Report on Accidents, Fires and Breakdowns in French Motorway Tunnels in fulfilment of requirements for the course Regression Methods

Pranav Krishna

Contents

1. Introduction

The data we analyse are the number of Accidents, Breakdowns and Fires that occurred in French motorway tunnels. Here, Breakdowns refers to the number of breakdowns of cars and Fires refers to the number of cars that had fire relates events and not to events related to the tunnel. The covariates were Traffic, HGV(proportion of heavy-goods vehicles), Slope of the tunnel, Urban(if the tunnel is in an urban area), Type(if the Tunnel is Bidirectional or Unidirectional), Length of the Tunnel, Limit(the speed limit inside the tunnel), Slope-Type(the type of the slope inside the tunnel), Tunnel(the name of the tunnel), Direction(the direction in which an event occurred) and Company(the Company which runs the Tunnel). The data on Breakdowns and Accidents had the year corresponding the count of events in addition to the covariates mentioned above. Furthermore, the data on Accidents had the Width and number of Lanes of the tunnel. We added these columns to the relevant columns of the Breakdowns and Fires data-sets. We observe that direction is not a useful co-variate and remove it from the data-set.

The question of interest is the relation between these covariates and the number of various events that occur. The most important co-variate for these type of events is obviously, the traffic. We

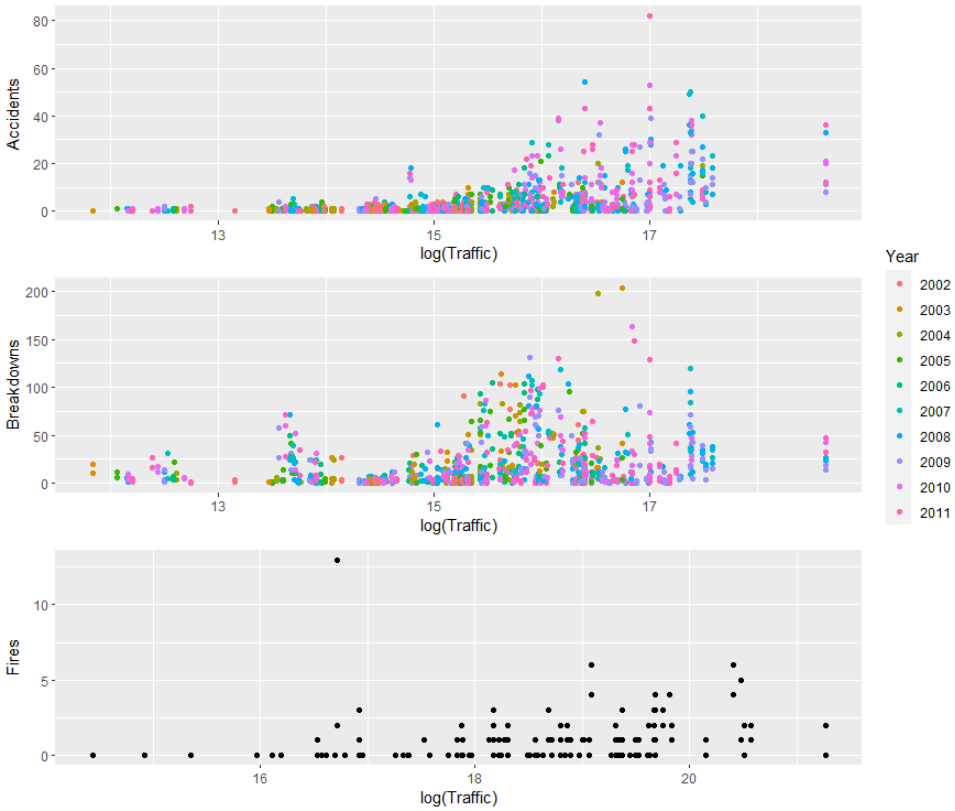

Figure 1: Scatter plots of Events vs log of traffic coloured by Year when available. Top Panel: Accidents vs log(Traffic). Middle Panel: Breakdowns vs log(Traffic). Bottom Panel: Fires vs log(Traffic)

observe in Figure 1 that the variability in the data rises with an increase in the Traffic. Furthermore, it is seen that there are a few outliers at a large value of Traffic and there are outliers for large values of events(1 outlier for 80 Accidents, 2 Outliers for about 200 Breakdowns and 1 outlier for more than 10 fires). In the Fires data, there are 3 outliers with a small value of Traffic.

The second most important co-variate is the speed limit. We remove the above-mentioned

outliers and make a scatter-plot of the number of events vs the speed limit. We observe the lack

