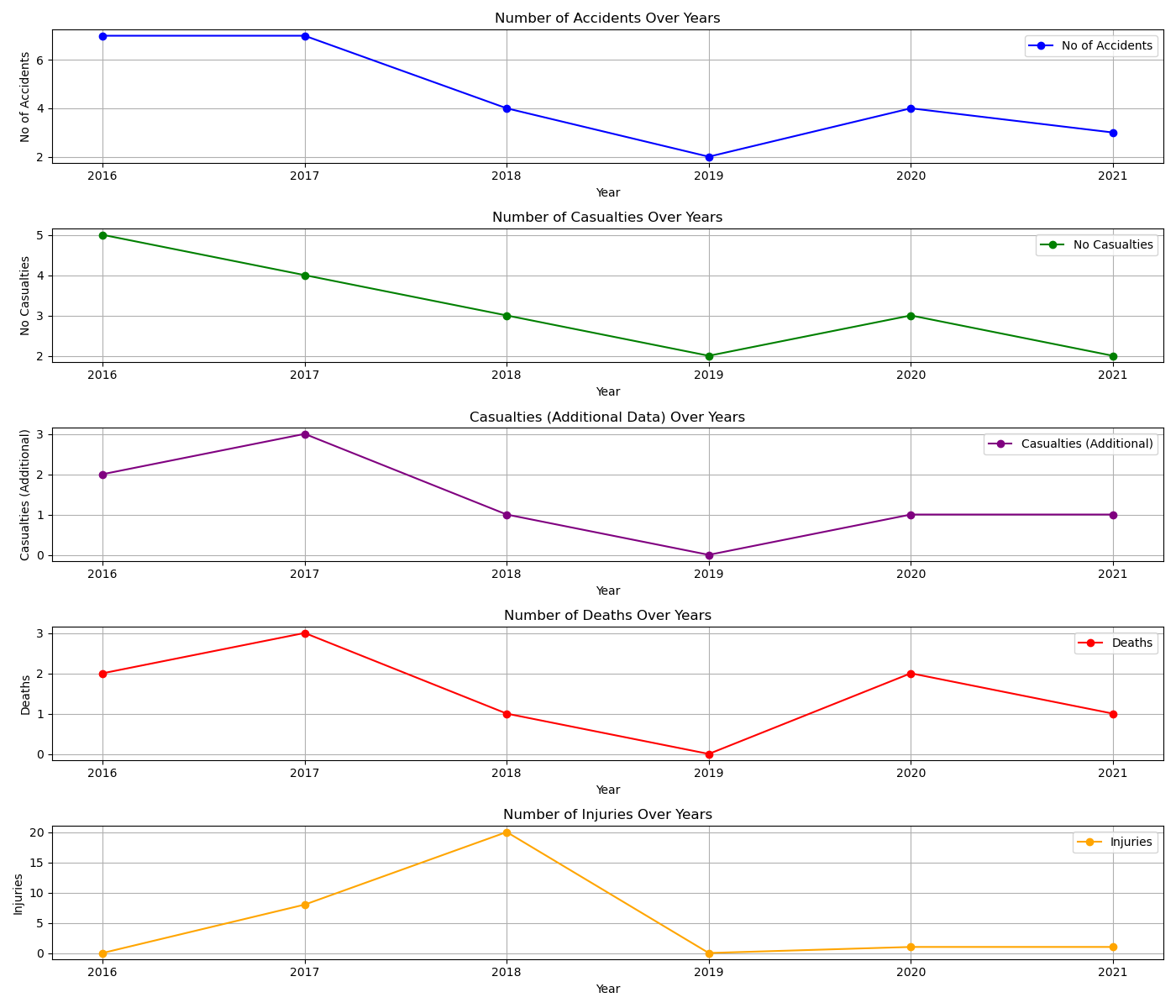
with a coefficient of 0.0064, very near to zero, it is not significant with p = 0.834. Thus, the number of injuries has an effect that is negligible, not significant, on the number of accidents.

## Model Metrics

The deviance is very low — 0.045 — thus showing a good fit of the model to the data, with the predictions very close to the actual values. The Pseudo R-squared equals 0.5481, so it explains a medium portion of the variance in accidents.

In this model, none of the predictors have statistically significant effects on the number of accidents. The deviance and Pseudo R-squared provide a good fit of the model to the data, but the individual predictors are not significant in accounting for the number of accidents.

## Graph Analysis



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

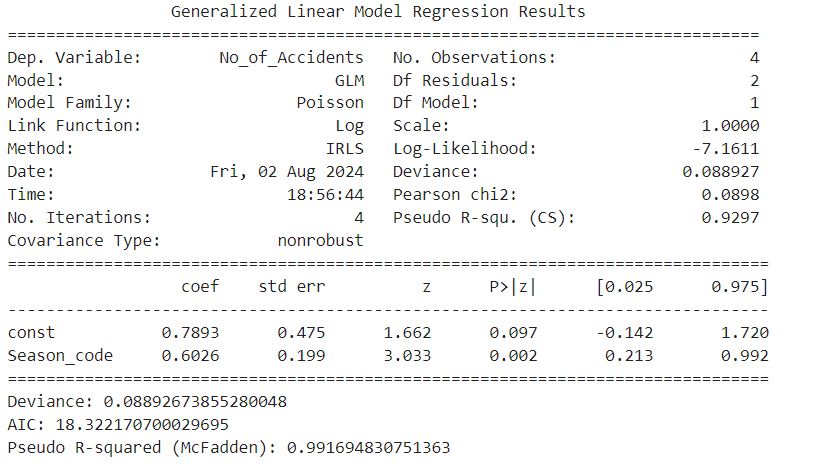
Number of Casualties Over Years: Casualties change greatly, peaking in 2017 and 2018, and then falling.

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:

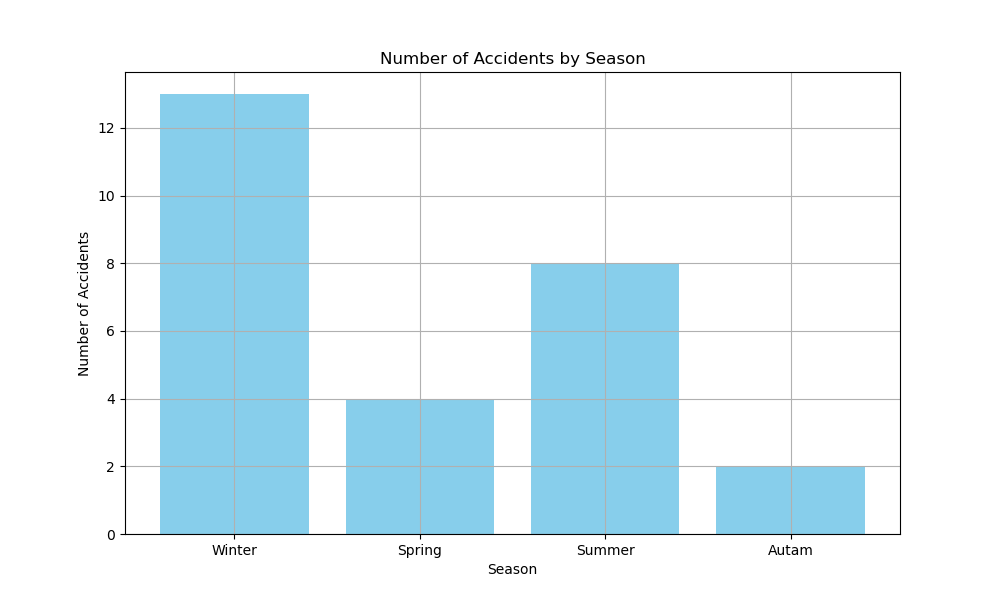
The coefficient for the intercept is 0.7893, with standard error 0.475; this is the expected log count of accidents when all other variables are zero. Its p-value of 0.097 suggests that it is not statistically significant at the 5% level but might be of further interest.

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.

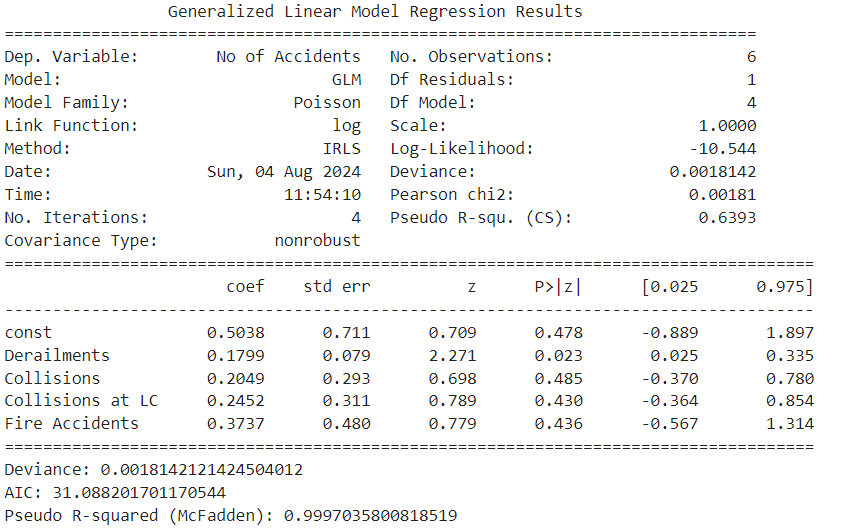
## Autumn:

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 and Interpretation

The GLM analysis indicates how different types of accidents add up to the total number of accidents in Quetta. Here are the main takeaways related to the model summary:

## Model Fit Metrics:

## Deviance:

The deviance value of about 0.0018 portrays good fitness between the model and the observed data.

## AIC

The value of Akaike Information Criterion of 31.088 gives the relative quality of the model; the lower, the better.

## Pseudo R-squared (McFadden)

The pseudo R-squared value of about 0.9997 shows excellent fit, indicating that the model explains almost all of the variability in the response variable.

## Coefficients:

Derailments For this coefficient, the coefficient value is 0.1799, and with a p-value significant at 0.023, there is a positive and statistically significant relationship with accidents.

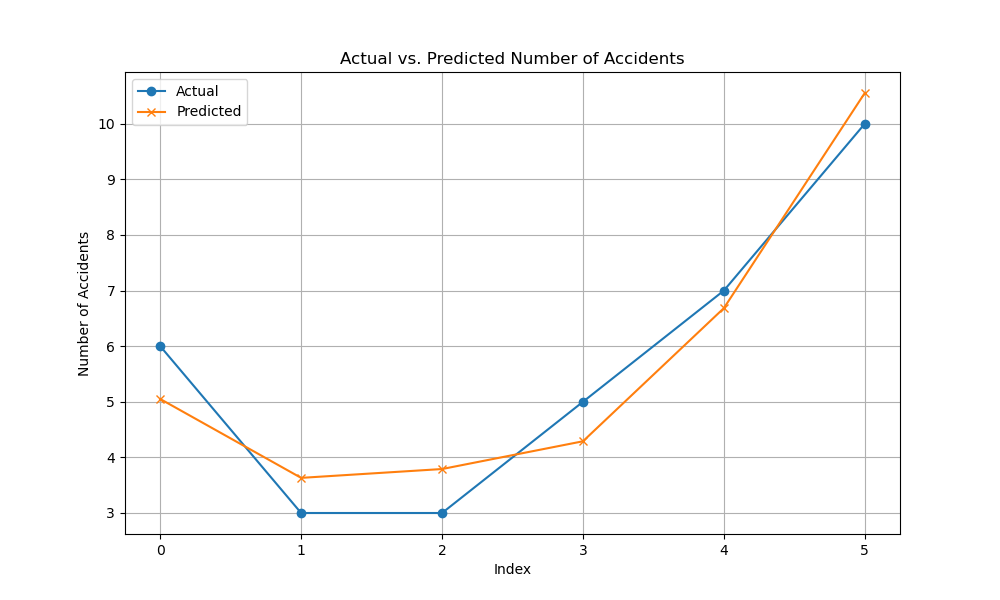
Collisions: The coefficient is 0.2049, but with a non-significant p-value of 0.485, it interprets a positive relationship but is not significant in terms of statistics.

Collisions at LC: It has a coefficient of 0.2452, with a non-significant p-value of 0.430, indicating a positive but not statistically significant relationship.

Fire Accidents: The coefficient here is 0.3737, having a non-significant p-value of 0.436, suggesting a positive but not significant relationship in terms of statistics.

