MATH-408: REGRESSION METHODS

---

**Group Project:**

# Regression models on count data

---

Emanuele SORGENTE, Yerkin YESBAY

30TH NOVEMBER 2022

EPFL

**Abstract**

This project aims to explore the use of regression methods on count data. This kind of data is characterized by its discrete and non-continuous nature, and there are some specific statistical models that are designed to handle it, such as Poisson regression and Negative Binomial regression. The process of model building, fitting, variable selection, and diagnostics will be conducted in order to identify the best-performing model. The final model will be selected based on its ability to accurately predict the count data and its overall performance as determined by various diagnostic measures.

# 1 Introduction

We are given three datasets containing simulated data based on events in French motorway tunnels. The dependent (target/endogenous) variables are the numbers of accidents, fires and breakdowns. The goal of the project is to analyse the dependence of target variables on the covariates and propose a prediction model.
The content of the report is as follows. In section 2 we describe the given data in more detail and provide a short exploratory data analysis. Section 3 is the main body of the report that contains the specification of the models used, argument for the use of certain models, variable selection, model diagnostics and performance analysis. In Section 4 we discuss the interpretation of the models and possible implications. In Section 5 we give a short conclusion. Section 6 is the appendix, where we provide a theoretical description of the frameworks used, as well as some additional figures and tables.

# 2 Data analysis

We are analysing three different datasets: Accidents.RData, Breakdowns.RData and Fires.RData. These datasets are partially simulated (numbers of events, tunnels and companies) and partially real data (the remaining variables).
The target variables are:

- Acc for Accidents.RData: The amount of accidents observed.
- Breakdowns for Breakdowns.RData: The amount breakdowns observed.
- Fires for Fires.RData: The amount of fires observed.

The predictors included in the datasets are:

- Tunnel: The name of the tunnel where the event took place.
- Company: The name of the company that runs the tunnel. One company may run a region containing numerous tunnels.
- Direction: Direction of traffic in which the event took place.
- Traffic:
  - For Fires.RData: Total number of vehicles passing through the tunnel accumulated over all years.
  - For Accidents.RData, Breakdowns.RData: Annual number of vehicles using the tunnel.
- HGV: Proportion of heavy goods vehicles, from 0 to 1, the latter being 100%.
- Slope: Slope of the tunnel in %.
- SlopeType: Type of slope in the tunnel.
- Urban: Indicates that the tunnel is in an urban zone.
- Type: One or two-directional tunnel.
- Limit: Speed limit in the tunnel (presumably) in km/h.
- Length: Length of the tunnel in metres.
- Just for Accidents.RData, Breakdowns.RData:
  - Year: Year when the events took place.
- Just for Accidents.RData:
  - Lanes: Number of lanes in the tunnel.
  - Width: Width of lanes in the tunnel in metres.

| | df Breakdowns | df Accidents | df Fires |
|---|---|---|---|
| Year | 10 | 10 | NULL |
| SlopeType | 5 | 5 | 6 |
| Tunnel | 91 | 92 | 92 |
| Company | 25 | 25 | 25 |
| Limit | 8 | 8 | 8 |
| Lanes | NULL | 5 | NULL |
| Width | NULL | 14 | NULL |

Table 1: Amount of values assumed by cathegorical variables

The purely categorical variables included in our datasets are: Direction, Urban, Type, SlopeType, Tunnel, Company, of which Direction, Urban, Type are binary. The quantitative variables are: Traffic, HGV, Slope, Length.
Limit, Lanes and Width appear quantitative variables, but they take just a few possible values so they can either be treated as quantitative or categorical.
In Table 1 are listed the amounts of values assumed by the categorical (non-binary) variables. We can also see from the table above, that the variable Year is specific to the dataset Breakdowns and Accidents, and the variables Lanes and Width are specific to the dataset Accidents only.
As we can see from Figure 1, there exist some outliers that can be detected visually, so we drop them from the dataframe. Specifically, it is the observation N$^o$ 62 from the dataset of the accidents, which reach the value 82. The observations N$^o$ 377,379 from the dataset of the breakdowns, taking the values 204, 198. The observation N$^o$ 149 from the dataset of the fires that reach the value 13.
In Table 2 are described the main statistics of the three dataframes (after having exluded the outliers).