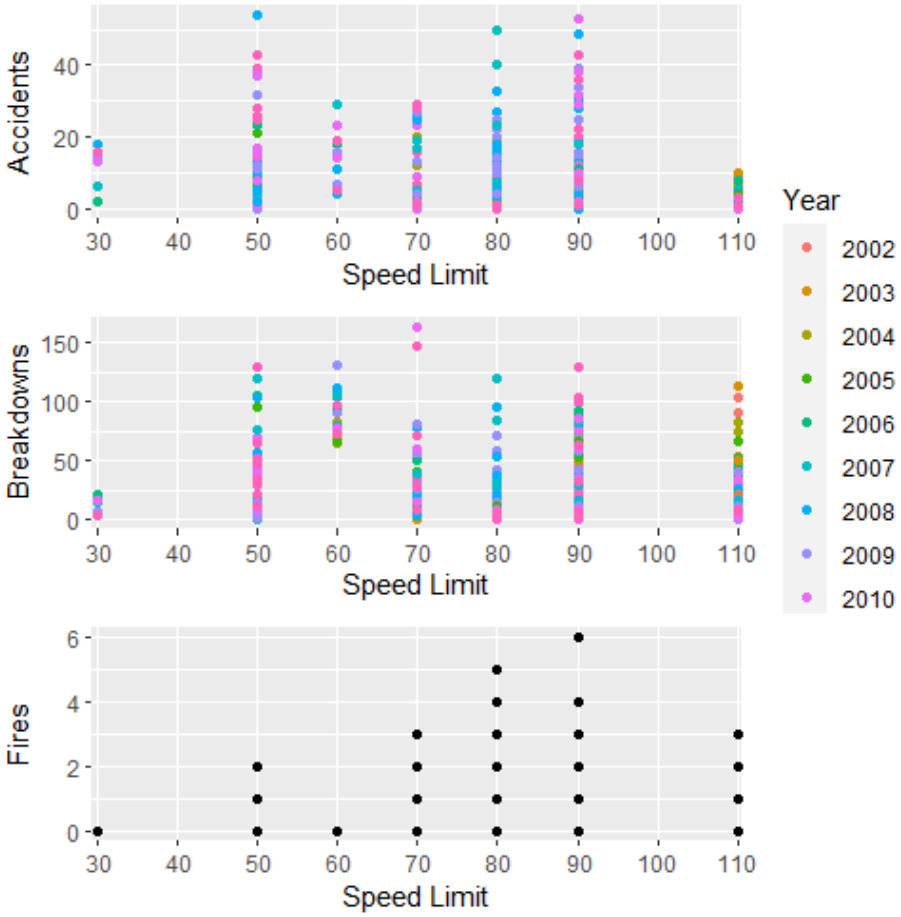

Figure 2: Scatter plots of Events vs Speed Limit coloured by Year when available. Top Panel: Accidents vs Speed Limit. Middle Panel: Speed Limit. Bottom Panel: Fires vs Speed Limit

of a clear pattern between the number of events and the speed limits in Figure 2. This suggests that there has been a revision in the speed limits of Tunnels based on the hypothesis that higher speeds might lead to higher accidents. On checking the data for the year 2002 (8 and 9), we observe that the same pattern holds, i.e., the speed limit does not have a monotonic relationship with the number of events.

Furthermore, we also observe that it would be better to treat the Year as a factor variable because we do not have a lot of temporal data and because it is hard to observe systematic variation in the data with respect to the year.

Another pattern we observed is that the tunnels with HGV zero are all urban. This suggests that they are a sub-class with no-entry for HGV vehicles, perhaps due to size or other structural limitations. We try to separate out these Tunnels whenever possible while doing the analysis. In our experience, such tunnels usually operate at capacity, allowing fast travel between residential and commercial municipalities during the rush hour.

In section 2, we detail the theory behind the models we choose to fit, the Poisson and Negative Binomial. In section 3, we detail the steps of the analysis . In section 4, we discuss the results of the analysis for each of the data-sets. Finally, we conclude in section 5.

2. Theory

The first model we try to fit is the Linear model. The linear model is the most basic model wherein the response y is assumed to be linear in terms of the parameters. i.e.,

$$Y = X\beta + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0_n, \sigma^2 I_n) \tag{2.1}$$

where Y denotes the vector of responses, X denotes the matrix of co-variates and β is the vector of coefficients in the model. We realise that the assumptions of the linear model are not satisfied here, but as mentioned in Slide 221 in (1), sometimes, taking a variance stabilising transformation, viz. $y \to 2\sqrt{y}$ for Poisson data, can be a fix. It is not the case for any of the data-sets so we look for more sophisticated models.

We choose to model the data using Generalised Linear models. They describe the dependence of a scalar variable y on a vector of co-variates x. The defining property is that

$$f(y; \theta, \phi) = exp\left\{\frac{y \cdot \theta - b(\theta)}{\phi} + c(y; \phi)\right\} \tag{2.2}$$

as mentioned on slide 246 in (1). If $\eta = x^\top \beta$, the covariates enter the model via the monotone link function as $\eta = g(\mu)$ where $\mu = \mathbf{E}(y)$. Another way to think about it is that $\mu_i = g^{-1}(\eta_i)$, that is $\mathbf{E}[Y_i|x_i] = g^{-1}(x_i^\top \beta)$.

We observe that the response is the count of the number of events that occur at a particular tunnel in one calendar year. Thus, the next model we try to fit is the Poisson regression model, a member of the GLM family in 2.2. The Poisson distribution has the pmf

$$P(y; \mu) = \frac{e^{-\mu}\mu^y}{y!}. \tag{2.3}$$

It is easy to see that the mean of a Poisson random variable is the same as its variance. This property is called equidispersion. As mentioned in (2) Section 10.5, the Poisson regression models Y as a Poisson variable with mean μ. Using the Canonical link, viz. log, we have that $\mu_i = e^{x_i^\top \beta}$, which ensures that μ_i is non-negative. This model implies that $\mathbf{E}[Y_i|x_i] = e^{x_i^\top \beta}$. This allows the effects of Traffic, Limit and Company etc to be incorporated in the model multiplicatively instead of being additive as in the Linear model.

Frequently, it is observed that the variance of the data exceeds the mean. This phenomenon is called over-dispersion. One possible reason why the data has over-dispersion is because of the possibility that the tunnels being operated by a company are only in one particular geographical area. To deal with over-dispersion, we model the response using Negative Binomial Regression.