## 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.

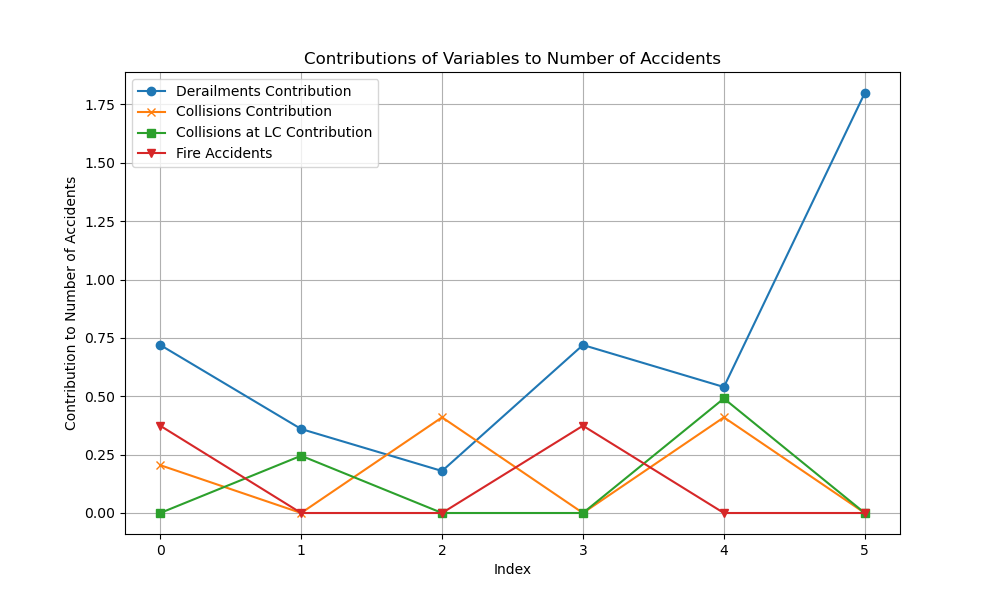
## Trend:

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



Contributions of Variables to Number of Accidents

This graph plots individual contributions of each AccidentType—Derailments, Collisions, Collisions at Level Crossings, and Fire Accidents—multiplied by the coefficient values for the GLM model. A more detailed breakdown is carried out as follows:

## Derailment Contribution

Pattern:

The plot for derailments shows a generally steady contribution over the years, with some noticeable increases in certain years. This pattern is consistent with the significant coefficient for derailments in the model.

Impact:

The constant contribution of derailment indicates that it is actually the leading cause of the total accidents. The high positive coefficient of 0.1799 indicates that as the variable of derailments goes up, that of the total accidents tends to increase, showing just how critical it is.

## Collisions Contribution

Pattern:

The contribution of collisions indicates a much more variable pattern. It doesn't seem to follow any trend and is less consistent when compared with derailments.

Impact:

The coefficient here is positive at 0.2049, but the p-value is not significant, so collisions do not strongly influence or consistently impact the total accidents. That means there is variability in what is returned by collisions. This contribution variability returned by collisions suggests that while it contributes to the number of accidents, compared to derailments, its effect is much less stable.

## Collisions at Level Crossings LC Contribution

Pattern: The contribution from collisions at LC is variable, with no clear trend. It impacts in some years but not consistently for all the years.

Impact: The coefficient would thus be positive but non-significant at 0.2452, which would imply that while collisions at LC take their share in contributing to the overall accidents, this contribution is not statistically robust or consistent across the dataset.

## Contribution of Fire Accidents

Pattern:

The contribution coming from fire accidents is erratic; some years contribute high, while others show a minimal effect.

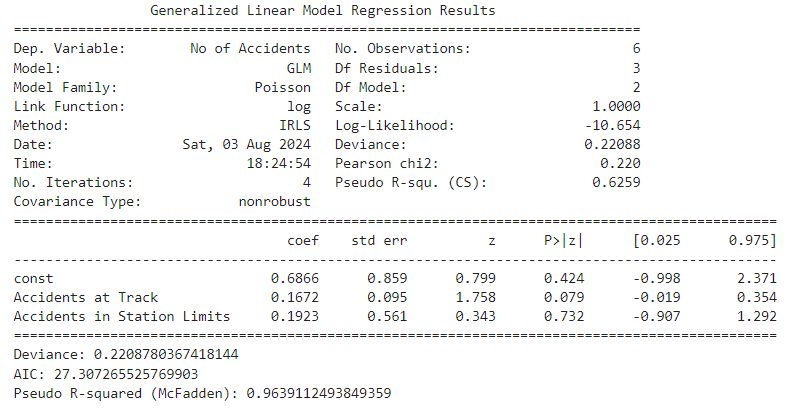
Impact:

The coefficient is positive - 0.3737, indicating that fire accidents do affect the total number of accidents, but similar to collisions at LC, the effect is not significant. The unevenness of the contribution mirrors the sporadic nature of fire accidents in the data.

## Summary

The second graph indicates that, over the years, derailments have been the most solid and highest contributor to the total accidents. This uniformity corresponds to the model results that clearly indicate that a reduction in derailment accidents means a large reduction in accident numbers overall. Other accident types, even though they do contribute, are less consistent and statistically significant, indicating that their impact is less predictable and therefore less critical as compared to derailments.

# 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.

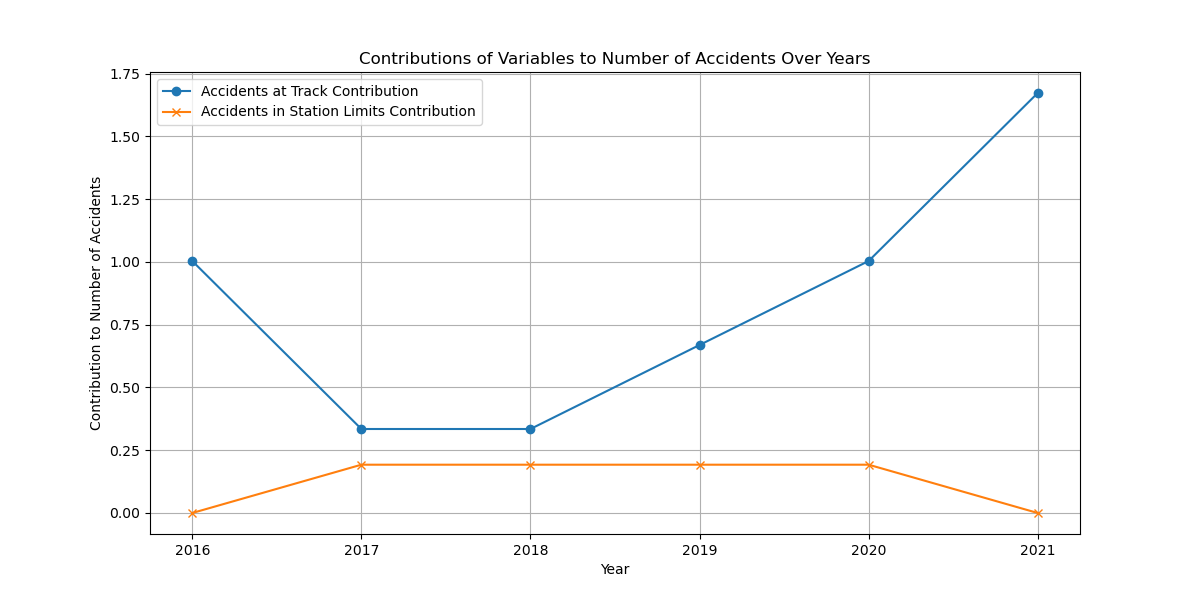
## 27.3073 is the AIC,

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 

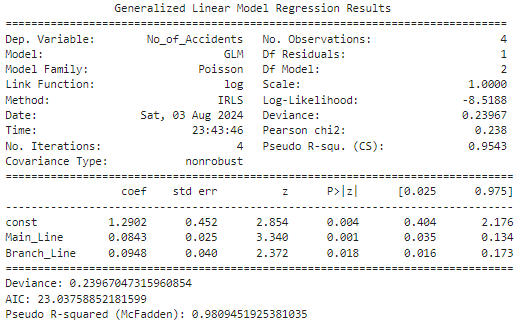
This is a graph that shows the contributions from the predictor variables of Accidents at Track and Accidents in Station Limits through the years to the number of accidents.

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:

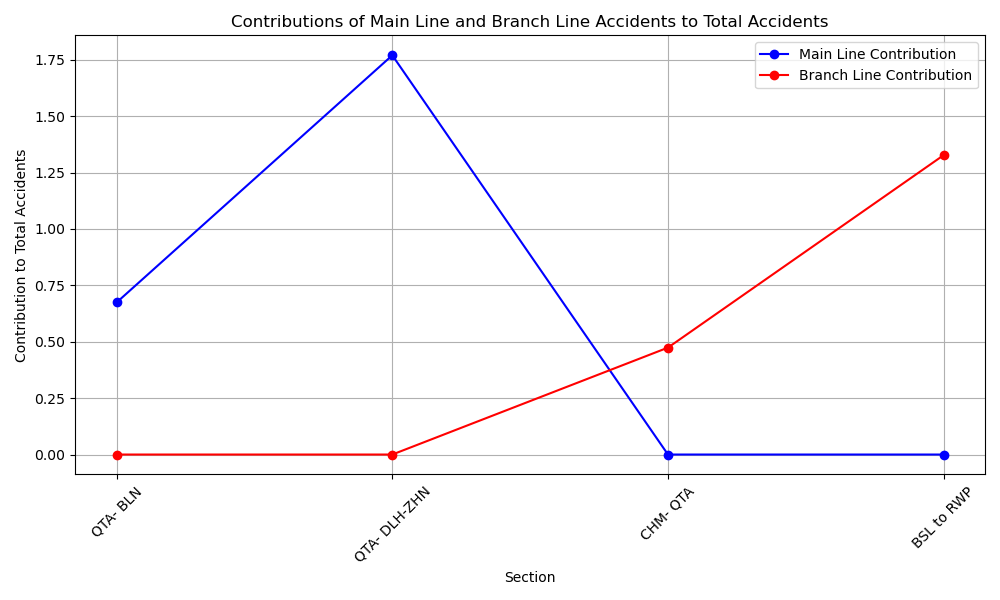
Akaike information criterion is 23.0376. AIC is a measure of how well a model fits the data; AIC values close to zero are indicative of very good fits, and smaller is always better.

## 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:

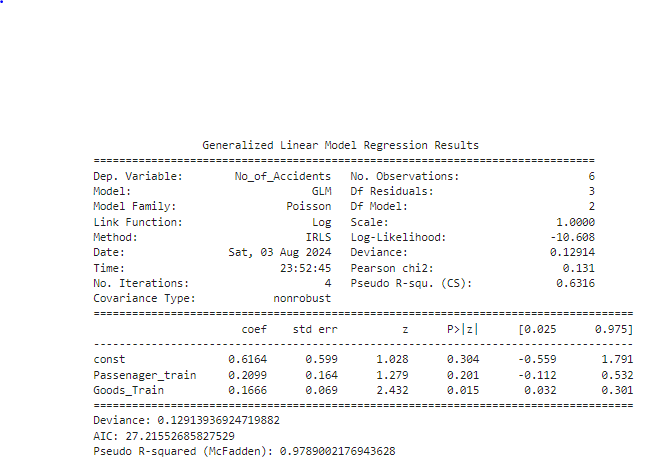
This contribution of the Main Line accidents is to the total number of accidents. Every point on the blue line is related to a section, and the contributions are computed by considering the number of Main Line accidents from the concerned section and the estimated coefficient from the Poisson model. 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:

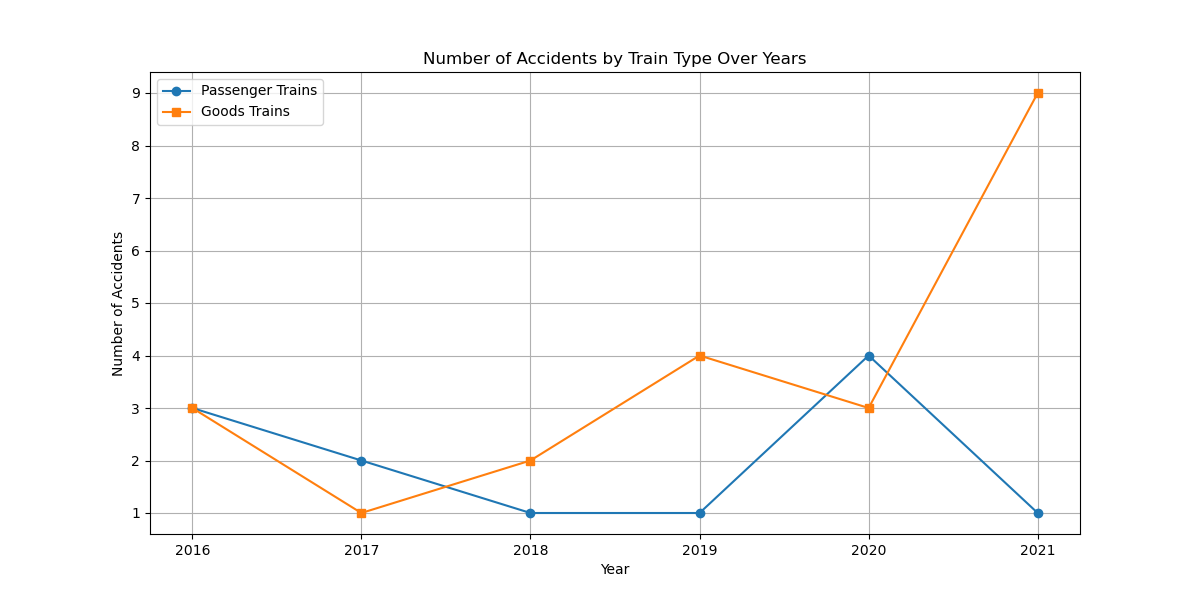
Akaike Information Criterion: The AIC value is 27.2155. It is used for model comparison. The model with a lower AIC value will have a better fit.