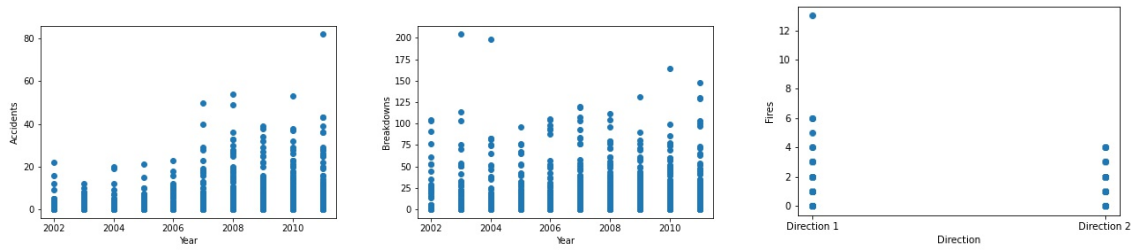


Figure 1: Scatterplot to check at first sight the outliers

| | Accidents | Breakdowns | Fires |
|---|---|---|---|
| Population | 1159 | 1040 | 169 |
| Mean | 4.11 | 16.75 | 0.75 |
| St. Deviation | 7.23 | 24.31 | 1.20 |
| Minimum | 0 | 0 | 0 |
| 25% | 0 | 2 | 0 |
| Median | 1 | 6 | 0 |
| 75% | 4 | 22 | 1 |
| Maximum | 54 | 164 | 6 |

Table 2: Main statistics of our datasets

As a consequence of the interpretation of the table above, we can say that the majority of the data are squeezed around zero. Indeed, the 25% quantile of Accidents and Fires are zero, and for Breakdowns 1. In addition the the median of Fires is still 0 and for Accidents is 1. It can be said that Fires are very scattered around zero, even the mean is less than 1 (0.75). Same results for Accidents and Breakdowns but with gradually less strengths than Fires.

Below are shown the scatterplots of the variables Company, HGV, Length, Limit, Slope, Traffic, Width compared to the target variable, number of accidents.

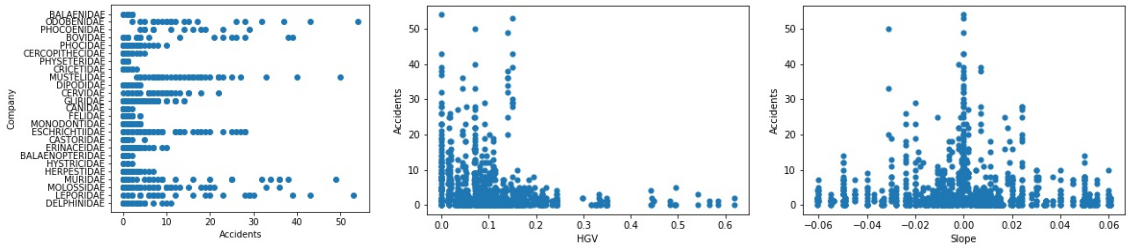


Figure 2: Scatterplots of Accidents against quantitative variables and Company

The plots above don't show very strong evidences of correlation with the target variable. However, we can see that some company tend to assume just few low values. HGV appears to be inversely related to Accidents. In the end, the plot showing the amount of accidents compared to the slope of the tunnel assume a bell shape. This can be seen since the slope become negative when it's taken the opposite direction to the positive one.
Underneath are diplayed the violinplot of the qualitative variables SlopeType and Type compared to the target variable, number of accidents.

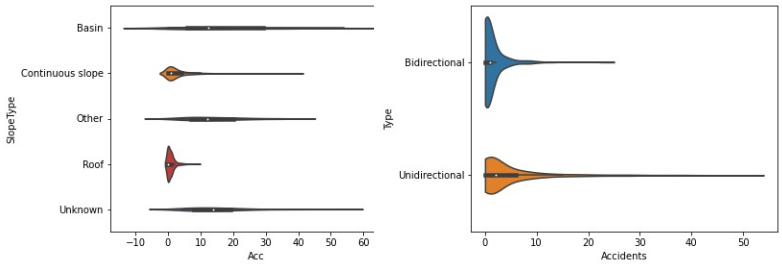


Figure 3: Violinplots of SlopeType and Type against Accidents

Even though Accidents is a discrete variable, violinplot provides smooth approximations of the discrete conditional densities, i.e. probability mass functions. One can see that these densities resemble Poisson or Negative binomial distribution.
Note that in general scatterplots are less informative than violinplots, because target variables are discrete and multiple observations are plotted in the same points, so one can not see the concentration in each point. However violinplots take reasonable space only for discrete covariates with not too many levels.

**Correlation Analysis** Below it is shown the correlation matrix heatmap of numerical variables' dependencies in the Accidents dataset.