The Negative Binomial Distribution has many parametrisations. While one can think of it in terms of the number of tosses until a set number of heads, another way it can arise is as a Gamma mixture of Poisson random variables. One parametrisation of the density with mean μ and shape parameter θ is

$$f(y; \mu, \theta) = \frac{\Gamma(y + \theta)}{\Gamma(\theta) \cdot \Gamma(y + 1)} \frac{\mu^y \cdot \theta^\theta}{(\mu + \theta)^{y+\theta}} \tag{2.4}$$

which is another special case of the GLM family in 2.2, for every fixed value of θ. Observe that if we set $\theta = \frac{\mu}{\xi}$, the variance function is $V(\mu) = \mu + \frac{\mu^2}{\xi}$, as mentioned in Slide 301 in (1). Again, the link function is the log, and this model also implies that $\mathbf{E}[Y_i|x_i] = e^{x_i^\top \beta}$, incorporating the effects

of the covariates multiplicatively. Clearly, the parameter θ allows us to model the variance in the data in a more effective manner.

Another observation we can make is that

$$\lim_{\xi \to \infty} V(x) = \mu. \tag{2.5}$$

Thus, as mentioned in (3) in Section 8.1 on Page 187, we see that Poisson Regression can be thought of as a special case of Negative Binomial Regression.

The AIC and BIC for GLMs are defined in terms of the Deviance as mentioned in slide 266 in (1). As mentioned on the slide, the BIC applies a stronger penalty than AIC but it can result in a model with poor predictions. We use the stepAIC function in R for model selection. We implemented backwards selection from the full model to reach a candidate model. Based on comments in slide 65 at (1), we know that AIC tends to select for complex models. Thus, we tried to counteract this by trying to remove covariates and asses the change in AIC. For assessing this change, we use the Table for AIC significance Levels on Page 70 in (3). When these methods didn't result in a meaningful change, we used BIC for model selection.

For evaluation the fit of the model, we utilise the Cook's Distance, Leverage and the plots of Deviance Residuals vs Fitted values. Cook's Distance and Deviance residuals for Generalised linear models is defined in section 10.2.3 (2) on page 477. For each plot of Cook's Distance, we use the threshold value of $\frac{8}{n-2p}$ as mentioned on page 396 in the first paragraph in (2).

3. Analysis

In this section, we detail the models that we fitted to the data-sets. In each case, we start with a linear model and go on to fit a Poisson model and finally, a Negative Binomial Model. For each data-set, we observed that the linear model was inadequate. It gave negative predicted values in Fires and Breakdowns and for the accidents data-set, even though the predicted values were positive, most of the data was outside a ± 2 band in a plot of Residuals vs Fitted values.

For the Fires data-set, we were not able to separate the data into zero and non-zero HGV components satisfactorily. For the other two data-sets, we separate them into zero and non-zero HGV portions. For the non-zero HGV portion, we fit two models. For one model, we supply the number of Heavy Goods Vehicles by multiplying the proportion with the Traffic and then taking the log of both quantities, where ever possible. In the other case, we use HGV as a proportion.

3.1. Fires Data

As mentioned earlier, we decide to fit the Poisson model. We tried to separate the data-set by the proportion of HGV vehicles so that the analysis on tunnels with 0 HGV vehicles could be done separately but we observed that all such tunnels have either 0 or 1 fires. Thus, the glm function in R refused to fit a Poisson model. While one can fit a logistic regression to this data, it would imply having an assumption that there can only be 0 or 1 accidents in these tunnels but no more. Clearly, such a modelling assumption is not justified.

First, we supply the HGV co-variate as a proportion. Now, because the different types of slopes were all represented by one number, we introduce the interaction between them. On adding this interaction, we observed that the decrease in the Residual deviance was insignificant but we kept it in the model.

We try to remove terms sequentially based on Table 5.1 in (3). As we have 163 observations, we need to look out for rise in AIC by more than 6. Thus, sequentially removing terms from the model, we end up with Traffic and Length as the only covariates.

On supplying the full model using count data in HGV to stepAIC, we again get a model with both Slope and Slope Type but without the interaction term. Again, we try to sequentially remove terms from the model based on Table 5.1 in (3) and finally end up with HGV, Length and Company as the remaining covariates. We recall that we cannot take log of HGV as it contains zeros.

Lastly, we tried to fit a negative binomial model but we observed that there were warnings that the iterations exceeded the threshold for the dispersion parameter . We believe this happened because of the limiting property of the negative binomial regression as mentioned earlier at 2.5. In the interest of having a simpler model and recognising that HGV is neither continuous nor discrete, we reject this model.

We look at the plot of residuals vs fitted values for the Poisson and remove observations where the residual was more than 6 and refit the model to get a Poisson model such that

$$\log(\mu) = -22.43 + 0.91 \cdot \log(\text{Traffic}) + 0.70 \cdot \log(\text{Length}) \tag{3.1}$$

3.2. Breakdowns data

Even though we fitted a Poisson model, it was with a view that the final model was likely to be Negative-Binomial. Our apprehension turns out to be correct in that the fit of the negative binomial model is significantly better for both zero and non-zero HGV tunnels.

First, we discuss the results of analysis on the tunnels with non-zero HGV. We transform the HGV proportion into the count and take the log of both the Traffic and the number of HGVs. Furthermore, we include the interaction term between Slope and Slope Type and use the stepAIC function for selecting covariates. Based on Table 5.1 in (3), we note that we are in the first row and removing a co-variate would leave the AIC essentially unchanged. We observed this in the analysis and we were not able to drop any covariates. Thus, we supplied this function to the step function using BIC as the model-evaluation criterion. This allowed us to drop the Limit, Year and the interaction between Slope and Slope Type. On inspecting the Analysis of Deviance Table for this model, we see that the drop in Deviance is not high for Direction and Lanes. On removing these two , we don't see a significant rise in the Deviance, even thought he AIC rises by a lot. Thus, we decide to drop these two co-variates.