## 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:

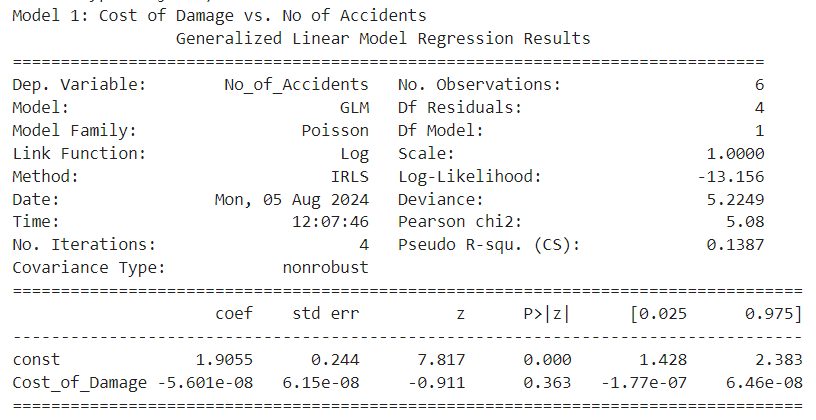
The blue line describes the contribution of Passenger Train accidents to the total number of accidents in the graph. Each point on the line corresponds to a different year and plots the contribution calculated by the number of accidents concerning the variable Passenger Train within that year and the 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.

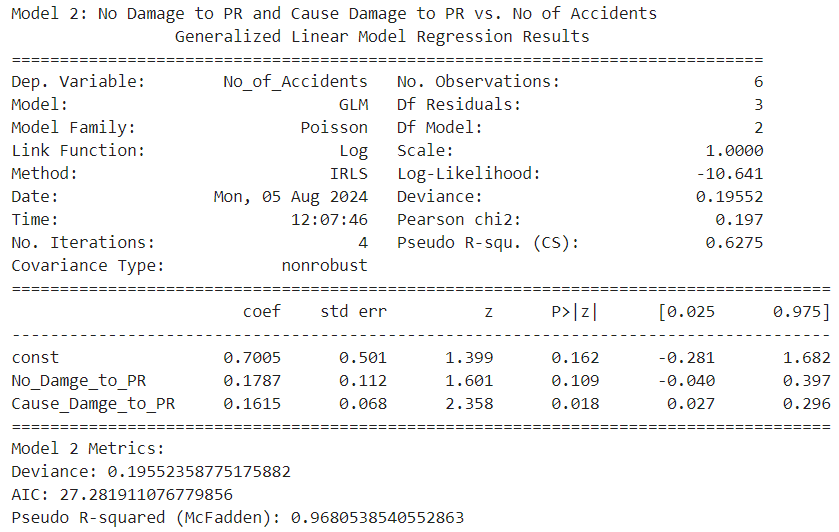
# Analysis of cost of damage/damage to pr/no damage to pr



## Model : Cost of Damage versus Number of Accidents

This Model explores the relationship between the cost of damage and the number of accidents in the dependent variable. A Poisson regression model will be applied. The dependent variables are the number of accidents, and there are six observations. The model is simple with 1 predictor; hence, the residuals have 4 degrees of freedom and the model also has 1 degree of freedom. It is in the form for Poisson family and the IRLS method is used in its estimation while adopting a log link function. In this case, the log-likelihood is -13.156. The estimated deviance for the model is 5.2249, which in some way presents a shape of the good fit of the model. The value of Pearson chi-square is 5.08, which shows a reasonable fit. The model took four iterations to converge. The value of the pseudo R-squared is 0.146, which says that the model explains approximately 14.6% of the deviation of the number of accidents.

The model's coefficients show that the more robust indicator is the constant, whose value is 1.9055, at a very high level of significance: p < 0.001. In other words, when the cost of damage equals zero, the average log number of accidents will be 1.9055. Also, the coefficient of the cost of damage—5.601e-08—proves to be insignificant, statistically, at p = 0.363. This gives a hint that the cost of damage is not strongly related to the number of accidents. AIC of the model is 30.311 (AIC is an Akaike Information Criterion, which is used in model comparison. Lower values of AIC means the best model.)



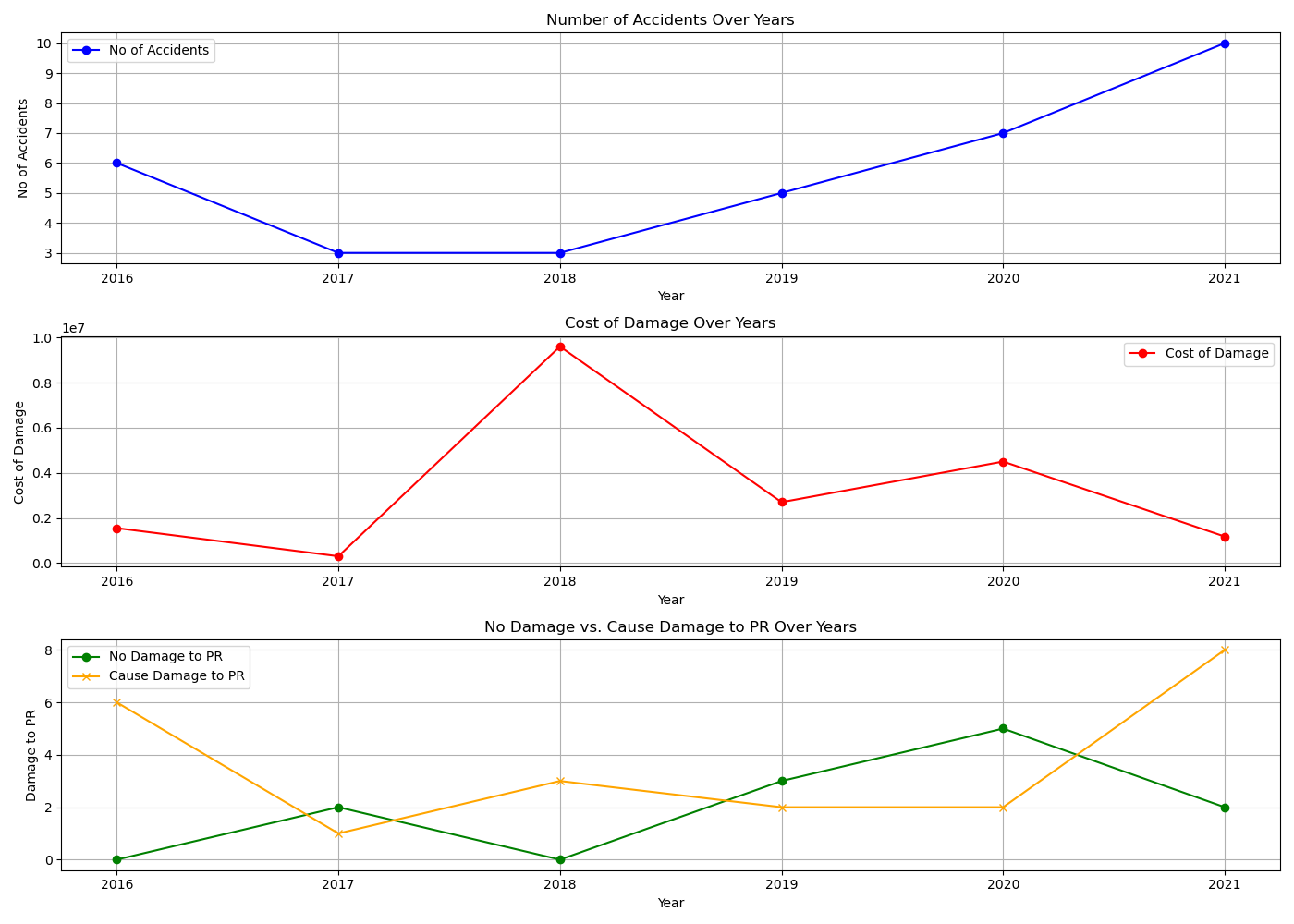
## Model : No Damage to PRO and Cause Damage to PRO vs. Number of Accidents

The following model looks into the impact of 'no damage to PR' and 'cause damage to PR' related to the number of accidents in a Poisson regression model.

This, too, holds the number of accidents as the dependent variable. It has six observations, that model has three degrees of freedom for residuals, and its degrees of freedom for the model are two, so this is a slightly more complex model with two predictors. The model comes from the Poisson family with the log as the link function and is estimated using the IRLS method. The log-likelihood of the model is -10.641, and deviance is 0.1955; both of them are very low in value, which can infer a good fit. The value of Pearson's chi-square is 0.197, and it indicated an excellent fit; it took the model four iterations to converge. The pseudo R-squared value is 0.968. This implies that about 96.8% of the variation in the number of accidents can be explained by the model.

From the coefficients of the model, we note that the intercept is 0.7005, which is not at all significant at p = 0.162. This clearly indicates that coefficient 0.1787 for 'No Damage to PR' is just marginally significant, having p = 0.109, and there may be the presence of a positive relation between no damage to PR and the number of accidents. The coefficient is 0.1615 for 'Cause Damage to PR', significantly related at p = 0.018, showing a positive relationship between causing damage to PR and the number of accidents. The AIC value for the model is 27.282, which can be compared with the other models.

## 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

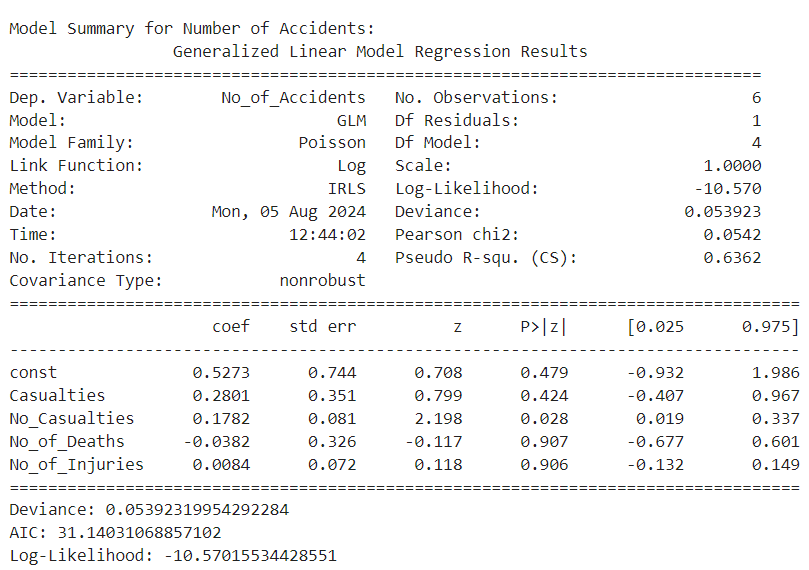
The cost of damage over the years is represented through the second subplot of the graph. In 2018, the cost of damage is the highest, accounting for 9,600,000. Other years portray varying costs that sometimes considerably change, hence proving that some of the years were financially impacted by accidents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



## Model Summary.

This is the interpretation of the Generalized Linear Model with number of accidents as dependent variables and casualties, no casualties, deaths, and injuries as the predictors:

## Intercept

This is the baseline estimate of the number of accidents in the absence of any predictor variables and comes to approximately 0.53, but this is not statistically significant (p = 0.479), indicating that the intercept alone does not provide any significant estimate of the number of accidents.

## Casualties:

The coefficient for the number of casualties is 0.2801, so it is positively related to the number of accidents, although the result is not statistically significant (p = 0.424). This means that changes in the number of casualties do not significantly affect the number of accidents in this model.

## No\_Casualties:

The coefficient, 0.1782, is statistically significant (p = 0.028). This means that if there is an increase in the cases of no casualties, it will raise the number of accidents. This relation is significant, and therefore, this factor is meaningful for explaining the number of accidents.

## No\_of\_Deaths:

The coefficient for death is negative but is not significant statistically as the p-value is 0.907. This interprets that the change in the number of deaths does not affect the number of accidents significantly.