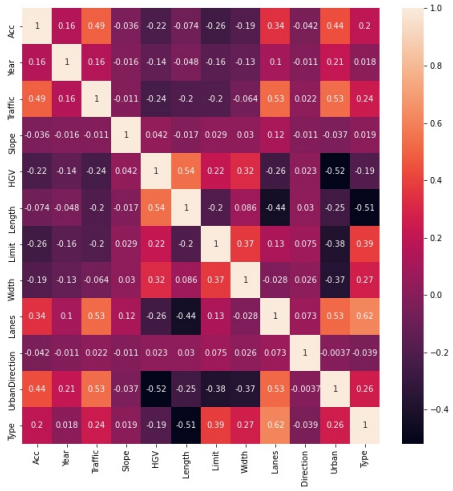


Figure 4: Heatmap of variables' linear dependencies in the Accidents' dataset

According to the heatmap above, there are not strong correlations between the covariates. Even though, we can mention slight dependencies between the variables Length and HGV, Lanes and Traffic, Lanes and Length, Urban and Traffic, Urban and HGV, Urban and Lanes, Type and Length, Type and Lanes.
As consequence, they shouldn't be considered together in the model. This consequence can be extended additionally to the other two models. Since it can be assumed that these dependencies don't vary with the dataframe.

# 3 Model

## 3.1 Model building

For each dataset we assume that the observations of target and explanatory variables are i.i.d. variables drawn from a joint distribution $(X, Y)$

$$(y_1, x_1), \ldots, (y_n, x_n) \sim (X, Y). \tag{1}$$

The main criterion for assessment of model accuracy is the (population) mean square error $\mathbb{E}(f(Y) - X)^2$. So the best prediction of $Y$ given $X$ corresponds to estimating the conditional mean

$$\widehat{y} \approx \mathbb{E}(Y|X). \tag{2}$$

For the sake of interpretability we stick to parametric models, specifically GLM. The models we consider for the analysis are: Poisson regression model and Negative binomial model, specifically NB2 model. The reason for such choice is the nature of the data: $y$ represents the number of events (e.g. accidents) of certain type appeared in presumably independent large number of trials (cars passing through a tunnel). We do not consider Binomial regression due to extremely small probabilities $p$ of the events and large number of trials $n$ (i.e. traffic), as a minor error in estimation of $p$ for $Binom(n, p)$ leads to significant error in the prediction $np$.

The Negative Binomial regression is a generalization of the Poisson regression as mentioned in the Appendix. It allows to take into account the effect of overdispersion, that may occur due to unobserbed heterogeneity.

To select the most useful explanatory variables to include in the model, we used K-fold cross validation with $K = 5$. A regression model makes sense only if it has better test performance than a constant mean estimator, i.e. if the dataset $D$ is split into $D = D_{train} \cup D_{test}$ then

$$\widehat{y}_j^{const} = \frac{1}{|D_{train}|} \sum_{i \in D_{train}} y_i. \tag{3}$$

Considering the shapes of pairplots and the values assumed by some observations we are choosing which functions of the variables are better to use. The most common transformations we used to check the models are: logarithmic, absolute value, quadratic.

To include the categorical variables in the model we considered two options:

- One-hot encoding or dummy encoding: instead of the variable $X^j$ with $k$ possible values (labels) $x^j_{(1)}, \dots, x^j_{(k)}$ we consider a set of binary variables $I\{X^j = x^j_{(i)}\}, \ \ i = 1, \dots, k$ and model the contribution of $X^j$ to the conditional mean as

$$\beta^1_j I\{X^j = x^j_{(1)}\} + \cdots + \beta^k_j I\{X^j = x^j_{(k)}\}. \tag{4}$$

- Target encoding or James-Stein encoding: we encode the label $x$ of the variable $X^j$ with an empirical estimate of $\mathbb{E}(Y|X^j = x)$, that is

$$\text{Target encoder} : X^j \mapsto \mathbb{E}_n(Y \mid X^j) = X^j_{TE}, \tag{5}$$

  where the expectation is taken w.r.t. the empirical distribution of $Y$ given $X^j$. This method is preferable for categorical variables with many levels, such as Tunnel or Company.

Note that when performing cross-validation, target encoding should be fitted on the train sample, otherwise the validation results might be biased. For categorical variables with many labels we can also group labels into bins, thus obtaining a new variable with less labels for which we can use one-hot encoding.