We observe that HGV did not factor into the final model for the non-zero HGV portion of the data-set. We applied a similar procedure to the data-set using HGV as a proportion and reached an identical model.

Observe that for the HGV zero data, we are in the third row of Table 5.1 in (3). On removing co-variates one-by-one, we see that we were able to remove all covariates except company. We removed company while keeping other covariates and observed that the AIC rose up sharply. Thus, we observe that for HGV zero tunnels, the main factor that affects Breakdowns is the Company operating the tunnel, irrespective of the Traffic, length and other covariates.

Finally, for zero HGV tunnels, we have a Negative Binomial model with

$$\log(\mu) = 4.22 \text{ - } 3.66\mathbb{1}_{\{\text{Castoridae}\}} \text{ - } 0.60\mathbb{1}_{\{\text{Eschrichtidae}\}} \text{ - } 0.08\mathbb{1}_{\{\text{Obobenidae}\}} + 0.27\mathbb{1}_{\{\text{Phocoenidae}\}} \tag{3.2}$$

with Company Bovidae as the baseline. For non-zero HGV tunnels, we have a Negative Binomial model such that

$$\begin{aligned}\log(\mu) = \text{ - } 17.94 + 0.84 \cdot \log(\text{Traffic}) + 14.62 \cdot \text{Slope} + 1.00 \cdot \log(\text{Length}) \\ + C_{\text{Slope-Type}}\mathbb{1}_{\{\text{Slope-Type}\}} + C_{\text{Company}}\mathbb{1}_{\{\text{Company}\}}\end{aligned} \tag{3.3}$$

3.3. Accidents Data

Similar to the Breakdowns data, we fitted a Poisson model with the understanding that the Negative Binomial model was likely to be better.

First, we discuss the analysis for non-zero HGV tunnels. For the first model, we use HGV as a count take the log. We add the interaction term for Slope and Slope Type and supply the model to stepAIC for selecting covariates. Again, we are in the top row situation of Table 5.1 in (3) and dropping the covariates is not possible. Thus, we try to select using BIC and are able to drop a few covariates. Thus, we have a Negative Binomial model such that

$$\begin{aligned}\log(\mu) = \; & -7.78 + 0.37 \cdot \log(\text{HGV}) - 3.67 \cdot \text{Slope} + 0.47 \cdot \mathbb{1}_{\{\text{Urban}\}} + 0.56 \cdot \log(\text{Length}) + \\ & 0.21 \cdot \text{Lanes} + C_{\text{Company}} \mathbb{1}_{\{\text{Company}\}} + C_{\text{Slope Type}} \mathbb{1}_{\{\text{Slope Type}\}} \end{aligned} \tag{3.4}$$

When we tried to analyse the data with HGV as a proportion, and supplied the model including all covariates and the interaction term to stepAIC, we were able to drop the interaction term between slope and Slope-Type. Again, we find it difficult to remove covariates and resort to using BIC for model selection. Thus, we end up with a Negative Binomial model such that

$$\begin{aligned}\log(\mu) = \; & -3.81 + 2.26 \cdot \log(\text{HGV}) - 4.11 \cdot \text{Slope} + 0.80 \cdot \mathbb{1}_{\{\text{Urban}\}} + 0.64 \cdot \log(\text{Length}) + \\ & 0.39 \cdot \text{Lanes} + C_{\text{Company}} \mathbb{1}_{\{\text{Company}\}} + C_{\text{Slope Type}} \mathbb{1}_{\{\text{Slope Type}\}} \end{aligned} \tag{3.5}$$

We observe that both the models have the same covariates and have similar coefficients. We use the model 3.4 because it utilises the number of HGV instead of utilising it as a proportion.

For the tunnels with zero HGV, we supply the full model, including the interaction between Slope and Slope Type to stepAIC for selecting the appropriate covariates. After this, we refer to Table 5.1 in (3) and are are able to remove Traffic, Slope and Year. Thus, our final model for tunnels with zero HGV is negative binomial such that -

$$\begin{aligned}\log(\mu) = \; & 2.58 + 1.80 \mathbb{1}_{\{\text{Slope Type Basin}\}} + 0.71 \mathbb{1}_{\{\text{Slope Type Continuous}\}} - 0.08 \mathbb{1}_{\{\text{Castoridae}\}} \\ & - 1.34 \mathbb{1}_{\{\text{Eschrichtidae}\}} \end{aligned} \tag{3.6}$$

4. Discussion of the Results

In this section, we discuss the results of the analysis. The models are indeed useful for explaining the variation for the data-sets with respect to the chi-squared test using Deviance as mentioned on slide 235 in (1). For each final model, we observed that the p-value associated to it was 0.

4.1. Fires Data

We see in Table 1, that Traffic and Length behave as expected with regards to the number of Fires that are observed in each tunnel. To investigate the fit of the model, we look at a few diagnostic plots in Figure 3. We see that most of the residuals are inside a ± 2 band. From the QQ-plot of Deviance residuals, we see that they are skewed to the right. Since we tried the Negative Binomial model, this suggests that fitting a Zero-inflated Poisson model would be worth exploring. From the plot of Cook's Distance vs leverage, we see that even though a few observations have a high Cook's Distance, they do not have high leverage and broadly speaking, the model appears adequately fitted to the data.