## No\_of\_Injuries:

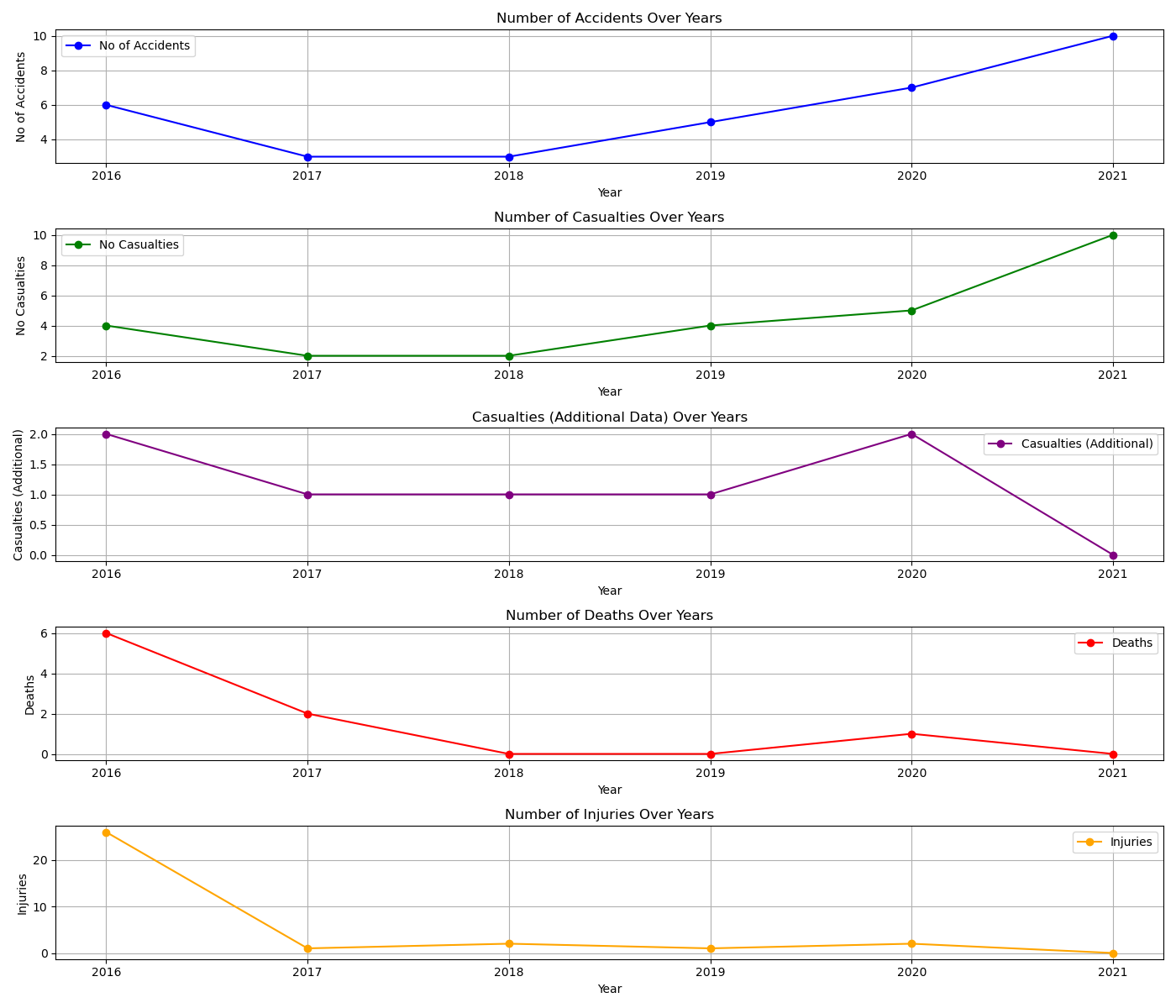
The coefficient for injuries is very close to zero and, in addition, is not significant statistically as the p-value is 0.906. This interprets that the number of injuries does not significantly affect the number of accidents.

## Model metrics:

By this, the model deviance is very low at 0.054, and the Pseudo R-squared equals 0.6362, indicating that the model explains a substantial portion of variance in the number of accidents. The low deviance means that the model's predictions are very close to the actual data; this, again, also indicates a good fit from the high Pseudo R-squared measure.

The overall large positive effect of no casualties on the number of accidents suggests that it may be an important factor. Of the remaining predictors, none of the effects—in particular, those for casualties, deaths, and injuries—come close to being significant in this model.

## 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

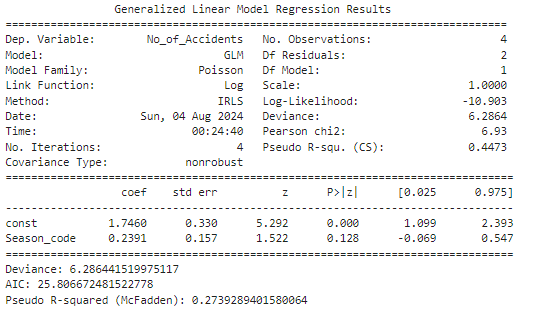
This plot shows a decreasing trend of the number of injuries over the years, with the count peaking in 2016 and almost nil in the previous years.

## 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:

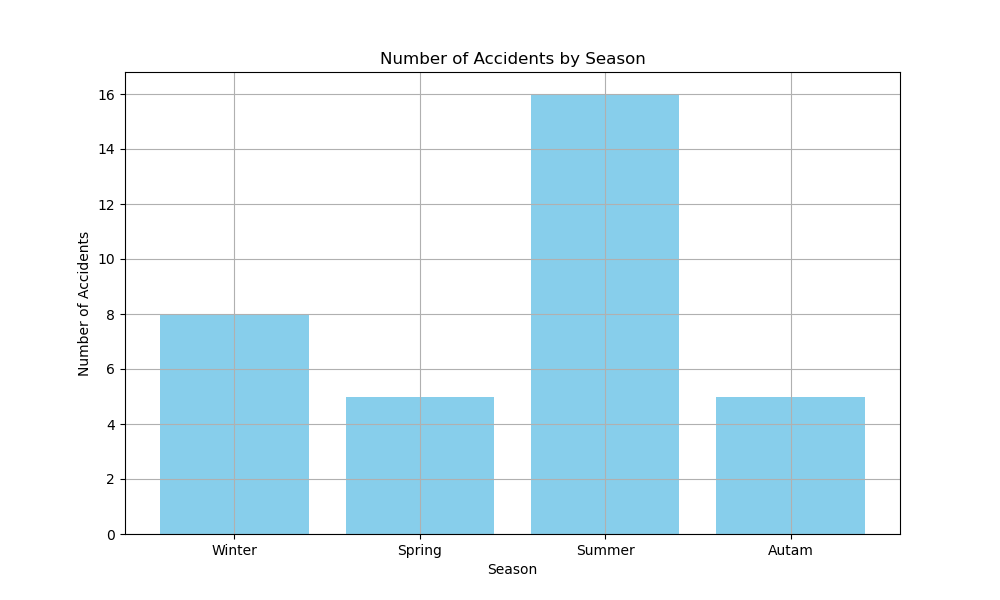
25.81, The lower the better the fit; used for model comparison.

## 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:

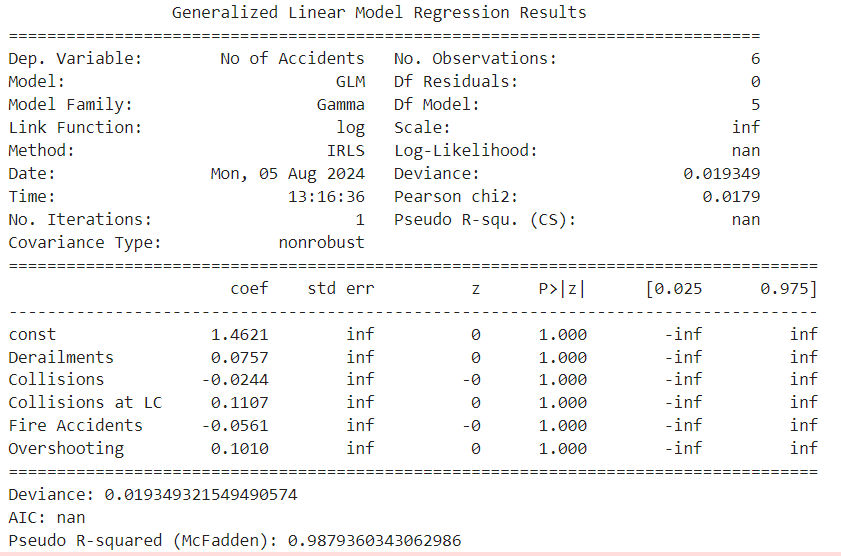
Summer tops with the highest number of accidents (16).

Winter is next, with 8 accidents.

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

# Analysis of Rawalpindi Accidents data

# Accident type Analysis



## Model Summary Interpretation:

Generalized Linear Model (GLM) Results:

Deviance: 0.0193

Pseudo R-squared (McFadden): 0.988 (Note: this is too high, model fit might have problem)

## Coefficients and Their Interpretation:

## Intercept (const):

The coefficient for the intercept is 1.4621 but its standard.

## Derailments:

The coefficient of the variable is 0.0757, and the standard error is infinite. It can be seen that the variable has an almost zero gradient.

## Collisions:

The variable has a gradient of −0.0244, but the standard error is infinite. This means the variable does not have a significant effect on model accident counts.

## Col LCs c:

The coefficient for the variable is 0.1107, and its standard error is infinite, so not significant.

## Fire Accidents:

The coefficient is -0.0561 and again has an infinite standard error; thus, it is statistically insignificant.

## Overshooting:

The coefficient is 0.1010, but again the standard error is infinite, thus another statistically insignificant variable.

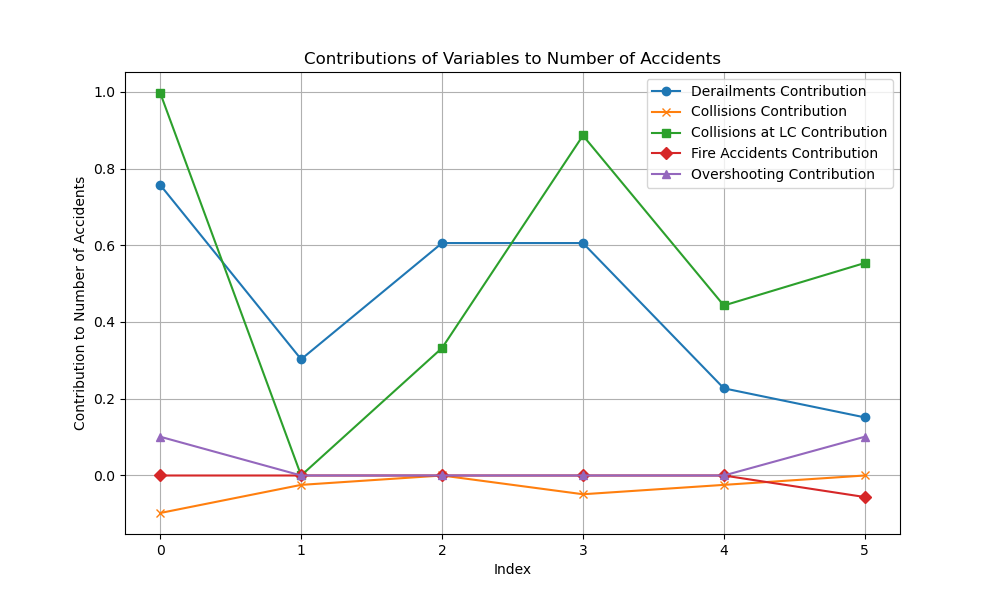
## Notes on the Model Fit:

Deviance: The very small value of deviance is just 0.0193, indicating the model explains variation in data very well, but infinite standard errors for coefficients may indicate some problem in model fit.

AIC: It could not be calculated because of the problems explained with the model fitting.

Pseudo R-squared: It got 0.988, which is extraordinarily high, probably due to the issues faced in model fitting, in general, or maybe due to overfitting with a dataset this small.

## Graph Interpretation:



The plotted graph, which was plotted in terms of each variable's contribution to accident occurrence, shows the way in which the predictors drive the accident:

## Derailments Contribution:

This train should show what the contributions of the derailments are to the populations of the accidents. However, because the issues run on the models, this contribution to the population would not be real.

## Collisions Contribution:

It indicates the impact of the collisions' effect on the accidents' population. The line should show whether there are occurrences of an increase in the number of the collisions that is associated with more accidents.

## LC Coll Contribution:

This shows the contribution of collisions at LC-level crossings. It indicates the way collisions at the LC-level crossing contribute to the number of accidents.

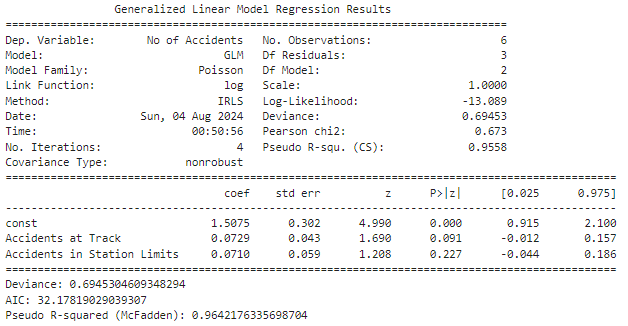
## The Contribution of Flaming Accidents:

This depicts the contribution made by flaming accidents toward the number of accidents. If the line does not fluctuate significantly throughout, it means that flaming accidents do not contribute significantly to the number of accidents.