We used one-hot encoding only for binary variables, i.e. Direction, Type and Urban. Other categorical variables have at least 5 labels, so using one-hot encoding adds too many dimensions to the parameters and leads to overfitting or even bad convergence of the likelihood minimization algorithm. Thus, we used target encoding for SlopeType and Company.
Taking into account the link function used in both Poisson and Negative Binomial models, it is reasonable to take logarithms of the target-encoded variables.

$$X^j_{TE} \approx \mathbb{E}(Y \mid X^j) \tag{6}$$

$$\log(\mathbb{E}(Y \mid X)) \approx \beta_j \log(X^j_{TE}) + \sum_{i \neq j} \beta_i X^i \tag{7}$$

Company can not be successfully included in the model for Fires even with target-encoding or grouped one-hot encoding. The low ratio of observations per label (169/25) does not allow for a reasonable estimation of the conditional mean $\mathbb{E}(y \mid X^j = x)$ or grouping labels into bins based on the effect on Fires. For the same reason Tunnel can not be included in any model, as there are 92 different tunnel names and up to 1159 observations.

Usefulness of each explanatory variable for a certain model was primarily assessed by comparing the prediction mean square error (MSE) before and after adding it to the model. If the accuracy on K-fold

cross-validation decreased or didn't show at least minor improvements after adding the variable, the latter was decided to be insignificant. To ensure the correctness of the decision, for most variables the K-fold cross-validation accuracies were estimated 3 times, each time fixing the random seed and applying random permutation to the indices. This is to reduce the influence of a particular train-test split on the decision.
Variables that showed very small improvements in cross-validation were then closer examined with statistical tests for significance.
After the final models were defined we ranked the included explanatory variables the from most to the least significant, again by measuring how much each variable improved cross-validation prediction MSE, when being added as the last one.

## 3.2 Useful variables based on cross-validation

| Importance rank | Breakdowns | Accidents | Fires |
|---|---|---|---|
| 1 | Company | Company | log(Traffic) |
| 2 | log(Length) | log(Traffic) | log(Length) |
| 3 | Slope | log(Length) | SlopeType |
| 4 | log(Traffic) | SlopeType | HGV |
| 5 | Urban | log(Year) | Slope |
| 6 | - | log(Limit) | log(Limit) |
| 7 | - | Type | - |
| 8 | - | Width | - |

Table 3: Useful variables ranked by significance

In Table 3 we listed only the useful explanatory variables ranked by their importance based on cross-validation. The resulting sets of variables and the ranks turned out to be the same for Poisson and Negative binomial models. Note that the variables Company and SlopeType were taken as logarithms of values produced by target-encoder following (7).
Width was transformed as $x \mapsto |x - 3.25|$, as it seems from the scatterplot that accident count takes higher values when Width equals $3.25$ and tends to decrease as Width goes further away from this value. More complicated transformations of Width or Lanes, such as e.g. target encoding, would lead to loss in interpretability of a numerical explanatory variable.
Urban and Type were dummy-encoded as Urban $= I\{\text{Yes}\}$ and Type $= I\{\text{Unidirectional}\}$. The variables Width, Urban and log(Limit) have shown very slight improvement of the prediction MSE for Accidents, Breakdowns and Fires datasets respectively. They may be excluded from the final models after test-based analysis, but can still be used if one wishes to get the best possible prediction accuracy.

## 3.3 Model Diagnostics

Model diagnostics is used to evaluate the model assumptions and investigate whether or not there are observations that influence more the analysis.
The only few assumptions of generalized linear models are independence of the observations from each other, and same with the errors. Since these, we can say that also limitations caused by assumptions are very few, as opposed to linear models which needs to confirm several assumptions such as linearity and homoscedasticity. The main diagnostics to analyze and visualize the fitted model are:

- Overdispersion: overdispersion occurs when the variance of the response variable is greater than the

mean. It can be checked by comparing the deviance residuals to the expected deviance residuals under the null hypothesis that the model is correct. The deviance residuals represent the difference between the observed deviance and the expected deviance, and the expected deviance residuals are calculated under the null hypothesis that the model is correct. If the deviance residuals are larger than the expected deviance residuals, it suggests that there is overdispersion, meaning that the variance of the response variable is greater than the mean.

- Goodness of fit: The goodness of fit of the model can be checked by looking at the patterns in the deviance residuals plot that can indicate a poor fit of the model to the data, such as U-shape or a funnel shape. Additionally, if there are residuals that are far away from the other residuals, it may indicate that the model is not capturing some important relationship in the data or that there are observations with extreme values that are not well captured by the model.