	Estimate	Std. Error	z value
(Intercept)	-22.436	2.480	-9.04
log(Traffic)	0.913	0.105	8.69
log(Length)	0.699	0.129	5.43

Table 1: Model Summary for Fires Data

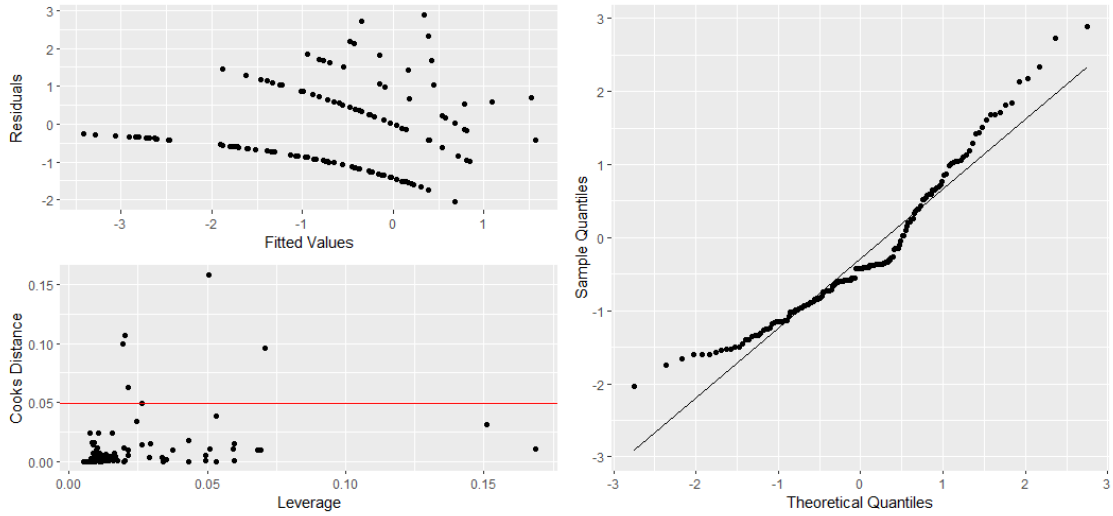

Figure 3: Diagnostic Plots for Fires data. Top Left: Plot of Deviance Residuals vs Fitted Values. Right: QQ plot of Deviance. Bottom Left: Plot of Cook's Distance vs Leverage

4.2. Breakdowns Data

4.2.1. Zero HGV Tunnels

We observe in Table 2 that while the sign for most companies is negative, it is positive for Phococenidae and the baseline Bovidae. This warrants a closer look a the tunnels run by these companies. It must be stated that without further information, a negative coefficient doesn't necessarily suggest issues with their maintenance practices. For instance, worse enforcement of vehicular regulations in the adjacent municipalities of the Tunnel can lead to such results and the Company would not be responsible for it.

	Estimate	Std. Error	z value
(Intercept)	4.219	0.154	27.37
CompanyCASTORIDAE	-3.659	0.452	-8.09
CompanyESCHRICHTIIDAE	-0.596	0.180	-3.30
CompanyODOBENIDAE	-0.083	0.183	-0.46
CompanyPHOCOENIDAE	0.272	0.178	1.53

Table 2: Model Summary for Breakdowns in zero HGV tunnels

In figure 4, we see the diagnostic plots for this model. We see from Residuals vs Fitted Values that all of the residuals are in the ± 1 error band. Similarly, the Cook's statistics are well below the threshold and suggest that the model fit is excellent. On the other hand, the QQ plot tells a different story. The residuals on the left tail appear to depart significantly from normality. We believe this is the situation as mentioned in page 447 in (2) in the second paragraph. We observe

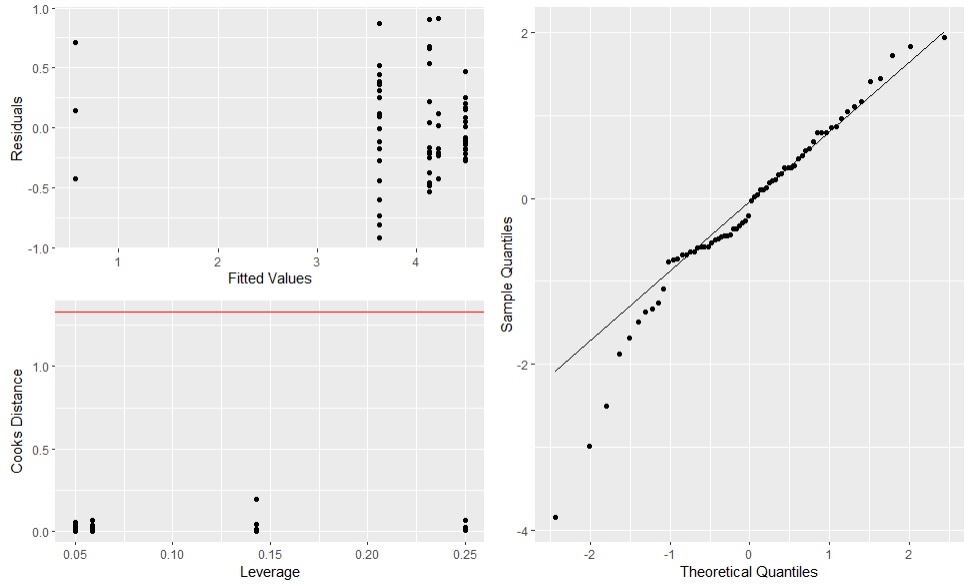

Figure 4: Diagnostic Plots for Breakdowns data with zero HGV. Top Left: Plot of Residuals vs Fitted Values. Right: QQ Plot of Deviance Residuals. Bottom Left: Plot of Cook's Distance vs Leverage

that the residuals are essentially discrete and thus, do not follow a normal distribution. Overall, we believe that the model fits adequately to the data.