## Overshooting Contribution:

This line of location is to further illustrate how overshooting contributes to the numbers of accidents. The relative impact with other variables will be shown.

# Analysis of accidents location wise



## Model Fit and Summary:

The Generalized Linear Model with a Poisson family and log link function was used to estimate the effect of "Accidents at Track" and "Accidents in Station Limits" on the total number of accidents.

## Deviance:

By this measure, the deviance of the model is 0.6945, so by this measure it is a good fit since it tells how well the values predicted by the model match with the observed values.

## AIC:

The Akaike Information Criterion is 32.1782 and one may use it for model comparison purposes; the smaller AIC, the better fit.

## Pseudo R-squared:

with a McFadden value of 0.9642, it is very large and thus indicates a fairly large proportion of the variation in number of accidents explained by the model, although it is not a formal measure of goodness-of-fit for a Poisson model.

## Coefficients:

## Constant:

The coefficient of the intercept is 1.5075. This means that it is the baseline level of accidents when both predictors are zero.

## Accidents at Track:

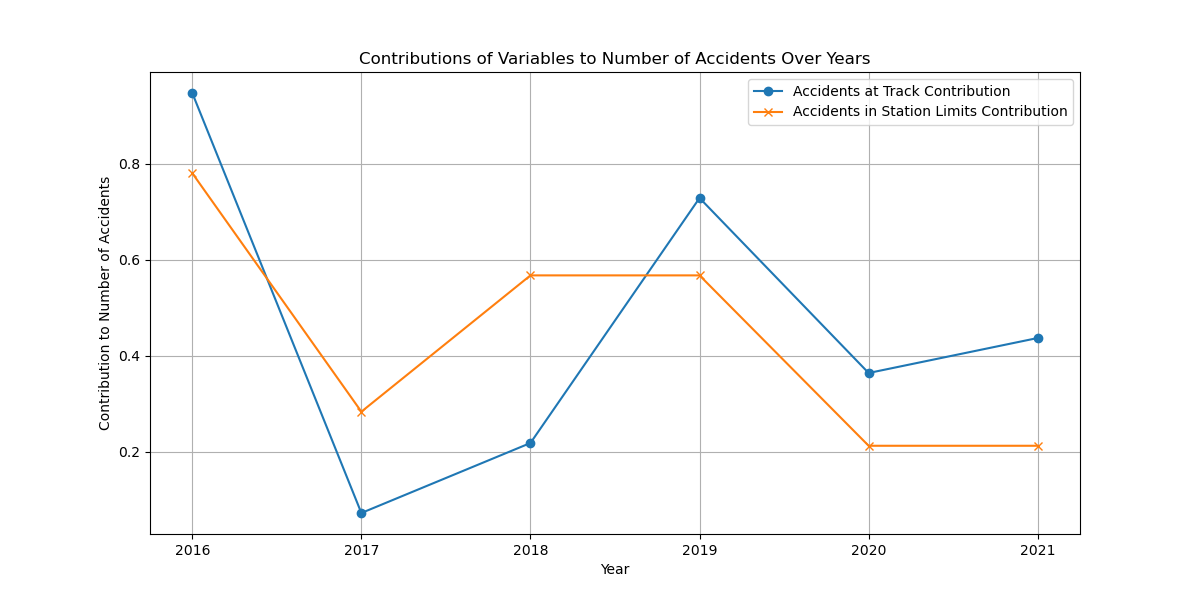
With a coefficient of 0.0729, it shows a positive effect on the number of accidents, though not statistically significant with p = 0.091. This might be interpreted to mean that though this factor tends to increase accidents, its influence is weak and hence not significant.

## Accidents in Station Limits:

This variable also shows a positive effect but is statistically insignificant, with a coefficient 0.0710 (p = 0.227). Similar to the case of "Accidents at Track," this implies that there is a weak influence on the number of accidents.

## Graph Analysis

Graph Analysis of Variable Contributions:



## Accidents at Track:

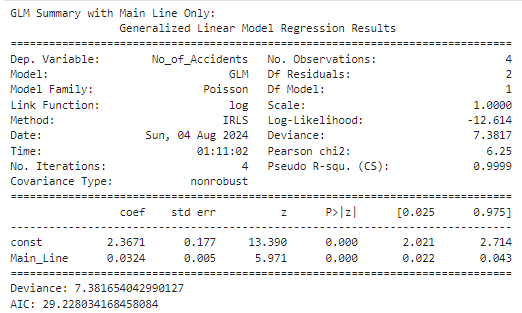
The graph shows a growing trend for "Accidents at Track" in their contribution to the number of accidents over the years, with peaks corresponding to years of higher total accidents. This agrees with the positive coefficient seen in the GLM results. A variation in the contribution from "Accidents at Track" appears to be more variable and responsive to changes in accident numbers.

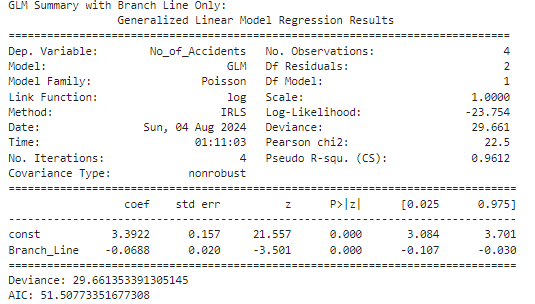
## Accidents in Station Limits:

The graph indicates that "Accidents in Station Limits" has always contributed a relatively stable share, with less fluctuation across the years compared to "Accidents at Track." This is reflected in its coefficient, positive but not significant. This component adds to the total number of accidents, although its influence is less strong and less time-variable.

The graph supports the GLM findings that "Accidents at Track" has more variable and high influence on accidents over time, in comparison with "Accidents in Station Limits," whose influence has higher consistency but is weaker. Both variables contribute positively to the number of accidents; neither of them shows a strong or statistically significant effect in the model. This very much suggests that, although this means that these factors do influence accident numbers, other variables not included in the model might play a more key role.

# Accident Analysis by Line Type





## Model Summary

## GLM with Main Line Only

The Generalized Linear Model GLM fitted to the variable "Main Line" accommodatingly reveals the following:

## Coefficient for Main Line:

0.0324

## Standard Error:

0.005

## Z-Value:

5.971

## P-Value:

0.000

The occurrence of a positive coefficient indicates a rise in accident frequency with a rise in the value of "Main Line". The impact is significant at statistical levels, thereby implying that a consistent positive relationship can be established with the occurrence of an accident and "Main Line".

## The deviance

7.3817

## AIC

29.2280

show a very good fit to the data. In fact, the exceptionally high pseudo-r value of

0.9999  is indicative of this model explaining almost all the variation in the number of accidents; however, this result may be sensitive to the small sample size.

## GLM with Only Branch Line

The GLM model with one variable only, "Branch Line", has

## Branch Line Coefficient:

0.0688

The negative coefficient for "Branch Line" means that increases in the "Branch Line" variable are associated with a decrease in accidents. This result is statistically significant, indicating that one may rely on this negative effect.

The model deviance is

29.6614

## AIC

51.5077

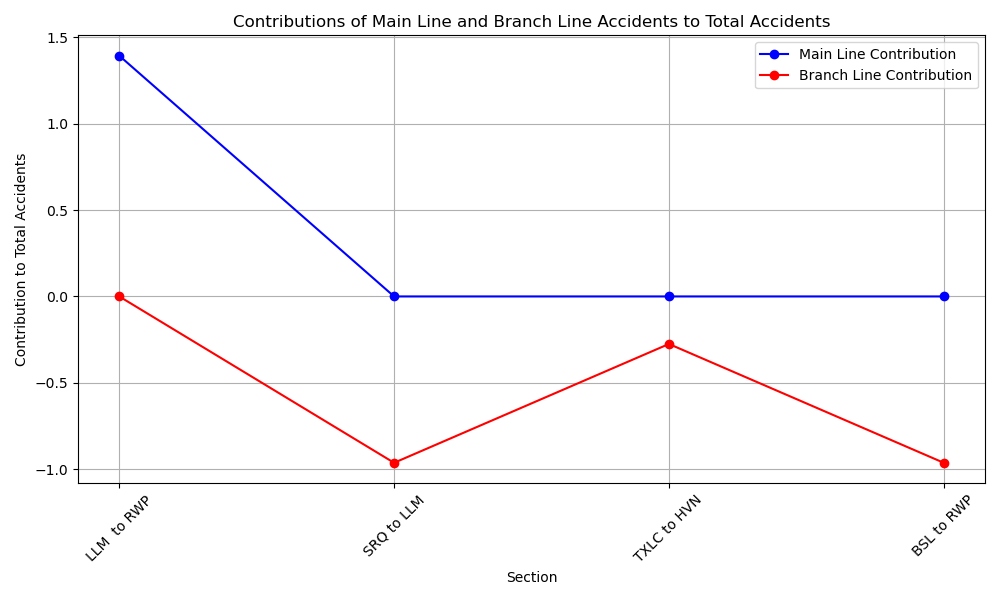
reflecting a good, but not as tight, fit as was obtained with the model for "Main Line."

The pseudo-r is 0.9612

Though it is still high, the fit is less precise in comparison with the "Main Line" model.

## Graph Analysis

The contribution from "Main Line" and "Branch Line" to the total number of accidents at different sections is clearly depicted in the graph. The observations are as follows:



## Main Line Contributions:

This graph is constantly positive, which means the greater the value of "Main Line," the more the accidents. This corresponds well with the positive coefficient observed for "Main Line" in the GLM and plays a further role in affirming the same—the "Main Line" variable contributes a lot to the rates of accidents.

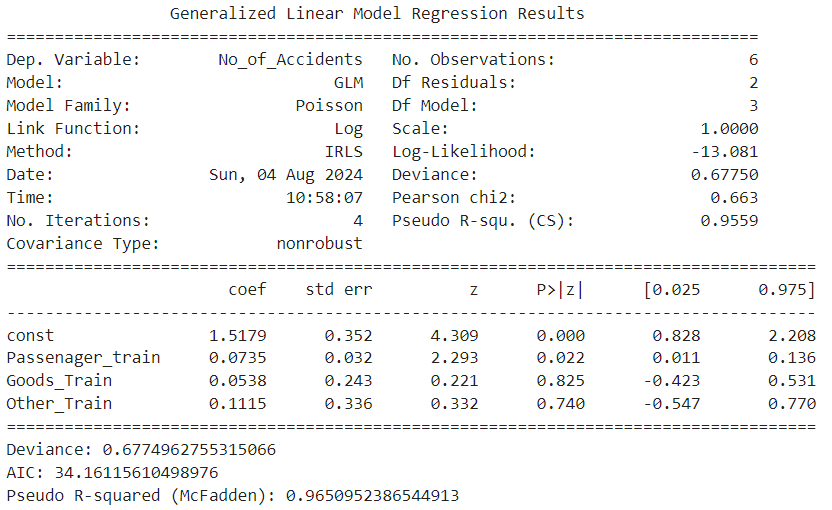
## Contributions of the Branch Line:

the scatter plot of branch line shows a highly varying effect. In some sections, branch line appears to have a positive impact while it shows a negative impact in other sections. This dispersion concurs with the negative coefficient from the GLM and gives further evidence that "Branch Line" positively does not have a very clear influence on the number of accidents occurring, with its influences probably differing in the different sections a great deal.

## Model Approach and Justification

The "Main Line" and "Branch Line" were modeled separately with a view to the accounting of the different effects of each kind of railway line on accident rates. The isolating of variables allows one to avoid confounding effects that may result from the combination of different variables under a single model. This will give clearer insight as to in what way each line type independently affects accident frequencies, and then make the model capable of doing further fine-tuning and explaining their individual contributions. Having separate models for these allows the possible biases to be accounted for and facilitates the proper interpretation of the influence of every variable in isolation, thereby providing a more precise and credible analysis of their effects.

# Train Type Accidents analysis



## Model Summary

Model The Generalized Linear Model applied on the dataset of number of accidents and train type (Passenger Train, Goods Train, and Other Train) informs us that:

## Passenger Train Coefﬁcient:

0.0735

## Standard Error

0.032

## Z-value

2.293

## P-value

0.022

The positive coefficient indicates that the higher the number of passenger trains, the higher is the rate of accidents. This effect is statistically significant, which means that there is a reliable positive relationship between the number of passenger trains and the frequency of accidents.

## Goods Train:

0.0538

## Standard Error:

0.243

## Z-Value:

0.221

## P-Value:

0.825

The positive coefficient indicates that the number of accidents may slightly increase with an increase in goods trains; however, the relation between goods trains and accidents is not statistically significant to say that it is robust with the given data.