We use then Cook's Distance to measure the influence of each observation on the regression coefficients. The Cook's distance statistic is a measure, for each observation in turn, of the extent of change in model estimates when that particular observation is omitted. Any observation for which the Cook's distance is close to 1 or more, or that is substantially larger than other Cook's distances (highly influential data points), requires investigation.
We produced the overdispersion test [2] on the fitted Poisson models. The p-values of the test for Accidents, Breakdowns and Fires are $2.2 \times 10^{-16}, 2.2 \times 10^{-16}$ and $0.204$ respectively. Also the Negative Binomial models estimated the overdispersion parameter $\alpha$ as $0.5641, 0.7158$ and $0.077$ respectively. Thus, we conclude that Accidents and Breakdowns are overdispersed, while Fires is not. Consequently, we choose Negative binomial for the first two datasets to treat the overdispersed data correctly.
Note that the target variable $y$ for Fires equals zero in the majority of observations, in fact $\sim 58\%$ are zeros. In this case one may consider using zero-inflated Poisson model [1], which can be useful if the ordinary Poisson model underestimates the probability of zero. However, the unconditional probability masses estimated by the ordinary Poisson model fit the actual frequencies almost perfectly (Figure 15 in the Appendix), and the zero-inflated counterpart showed a slightly worse test prediction. Hence, we stick to the ordinary Poisson model for Fires dataset.
On Figures 5, 6 and 7 we show the diagnostics plots for the models in analysis. According to these plots, we can clearly see that target variables are more overdispersed in Accidents and Breakdowns than in Fires. So, this is another interpretation of the use of Poisson model in Fires. Regarding the ordered deviance residuals, they are very close to standard normal in Accidents and Breakdowns, suggesting that they have a good fit. In Fires, also the majority of the points seem to follow the gaussian distribution, not telling us anything particular to worry about. In reference to Cook's distance, none of the plots indicate possible high leverage values due to large measures, since we would reject only observations with $D \geq 1$.