	Estimate	Std. Error	z value
(Intercept)	-17.936	0.993	-18.06
log(Traffic)	0.839	0.059	14.05
Slope	14.622	1.058	13.82
log(Length)	1.002	0.045	21.88
SlopeTypeBasin	0.824	0.282	2.92
SlopeTypeUnknown	1.160	0.303	3.82
SlopeTypeContinuous slope	0.024	0.267	0.09
SlopeTypeRoof	0.030	0.288	0.10

Table 3: Section of Model summary for Breakdowns in non-zero HGV tunnels

4.2.2. Non-zero HGV Tunnels

We see from Table 3 that the slope has a large effect on the Number of Breakdowns. The positive sign is to be expected as a vehicle going downhill does not strain the engine while going uphill does. On inspecting the Tunnel named Eubalaena glacialis, we see that it as the largest value for the slope, 0.065. Since this is bidirectional, we have the opportunity to compare the effect of slope while keeping other factors constant. We see that the direction with positive slope has 32, 23 and 33 breakdowns or the years 2009, 2010 and 2011. On the hand, the direction with negative slope has 5, 6 and 5 Breakdowns respectively. This is borne out by our model which says that this change in slope would be associated with $e^{14.622*2*0.065} \approx 7.38$ times an increase in the number of breakdowns. Owe observe that a similar trend is observed in our data. The effect of the length and traffic are as would be expected and do not raise any serious questions about the validity of the model.

In Figure 5, we see that most of the residuals are within a ±2 band. While the Cook's Distance is large for a few points, we see that they do not have high leverage. On inspecting the observations with large Cook's Distance(observations above the red line), we see that the deviance residuals are unremarkable except for the one observation which has a residual more than 3. This suggests that this observation is an outlier. From the QQ plot, we see that the Deviance Residuals appear to be normal. Thus, the model fits to the data very well.

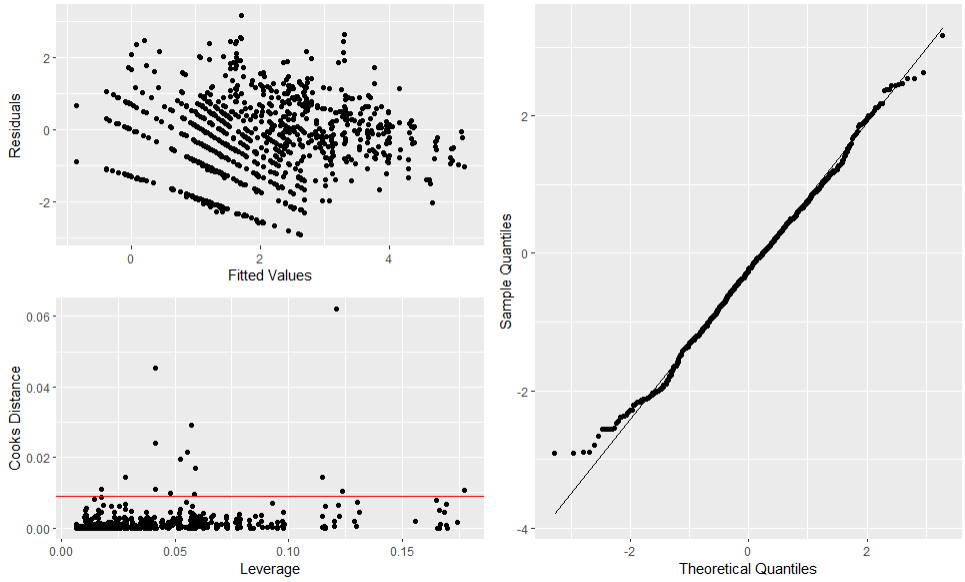

Figure 5: Diagnostic Plots for Breakdowns data with non-zero HGV. Top Left: Plot of Deviance Residuals vs Fitted Values. Right: QQ Plot of Deviance Residuals. Bottom Left: Plot of Cook's Distance vs Leverage

4.3. Accidents Data

4.3.1. Zero HGV Tunnels

We observe from the data all tunnels run by the company Phocoenidae have Slope Type other and all tunnels with HGV zero and Slope Type other are operated by it. Thus, the coefficients for this combination aren't estimable separately. In the output, we see that the coefficient for the company Phocoenidae is NA and the Slope-Type other is taken as the baseline. The output is as seen in Table 4. The fact that some coefficients for this model are inestimable make interpretation difficult. We see that the Continuous-Castoridae combination is associated to the least number of accidents whereas Basin-Obobenidae is associated with the most number of accidents.

	Estimate	Std. Error	z value
(Intercept)	2.577	0.196	13.14
SlopeTypeBasin	1.796	0.425	4.22
SlopeTypeContinuous slope	0.712	0.314	2.27
CompanyCASTORIDAE	-3.760	0.555	-6.77
CompanyESCHRICHTIIDAE	-1.339	0.281	-4.76
CompanyODOBENIDAE	-1.054	0.309	-3.41

Table 4: Model summary for Accidents in zero HGV tunnels

We see the diagnostic plots for this model in Figure 6. We see from the plot of deviance residuals that all but 3 observations are in a ± 2 band. From the plot of Cook's Distance, we see that all observations are below the threshold for a large Cook's Distance. While the QQplot does appear to depart from normality, it is only in the left tail. Similar to the model for Breakdowns, we believe this is the situation as mentioned in page 447 in (2) in the second paragraph. We observe that the residuals are essentially discrete and thus, do not follow a normal distribution. This suggests that the the fit of the model is adequate.