## Coefficient for Other Train:

0.1115

## Standard Error:

0.336

## Z-Value:

0.332

## P-Value:

0.740

The positive coefficient indicates an increase in the number of accidents with an increase in other trains. Like for goods trains, this effect is not statistically significant; hence there is a weak relationship between other trains and the number of accidents.

## The model deviance of

0.6775

## AIC of

34.1612

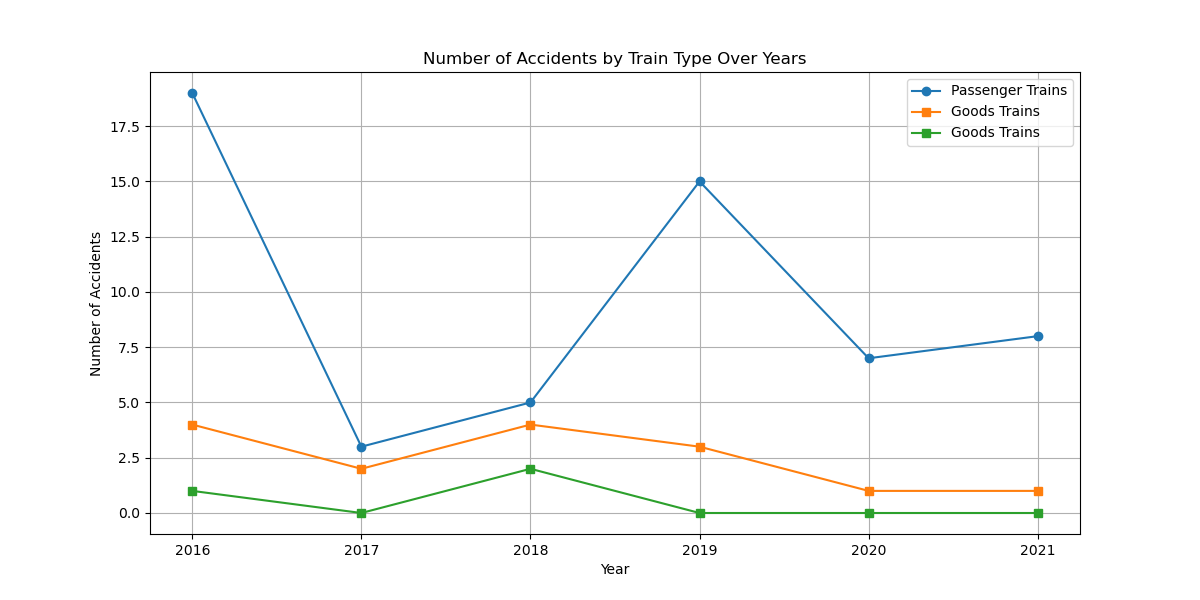
shows a good fit to data.

The pseudo R-squ = 0.9651

The measure 0.9651 indicates that the model explains a large part of the variation in the number of accidents, thus having strong explanatory power.

## Analysis of graph

The graph shows the trend in the no. of accidents of different types of trains with the change in years. Key observations are:



## Passenger Trains:

The graph indicates that passenger trains generally have the highest accidents over the years. The trend shows a constant positive effect on the total number of accidents, corresponding to the significant positive coefficient estimated in the GLM for passenger trains. The spikes observed in 2016 and 2019 indicate that these specific years had considerably high rates of accidents involving passenger trains.

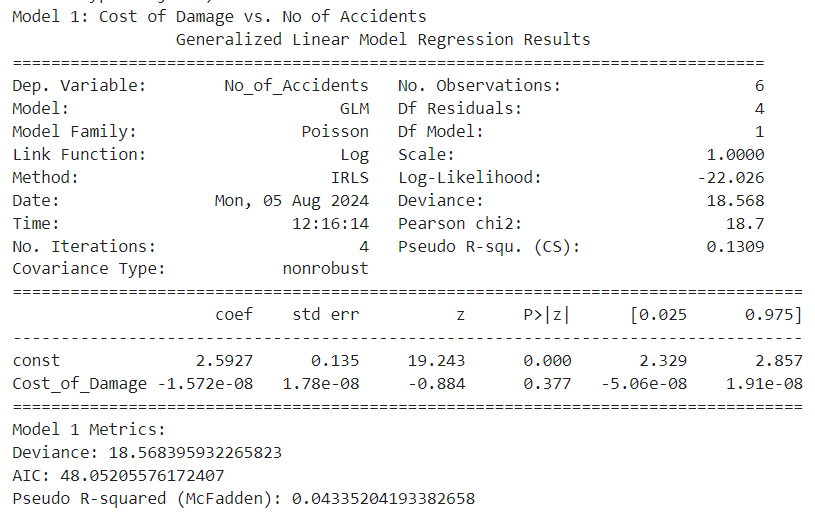
## Goods trains:

The number of accidents associated with goods trains is stable over the years, with a slight increase in 2018. However, according to the result of GLM, it does not seem that goods trains statistically significantly contribute to accident rates. Nevertheless, the trend line for goods trains still stays lower than that of passenger trains, suggesting that it contributes less to the total numbers of accidents.

## Other trains:

The graph indicates that other trains accidents are quite rare, with small outbreaks in 2016 and 2018. The general trend is still low and also confirmed by GLM results with a positive coefficient, not significant. This is evident from the graph that the other train factor has a very little effect on the total number of accidents.

# Cost of damage analysis/damage to pr/no damage to pr



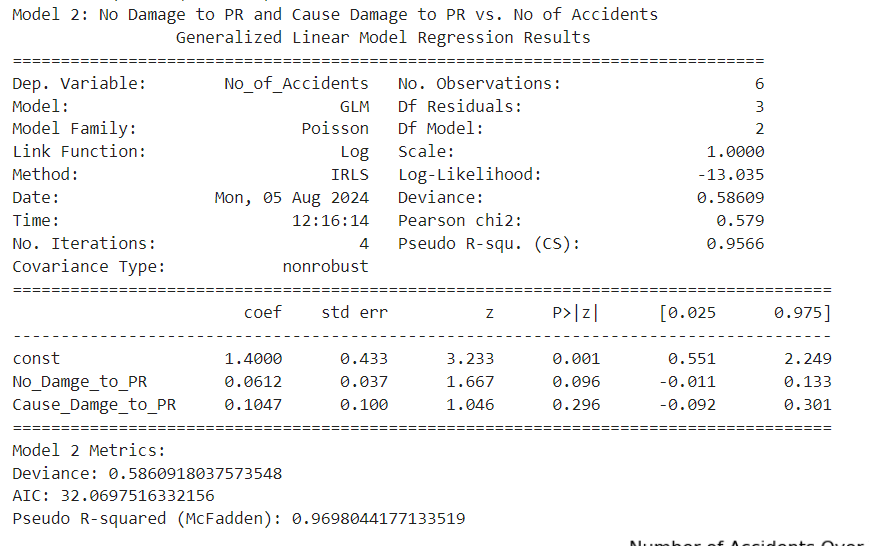
## Model : Cost of Damage vs. No. of Accidents

The first model relates the cost of damage and a number of accidents with each other based on a Poisson regression.

The response variable here is the number of accidents, and there are six observations in all. The residual degrees of freedom are four, and the model degree of freedom is one, showing this is a simple model with one predictor.  The family of the model is Poisson with a log link function, estimated via the Iteratively Reweighted Least Squares method. The log-likelihood of the model is -22.026. The deviance, which is a measure for goodness of fit of this model, is 18.568. A Pearson chi-square value of 18.7 indicates that the model is not fitting well. It took the model four iterations to converge. Using the pseudo R-squared value of 0.043, the model explains around 4.3 percent of the variance in number of accidents.

The coefficients of the model show that its intercept—constant—is 2.5927, with a very high significance of p < 0.001. This means it predicts the average log number of accidents when the cost of damage is zero. The coefficient for cost of damage is -1.572e-08, not statistically significant as p = 0.377. This suggests that cost of damage does not have much relationship with no. of accidents. The Akaike Information Criterion (AIC) for this model is 48.052. AIC is used for comparing models and a smaller value is a better fit.

## Model : Cause No Damage to PR and Cause Damage to PR vs. Number of Accidents



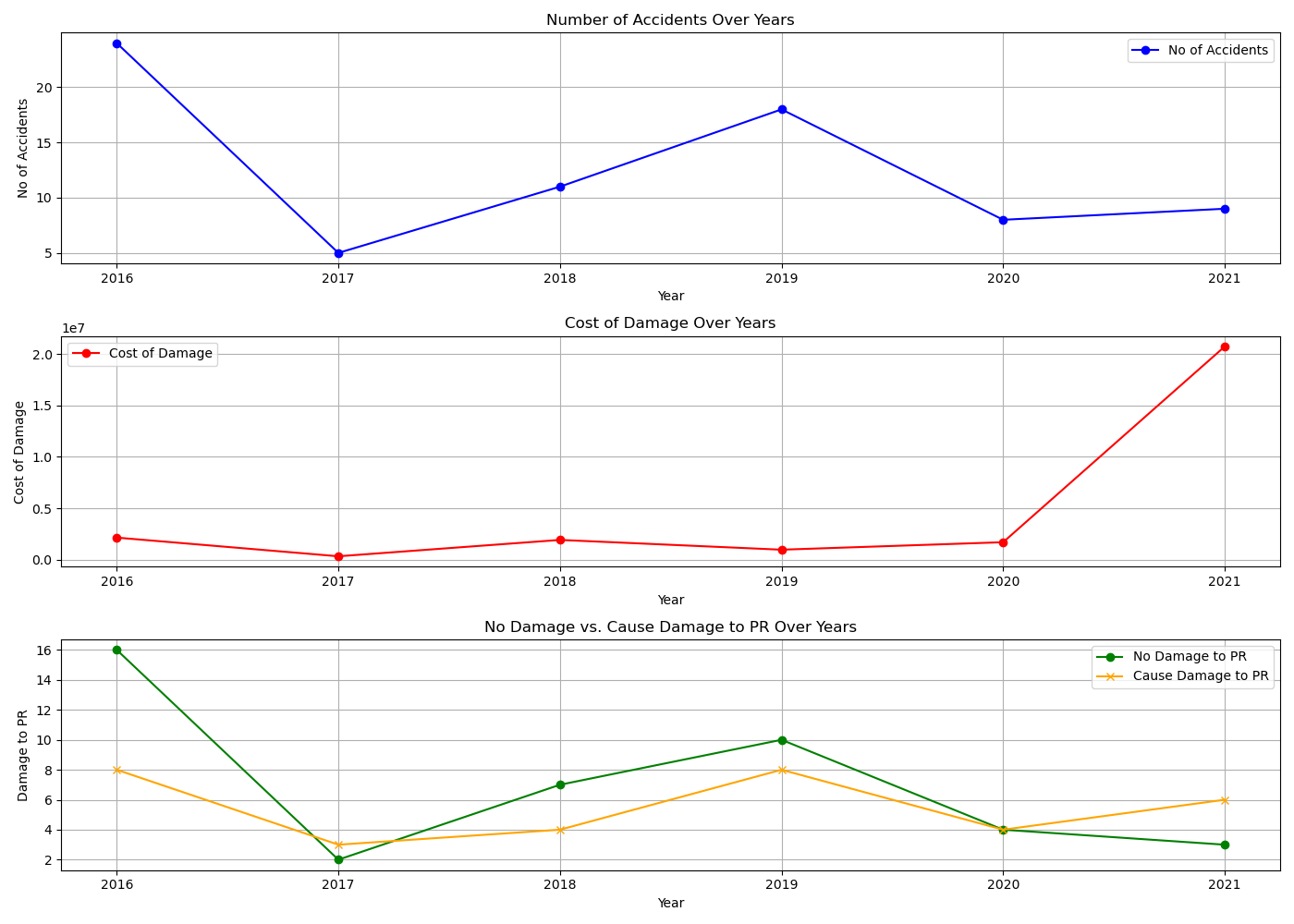
The second model investigates the influence of 'No Damage to PR' and 'Cause Damage to PR' on accidents through Poisson regression.

This model also uses a dependent variable of number of accidents with six observations. The degrees of freedom for residuals are three, and that for the model is two, which again makes this model a bit complex with two predictors. The model has a Poisson family with a log link function and will be estimated using the IRLS method. The log-likelihood of this model is −13.035, and the deviance is 0.5861, hence very good. The Pearson chi-square value is 0.579, again indicating excellent fit, and it converged after four iterations. The pseudo R-squared value is 0.970, indicating that the model explains about 97% of the variance in the number of accidents.

The coefficients for this model give an intercept of 1.4000, which is statistically significant at a p-value of 0.001. The coefficient for 'No Damage to PR' is 0.0612, which, with p = 0.096, suggests a possible positive relationship between the absence of damage to PR and the number of accidents. The coefficient for 'Cause Damage to PR' is 0.1047 and is not statistically significant with p = 0.296, indicating there is not very strong evidence that causing damage to PR is related to the number of accidents. The result of AIC for the model is 32.070, which may be used in comparison to other models.