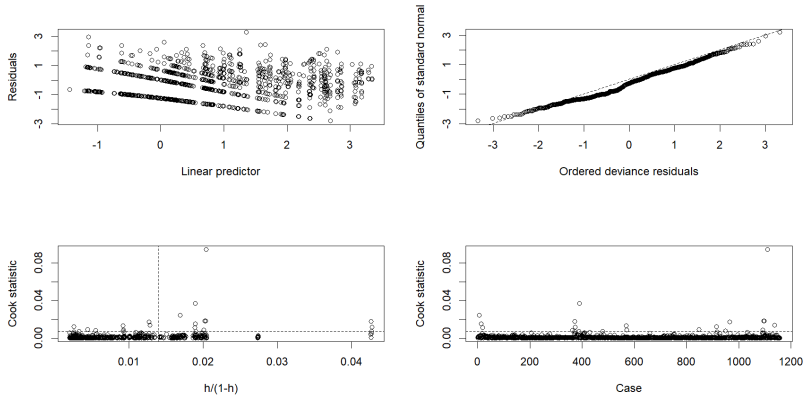


Figure 5: Diagnostics plots for Negative binomial model in Accidents

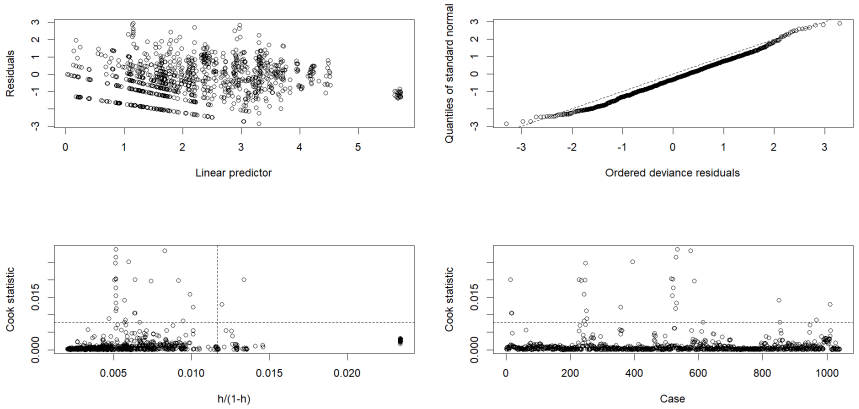


Figure 6: Diagnostics plots for Negative binomial model in Breakdowns

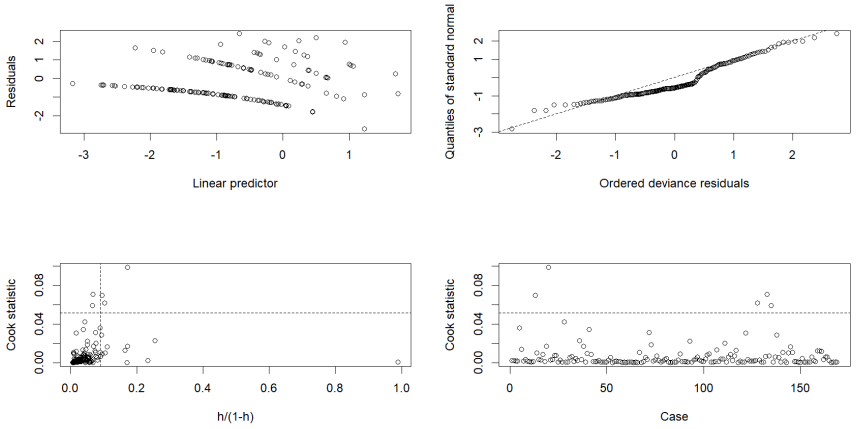


Figure 7: Diagnostics plots for Poisson model in Fires

## 3.4 Deviance Analysis

| Analysis of deviance | | | |
|---|---|---|---|
| Terms | df | Deviance reduction | p-value |
| log(Traffic) | 1 | 3451.8 | 0 |
| SlopeType | 1 | 4086.2 | 0 |
| log(limit) | 1 | 4641.4 | 0 |
| Company | 1 | 4766.9 | 0 |
| log(Year) | 1 | 4848.6 | 0 |
| log(length) | 1 | 4877.8 | 0 |
| Unidirectional | 1 | 4902.9 | 0 |
| Width | 1 | 4908.8 | 0.0155 |
| Rediduals | 1151 | 1240.2 | |

Table 4: Analysis of deviance for negative binomial model in Accidents, using $\chi^2$ test

| Analysis of deviance | | | |
|---|---|---|---|
| Terms | df | Deviance reduction | p-value |
| Company | 1 | 671.44 | 0 |
| log(Traffic) | 1 | 963.55 | 0 |
| Slope | 1 | 1180.40 | 0 |
| log(length) | 1 | 1381.76 | 0 |
| Urban | 1 | 1382.92 | 0.2821 |
| Rediduals | 1034 | 1183.2 | |

Table 5: Analysis of deviance for negative binomial model in Breakdowns, using $\chi^2$ test

| Analysis of deviance | | | |
|---|---|---|---|
| Terms | df | Deviance reduction | p-value |
| log(Traffic) | 1 | 57.57 | 0 |
| log(Length) | 1 | 86.31 | 0 |
| Slope | 1 | 93.65 | 0.0075 |
| HGV | 1 | 100.87 | 0.0067 |
| SlopeType | 1 | 108.01 | 0.0072 |
| log(Limit) | 1 | 108.95 | 0.3312 |
| Rediduals | 162 | 164.54 | |

Table 6: Analysis of deviance for poisson model in Fires, using $\chi^2$ test

We can see from Table 4 that all the predictors, before selected for Accidents' model, are relevant for Deviance reduction. In addition, looking at the p-value they all seem to be significant.
In Table 5 we can immediately detect that Urban doesn't produce a pertinent reduction for the Breakdowns' model, just $1.16$. Moreover, checking the p-value, it seems to be not significant. Then, we remove it from our final model for Breakdowns.
Finally, we look at Table 6. It can be perceived that log(Limit) appears not to produce an appropriate variance reduction for Fires, the decrease is only of $0.94$. Furthermore, even consulting the p-value it looks insignificant. So, it will be dropped in order to produce the final model for Fires.

## 3.5 Final Models

In Tables 7, 8 and 9 are shown the estimated coefficients $\widehat{\beta}$, intercepts and overdispersion parameters $\alpha$ of the final models, as well as the corresponding $95\%$ confidence intervals.

| Terms | coef | $[0.025$ | $0.975]$ |
|---|---|---|---|
| Company | 0.8105 | 0.734 | 0.887 |
| log(Length) | 0.7430 | 0.673 | 0.813 |
| Slope | 15.6657 | 13.524 | 17.808 |
| log(Traffic) | 0.4327 | 0.376 | 0.489 |
| intercept | -11.6734 | -12.824 | -10.523 |
| $\alpha$ | 0.7211 | 0.644 | 0.798 |