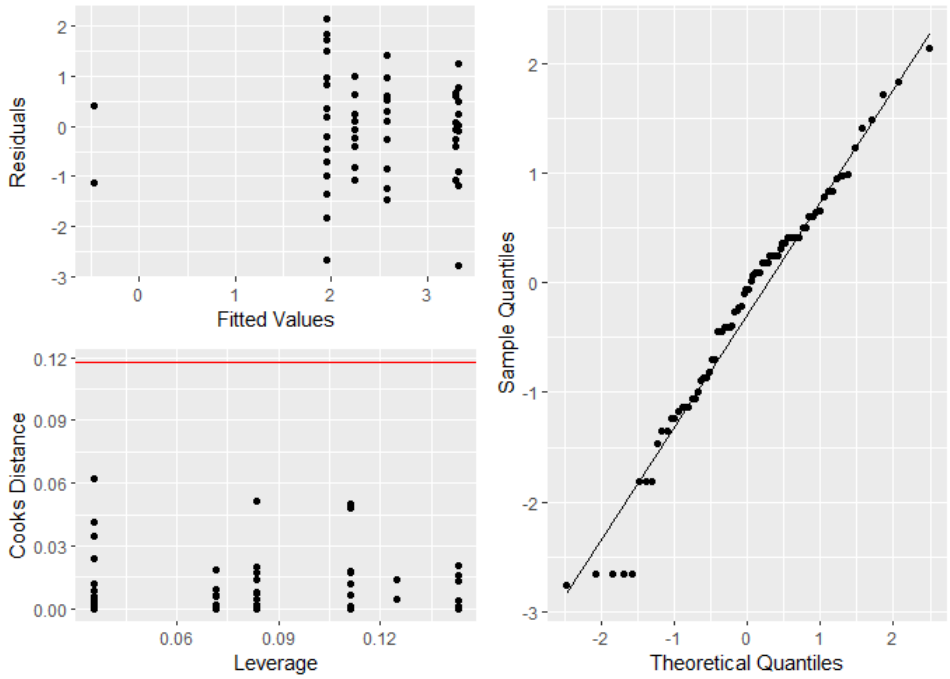

Figure 6: Diagnostic Plots for Accidents data with zero HGV. Top Left: Plot of Deviance Residuals vs Fitted Values. Right: QQ Plot of Deviance Residuals. Bottom Left: Plot of Cook's Distance vs Leverage

4.3.2. Non-zero HGV Tunnels

We see in Table 5 that instead of Traffic, the important co-variate is the number of heavy goods vehicles passing through a tunnel. This suggests that most accidents would have involved such vehicles, either directly or indirectly. Furthermore, we think that a re-assessment of which tunnels are safe for HGVs would also help to reduce the number of accidents. The fact that the coefficient for Urban is positive reflects the chaotic nature of Traffic in Urban areas, which can lead to a higher number of accidents, especially with the presence of HGVs which are not as manoeuvrable as cars. The fact that the coefficient for lanes is positive is surprising because one would expect more leans to lead to an orderly flow of traffic. One possible interpretation is that the number of lanes is correlated to traffic, which leads to higher number of accidents overall.

Finally, we see that the coefficient for slope is negative. This is unexpected in light of the model for Breakdowns in non-zero HGV tunnels as one would expect a more breakdowns to also lead to more accidents. One possible explanation is that once a vehicle breaks down while going up a slope, the hand brake can be used to bring it safely to a halt as gravity would also be slowing it down. But, for a vehicle going down a slope, the hand brake would take longer to bring the vehicle to a halt as gravity would be making is go downwards faster. This difference in the situations can lead

to an opposite effect of slope on Breakdowns and accidents.

We see the diagnostic plots for this model in Figure 7. We observe that most of the residuals are

	Estimate	Std. Error	z value
(Intercept)	-7.777	0.873	-8.90
log(HGV)	0.367	0.058	6.32
Slope	-3.671	1.189	-3.09
UrbanYes	0.472	0.124	3.79
log(Length)	0.556	0.056	9.79
SlopeTypeBasin	-0.471	0.240	-1.96
SlopeTypeUnknown	-0.604	0.329	-1.83
SlopeTypeContinuous slope	-1.105	0.219	-5.03
SlopeTypeRoof	-1.272	0.265	-4.79
Lanes	0.217	0.070	3.07

Table 5: Section of Model summary for Accidents in non-zero HGV tunnels

within a ± 2 band. For the Cooks distance, we see that a few observations are above the threshold but none of them have high leverage. We look at the observations with a high Cook's Distance and see that their Residual Deviances are less than 2 in absolute value. Thus, none of these observations appear to be outliers. From the QQ plot of Deviance Residuals, we see that the left tail of the distribution appears to depart significantly from normality.