## Analysis of Graphs

These graphs are showing the trends of accidents and related factors over the years.



The first graph shows the fluctuating trend of accidents over the years, peaking in the years 2016 and 2019. The general trend indicated a decrease in accidents after 2016, with a slight increase in 2019.

## Cost of Damage Over Years:

The second graph shows the cost of damage over the years. There is a spike in 2021, which shows that in this very year, the cost of damages increases drastically. Other years are rather stable with small cost fluctuations.

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

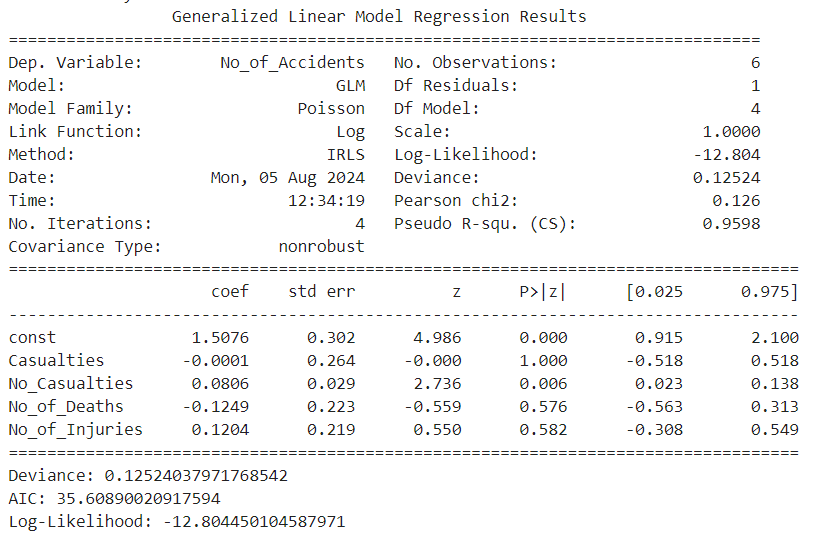
The third graph compares the number of incidents that do not result in damage to PR with those that do over the years. Both trends are relatively stable, with some ripples at times. Notably, the trend of incidents causing damage to PR does not follow any particular pattern. This indicates that they happen randomly.

## Discussion

The peak in the cost of damage in 2021 stands completely off-trend and does not relate to an increase in accidents, further weakly relating these variables. Overall, these findings indicate that other variables may bear a stronger relation in determining the number of accidents and warrant further investigation.

It is the use of a combined GLM model for all related factors that can explain how each of the factors contributes to the total number of accidents. This will present the contribution of every predictor variable towards the overall accident rate, hence giving insights into possible variables of influence useful for further research and prevention.

# Accident Severity Analysis



For the GLM related to the number of accidents, considering the occurrence of casualties, the number of casualties, deaths, and injuries as predictors, the findings are as follows:

## Intercept (Constant) for all predictors being zero:

the model tends to be predictive to the scenario of 1.51 accidents in a positive direction: this is taken as the expected count in a baseline condition.

## Casualties:

The coefficient for casualty number, which is almost zero, is associated with a probability of about p = 1.000, agent to insignificance. This courtesy shows that the number of casualties interlinked is less on the number of accidents taking place.

## No\_Casualties:

Has a positive coefficient value of 0.0806, meaning that there is a general trend for the number of cases without casualties to go up as the number of accidents increases. These are statistically significant at p = 0.006, hence confirmation of a meaningful effect of no casualties cases on the number of accidents.

## No of Deaths:

The coefficient of the no of deaths is negative but also not significant at all (p = 0.576), indicating that the model suggested changes in the number of deaths do not significantly affect the number of accidents.

## No\_of\_Injuries:

The number of injuries is positively but insignificantly significant at p = 0.582. Thus, it positively influences the accident but does not significantly vouch to vary in the number of accidents.

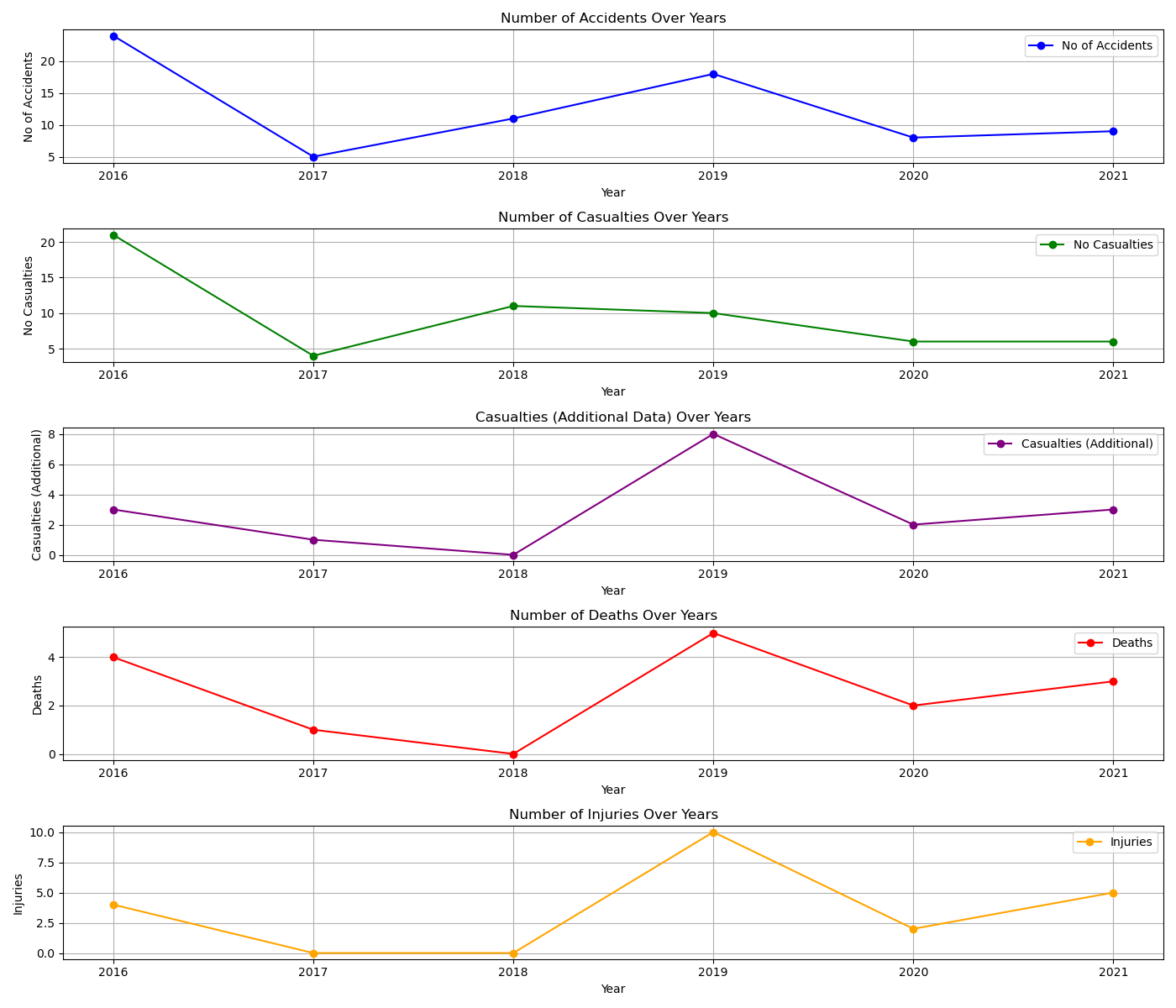
## Model Metrics:

The model shows very low deviance and high Pseudo R-squared, specifically at 0.125 and 0.9598, respectively; hence, it fits the data highly. The low deviance would represent low deviance, meaning the predictions of the model are quite closer to the observed data, while a high Pseudo R-squared indicates a very huge proportion of variance in the number of accidents being defined by the model.

In sum, the big relationship between the incidence at which there were no casualties and the accidents indicates that the former factor has a meaningfully predicted outcome, while casualties, deaths, or injuries do not present any significant influences that are present in the model.

## Analysis of Graphs

The graphs show the pictorial of trends of the number of accidents, casualties, deaths, and injuries over the years.



## Number of Accidents Over Years:

This graph indicates that the number of accidents has both increased and decreased over the years, peaking in 2016 and decreasing before increasing again in 2019.

## Number of Casualties Over Years:

This graph indicates that the number of casualties will, in most cases, take the same trend as the number of accidents, peaking in similar years.

## Casualties Extra Data Over Years: (NO casualties)

This graph shows the excess deaths data, which is more erratic but still has peaks in 2016 and 2019.

## Number of Deaths Over Years:

The graph shows ups and downs in the number of deaths, with peaks in 2016 and 2019.

## Number of Injuries Over Years:

This graph depicts ups and downs in the number of injuries, peaked significantly in 2019.

## Discussion

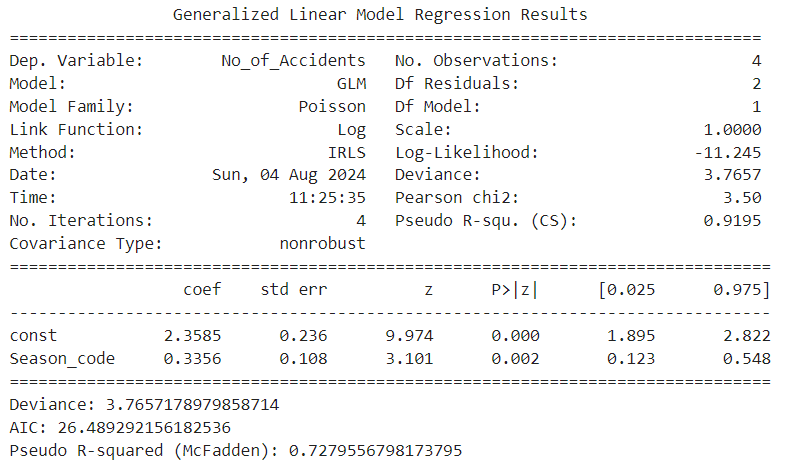
Using the GLMs analysis, the number of accidents comes out as a significant predictor for the number of casualities and injuries. The positive coefficients for these relationships suggest that an increase in the number of accidents would predict an increased severity of accidents in terms of casualties and injuries.

These are further confirmed by graphs of trends in the number of accidents and the measures of severity over the years. The peaks in 2016 and 2019 can be noted in more than one measure of severity as periods of higher severity in accidents.

The models for the number of deaths and the number of casualties have weaker relationships, indicative of the involvement of other factors in these outcomes. Further investigation into additional variables could give a fuller understanding of the factors affecting accident severity.

The general implication of this analysis is that addressing the causative factors for the number of accidents will go a long way in mitigating their severity in terms of casualties and injuries.

# Seasons wise accidents analysis



## Model Summary

The model takes into account the relationship between the number of railway accidents and the seasons. Discussion on the output is provided below:

The response variable is the number of accidents and the four observations are the seasons: Winter, Spring, Summer, Autumn. In this Poisson family with the log link function, the IRLS fits the data.

The highly significant large intercept coefficient of 2.3585, at p-value 0.000, denotes that the model is postulating a high baseline log count of accidents where the season code is zero. In other words, the coefficient of the variable Season\_code is positive—0.3356, at p-value 0.002—so this variable is also highly significantly, statistically positively related to accident counts.

The deviance of 3.7657 and an AIC of 26.4893 present the goodness of fit of the model; the Pseudo R-squared of 0.7280 tells that about 72.80% of the variation in the number of accidents is explained by the model. It shows that there is a good fit of the model.

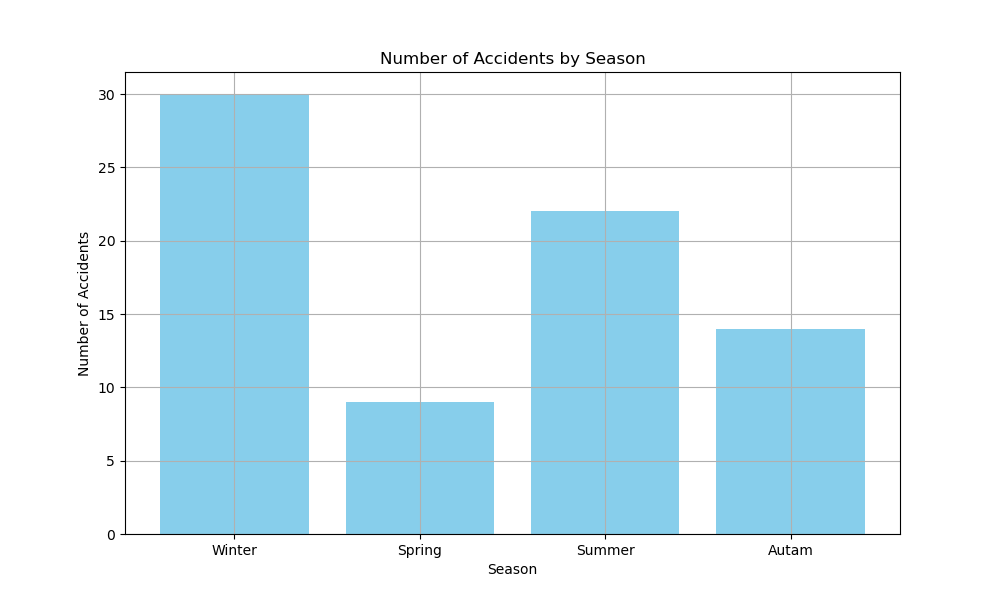
## Interpretation of Coefficients

The coefficient for the intercept is 2.3585, which provides the baseline log count of accidents when the season code is zero. This magnitude of the coefficient is very significant, showing a strong baseline effect. The coefficient on Season\_code, 0.3356, indicates that moving from Winter through Autumn seasons, the log count of accidents increases. That is to say, there is a positive relationship, and it is statistically significant, meaning different seasons have a marked impact on the number of railway accidents.