Table 7: Negative Binomial model for Breakdowns

| Terms | coef | $[0.025$ | $0.975]$ |
|---|---|---|---|
| log(Traffic) | 0.4203 | 0.330 | 0.511 |
| SlopeType | 0.2520 | 0.149 | 0.355 |
| log(length) | 0.4384 | 0.357 | 0.520 |
| log(limit ) | -0.7369 | -1.041 | -0.433 |
| Company | 0.4869 | 0.379 | 0.595 |
| Width | -0.6876 | -1.245 | -0.130 |
| Type Unidirectional | 0.5218 | 0.319 | 0.725 |
| log(Year) | -0.8474 | -1.078 | -0.617 |
| $\alpha$ | 0.5634 | 0.478 | 0.649 |

Table 8: Negative Binomial model for Accident

| Terms | coef | $[0.025$ | $0.975]$ |
|---|---|---|---|
| log(Traffic) | 0.6251 | 0.425 | 0.825 |
| log(Tength) | 0.5328 | 0.287 | 0.778 |
| SlopeType | 0.5640 | 0.225 | 0.903 |
| HGV | 3.4276 | 0.979 | 5.876 |
| Slope | 11.4901 | 3.071 | 19.909 |
| intercept | -16.6631 | -21.182 | -12.144 |

Table 9: Poisson model for Fires

We didn't include the intercept in the models for Accidents, as it was insignificant according to the z-test and spoils the CV prediction. On the figures below are shown 60 randomly picked observations of the target variable for each dataset and their estimated means.

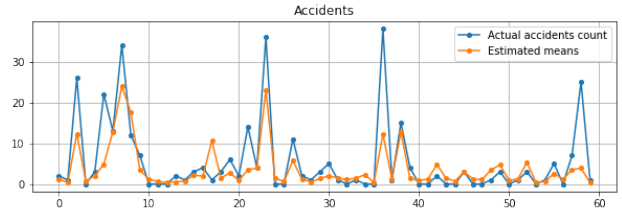


Figure 8: Fitted values vs. actual values of accidents

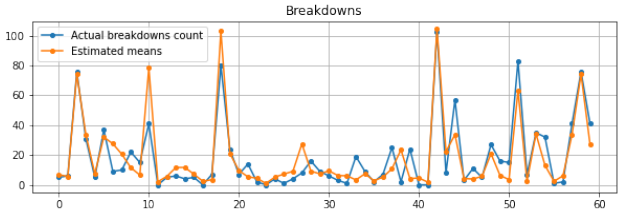


Figure 9: Fitted values vs. actual values of breakdowns

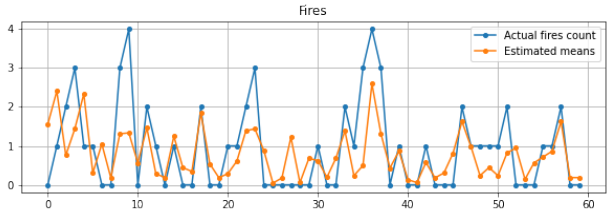


Figure 10: Fitted values vs. actual values of fires

### 3.6 Prediction

We also wish to demonstrate the predictive power of the built models. We have split the datasets into train and test subsets in $80/20$ ratio in a (pseudo-)random order and fitted the models on the train subset. On Figure 11 are shown the actual values of the target variable from the test subset and the values predicted by the models for Accidents and Fires, that were fitted on the train subset. The analogous plot for Breakdowns can be found on Figure 17 in the Appendix.

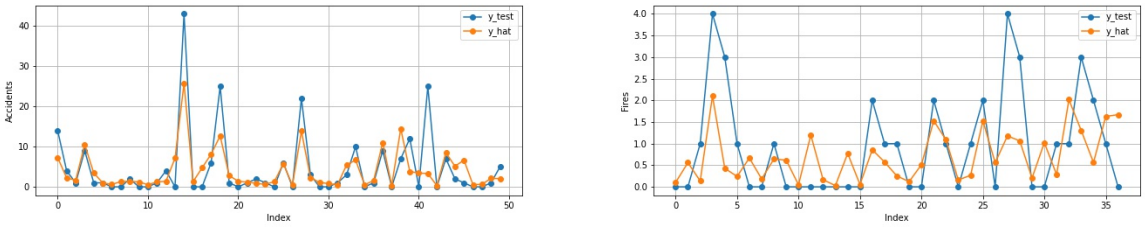


Figure 11: Actual accidents and fires counts from test subset vs predicted values

Comparing the prediction root mean square error (RMSE) on 5-fold CV of the models in question with that of the constant mean predictor (defined as in equation (3)), we have the following results:

- $4.858$ and $7.174$ for Accidents,
- $14.999$ and $24.259$ for Breakdowns,
- $0.941$ and $1.194$ for Fires.