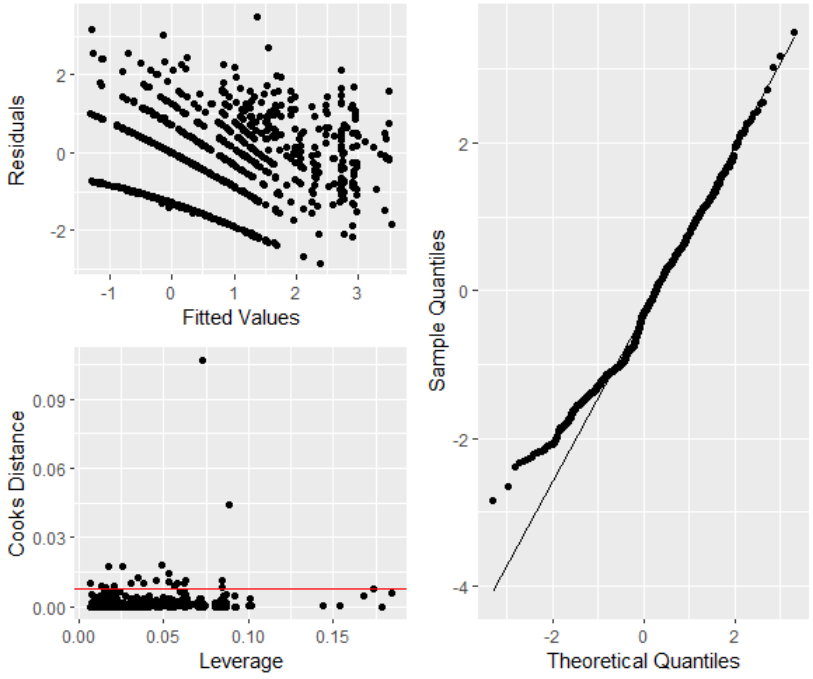

Figure 7: Diagnostic Plots for Accidents data with non-zero HGV. Top Left: Plot of Deviance Residuals vs Fitted Values. Right: QQ Plot of Deviance Residuals. Bottom Left: Plot of Cook's Distance vs Leverage

5. Conclusion

This report analyses data on Accidents, Breakdowns and Fires in French Motorway Tunnels. We analysed the data for zero and non-zero HGV separately as all such tunnels were in Urban areas. This suggested that these tunnels are a part of the arterial pathways within cities in France which facilitate traffic between residential and commercial municipalities within a city. This supposition of being a sub-class was corroborated by the results of the analysis for Breakdowns and Accidents. We could not do a similar analysis on the Fires data-set due to a lack of non-binary observation for zero-HGV tunnels.

We observe that the Traffic and Length of the tunnel are crucial factors which explain the variability in the number of Fires reported for each tunnel. We observed from the QQ plot that the data appears to be over-dispersed but we could not model it satisfactorily using a Negative Binomial model. We expect that a zero-inflated Poisson or Negative Binomial model might be more appropriate for this data-set.

For the HGV-zero tunnels, we observed that the Type of Slope in a tunnel was an important co-variate for the number of Accidents in addition to the Company operating the tunnel for both Breakdowns and Accidents. Without further analysis, the coefficients do not provide enough evidence to raise questions about the operational and maintenance practices observed the the Companies which are associated with more Breakdowns and Accidents.

For the non-zero HGV tunnels, we had to use more complicated models in that they utilised more covariates. The effect of a positive slope was clear in the model for breakdowns, whereas it had the opposite effect for Accidents. Lastly, we observe that the model for Breakdowns utilises the Traffic while the model for Accidents utilises the count of HGVs. This suggests that a review of the classification of zero and non-zero HGV tunnels would be beneficial to decrease the number of accidents.

Overall, we believe that the fit of the models was adequate and they explain the variation in the data adequately. While the model for Fires had some over-dispersion, which we could not incorporate into the model using a negative binomial model, the other data-sets were adequately modelled using it. Further directions of enquiry would be to fit zero-inflated models to each of these data-sets and asses the change in the significant covariates.

Lastly, we would like to mention that we were able to reach the same model for Breakdowns and Accidents in two different but related ways, we expect the conclusions about the important covariates and the direction of their relationships with the response variables to be robust.

References

[1] A. C. Davison. Course Notes.

[2] A. C. Davison. Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2003.

[3] Joseph M Hilbe. Negative binomial regression. Cambridge University Press, 2011.

6. Appendix

We referred to the plots for Accidents and Breakdowns in the year 2002 in the introduction. We present them here.

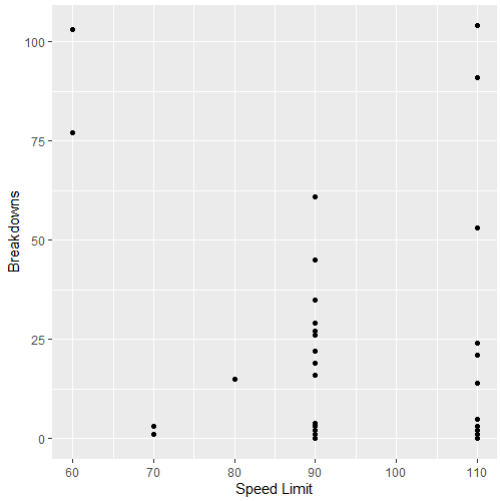

Figure 8: Scatter plot of Breakdowns vs Speed Limit in 2002

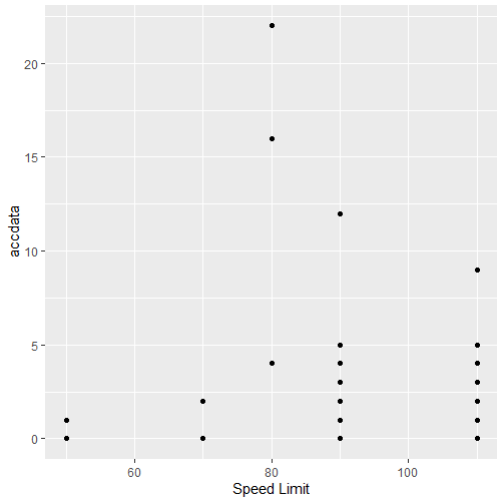

Figure 9: Scatter plot of Accidents vs Speed Limit in 2002