## Logic of Coding the Seasons in Codes

The seasons were coded in number codes being they were part of the GLM. This implied assigning every one of the unique seasons a numerical code, like Winter = 0, Spring = 1, Summer = 2, Autumn = 3. So now the categorical data is compatible, and the model will give an estimate on how the number of accidents vary with the changes in seasons. These numerical codes will enable the model to interpret and even quantify the effect of the different seasons on the number of accidents hence giving substantial insights.

Discussion of Graphs



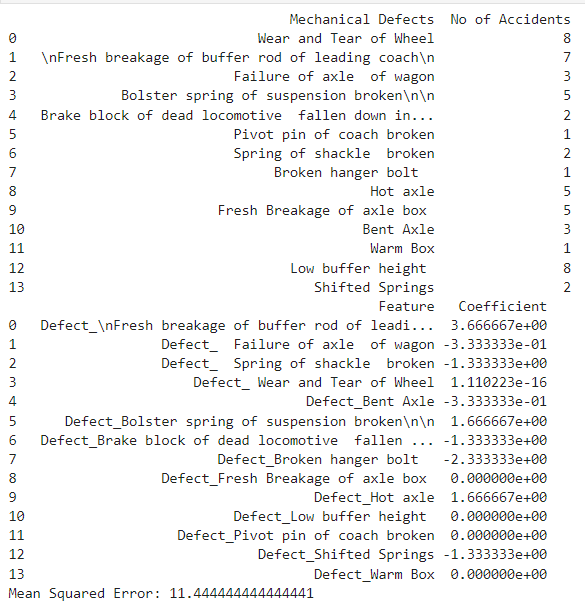
Graph of number of accidents against the different seasons shall be there to be a means of representing the data in an illustrative format. The bar chart will highlight the variation of the number of accidents against winter, spring, summer and autumn.

It is pretty apparent from the above chart that winter had maximum number of the cases at 30, followed by Summer at 22, Autumn at 14, and Spring at 9. In this sense, the regression model and barplot clearly correspond, because the bar graph also presents an increase in accidents from winter to autumn. The Season code in the model is positive, meaning that the farther the model proceeds from the winter to autumn in the seasonal progression, there will be an increasing accident trend, which is being visualized by the barplot.

Hence the graph supplements the statistical findings and evidences that seasonal variation is also one most important factor for railway accident analysis.

# Causes of Accidents

## Mechanical Defects



### Interpretation of Mechanical Defects

The analysis of mechanical defects and their association with the number of accidents reveals a range of impacts.

**Mechanical defects with positive coefficients** are associated with an increase in the number of accidents. Notably, defects like "Fresh breakage of buffer rod of leading coach," "Bolster spring of suspension broken," and "Hot axle" have coefficients of 3.67, 1.67, and 1.67 respectively. This indicates that these defects are significant contributors to accident rates. For instance, the high coefficient for the "Fresh breakage of buffer rod of leading coach" suggests that this defect leads to a substantial increase in accidents.

In contrast, **mechanical defects with negative coefficients** show a reduction in the number of accidents. Defects such as "Failure of axle of wagon," "Spring of shackle broken," and "Broken hanger bolt" have coefficients of -0.33, -1.33, and -2.33. These negative coefficients suggest that these defects are associated with fewer accidents. The defect "Broken hanger bolt," for example, shows a strong negative impact, indicating it is less likely to cause accidents.

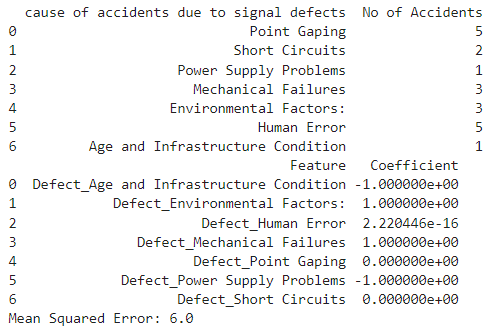
Some defects, including "Wear and Tear of Wheel," "Fresh Breakage of axle box," and "Low buffer height," have coefficients close to zero. This suggests that these defects do not have a significant impact on the number o

### Key Takeaways

1. **High Impact Defects**: Mechanical defects such as "Fresh breakage of buffer rod of leading coach," "Bolster spring of suspension broken," and "Hot axle" are significantly associated with higher accident rates. Addressing these defects should be a priority to improve safety.
2. **Low Impact Defects**: Some defects, like "Wear and Tear of Wheel" and "Low buffer height," show negligible impact on accident rates. These may not need as urgent attention in terms of safety interventions.
3. **Negative Impact Defects**: Defects such as "Failure of axle of wagon" and "Broken hanger bolt" are associated with fewer accidents. This counterintuitive finding suggests that these defects might not be as critical in causing accidents, though further investigation may be needed to understand the underlying reasons.
4. **Variable Effects**: The impact of mechanical defects on accident rates varies considerably, indicating that a targeted approach to defect management is necessary. Each defect affects accident rates differently, highlighting the need for tailored safety meaaccidents.f accidents.

In summary, mechanical defects such as the breakage of buffer rods and hot axles are crucial in increasing accident rates, while others like broken hanger bolts and shifted springs are less impactful or even reduce the accident rate. This analysis highlights the importance of addressing specific mechanical issues to enhance safety and reduce accidents.

## Signal Defects



### Interpretation of Signal Defects and Their Impact on Number of Accidents

The analysis of signal defects and their impact on the number of accidents is based on a Generalized Linear Model (GLM) with coefficients indicating the relative effect of each defect type. Here’s a breakdown of the findings:

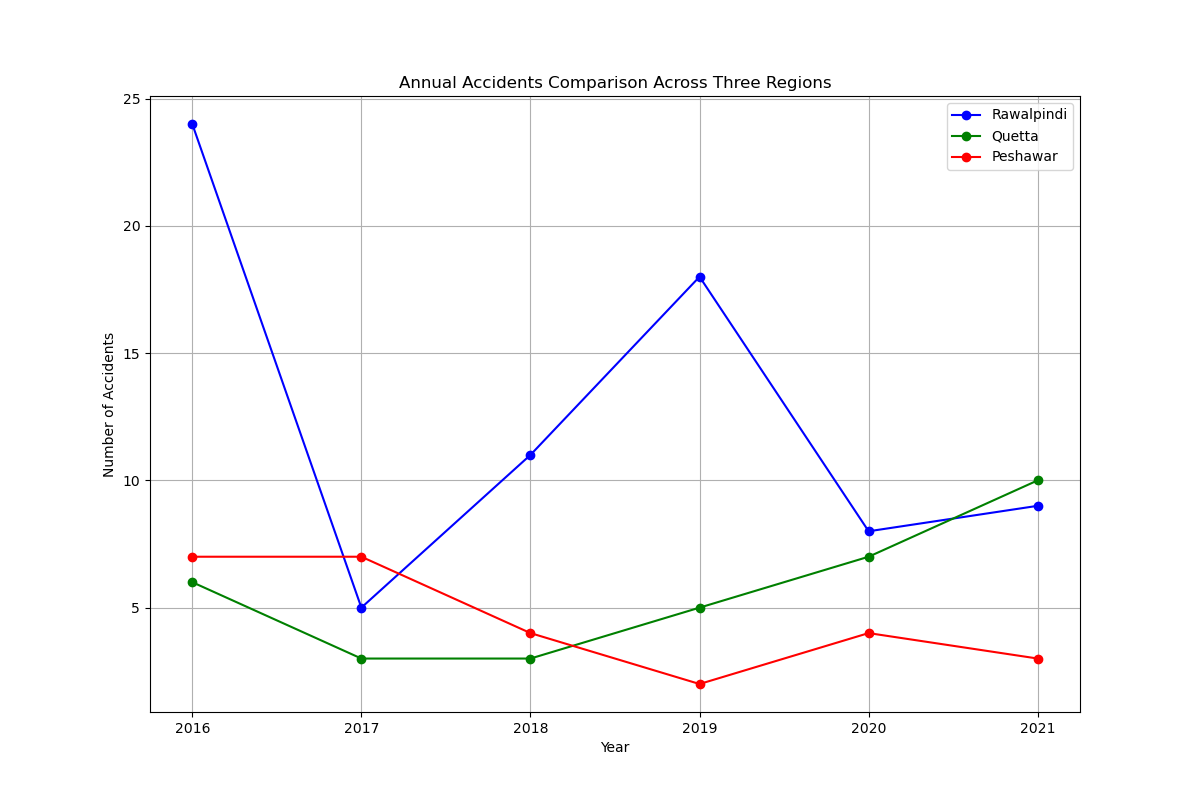
1. **High Impact Defects**:
   * **Human Error**: Despite a coefficient close to zero, the highest number of accidents are associated with human error. This suggests that while the model's coefficient may indicate a minimal direct effect, in practice, human error remains a significant factor due to its high frequency.
   * **Point Gaping**: Associated with a coefficient of zero, this defect appears to have a neutral effect in the model, yet it corresponds to a moderate number of accidents. This might indicate that while it doesn't show a strong statistical relationship in this model, it could still be a significant factor in real-world scenarios.
2. **Defects with Positive Coefficients**:
   * **Environmental Factors** and **Mechanical Failures**: Both defects have positive coefficients of 1.0000, indicating that they contribute significantly to the number of accidents. This suggests that these factors have a notable impact and should be prioritized in safety improvements.
3. **Defects with Negative Coefficients**:
   * **Age and Infrastructure Condition** and **Power Supply Problems**: These defects have negative coefficients, suggesting a decrease in the number of accidents associated with them. This could imply that improvements in these areas might have led to fewer accidents, or that these defects are less impactful in the context of this analysis.
4. **Neutral or Minimal Impact Defects**:
   * **Short Circuits** and **Point Gaping**: Both have coefficients of zero, indicating that these defects do not have a significant direct impact according to the model. However, this does not necessarily mean they are not important; their effect might be context-specific or less pronounced in the data used.

### Key Takeaways

* **High Priority Areas**: Focus on **Human Error**, **Point Gaping**, **Environmental Factors**, and **Mechanical Failures** for targeted interventions to reduce accidents. These factors show a significant relationship with accident rates.
* **Less Immediate Concern**: **Age and Infrastructure Condition** and **Power Supply Problems** appear to have a negative impact in this model, which might suggest these issues are being effectively managed or less critical in this dataset.
* **Model Limitations**: The coefficient values alone might not fully capture the practical significance of each defect. Further investigation and additional context might be needed to comprehensively address safety improvements.

Overall, the analysis suggests that specific defects, particularly **Human Error** and **Mechanical Failures**, should be addressed as priorities in safety management to effectively reduce the number of accidents.

# ANNUAL COMPARISON OF ACCIDENTS



## Discussion of the Graph

The graph portrays a comparison of the number of railway accidents on a yearly basis that occurred in Rawalpindi, Quetta, and Peshawar from the year 2016 to 2021. This comparative graphical representation depicts different trends and patterns in the three aforementioned areas.

## Rawalpindi:

The highest number of accidents occurred in 2016, but it drastically fell in 2017.

The years 2018 and 2019 reflected a steep rise and reached 18 accidents, but the numbers declined in the next two years.

## Quetta:

The accidents were quite low in number and stable from 2016 to 2018.

There is an increasing trend starting from 2019 and reaching as high as 10 in 2021, so that depicts an emerging concern for safety over the years.

## Peshawar:

Accidents were always very low with a maximum count in 2016 and 2017.

After decreasing in 2019, the accidents remain minor with a low count moving through 2021.

This comparative analysis shows that Rawalpindi had the highest variability and peaks in accident numbers as compared to Peshawar, which generally had more stable and lower rates of accidents. On the other hand, Quetta presents quite a concerning upward trend in the final years and hence is another area that could have higher safety interventions. Regional differences are well brought out in this graph and thus helpful in targeted decision-making about railway safety in these areas.