Thus, the explanatory variables are indeed useful for prediction purposes, especially for Accidents and Breakdowns. Regarding Fires, the predictive power is not so significant as the sample size is very moderate, and some variables (e.g. Company) could not be included for such a small sample size.

## 4 Interpretation

For GLM with logarithmic link function, including Poisson and Negative binomial models, differentiation yields

$$\frac{\partial \log(\mathbb{E}(Y \mid x))}{\partial x^j} = \beta_j. \tag{8}$$

Thus the coefficient $\beta_j$ should be interpreted as how much the logarithm of the conditional mean changes when the variable $x^j$ changes by one unit. However, if $x^j = \log(\widetilde{x}^j)$ and we are interested to interpret the influence of $\widetilde{x}^j$, we have

$$\mathbb{E}(Y \mid x) = (\widetilde{x}^j)^{\beta_j} \exp\left(\sum_{k \neq j} \beta_k x^k\right), \tag{9}$$

$$\frac{\partial \mathbb{E}(Y \mid x)}{\partial \widetilde{x}^j} = \beta_j (\widetilde{x}^j)^{\beta_j - 1} \exp\left(\sum_{k \neq j} \beta_k x^k\right). \tag{10}$$

Equation (9) means that the conditional mean is proportional to $(\widetilde{x}^j)^{\beta_j}$, and equation (10) means that at any fixed point $x$ a small increment $dx^j$ of $x^j$ causes the conditional mean to change proportionally to $\beta_j dx^j$. Following this reasoning, we interpret the final models as below.

- All three target variables are proportional to some positive powers of Traffic and Length. This corresponds well with the common sense, that the more cars there is and the longer they drive, the higher are the chances for an unfortunate event.
- The number of accidents is proportional to a negative power of Limit, i.e. the speed limit reduces the rate of accidents, as it is supposed to do.
- The model for Accidents shows that the rate of accidents decreases with the year.
- Accidents are $\sim e^{0.5218}$ times more likely to happen on unidirectional tunnels on average.
- Company name has shown to be a very relevant factor for Accidents and Breakdowns, in fact it gives the largest CV MSE improvement. That means that even conditional on other explanatory variables, the rates of these two events are very dependent on the company that runs the tunnel. The values associated to the companies by target encoders allow to assume which companies' tunnels are safer conditional on other important characteristics (Table 10).
- Similarly, one can assume which slope type corresponds to higher rates of accidents and fires by looking at Table 11.
- Positive slope increases the probability of a breakdown, supposedly, because it puts more pressure on the engine. It also increases the probability of fires.
- Higher proportion of HGV is associated with the higher rate of fires.
- Keeping in mind the way we transformed the variable Width, one can roughly suppose that accidents are more likely to take place on a tunnel of medium width and less likely to happen in wider or more narrow tunnels. Although the Width's significance passes through the $5\%$ level of $\chi_2$- and $z$-tests, it produces but a tiny deviance reduction and CV MSE improvement. Thus, one needs to conduct a more domain-specific analysis to make correct inference.

Some variables may seem to have a strong association with some target variables, e.g. Urban in Accidents (Figure 13), but they are hardly significant (large p-values in z-test) conditioned on the included variables and adding them to the model leads to overfitting (CV performance gets worse).

# 5 Conclusion

In conclusion, the use of Negative binomial and Poisson regression models for analyzing count data has been shown to be effective. The predicted values from these models were well-fitted to the observed data, and the deviance and diagnostic analyses indicated that the models were a good fit for the data. Additionally, the results of these models were easy to interpret and provided valuable insights into the underlying relationships in the data. Overall, Negative binomial and Poisson regression are useful tools for statisticians working with count data.

# 6 Appendix

## 6.1 Poisson regression

Poisson regression is a generalized linear model form of regression analysis to model count data and contingency tables [3]. Poisson regression assumes that the response variable $Y$ follows a Poisson distribution and that the logarithm of its expected value can be modeled by a linear combination of the explanatory variables.
Poisson regression models are generalized linear models with the logarithm as the link function and the Poisson distribution as the assumed response probability distribution.
If $X \in \mathbb{R}^p$ is a vector of explanatory variables (covariates), then the model takes the form

$$\log(\mathbb{E}(Y \mid X)) = \beta_0 + \beta^T X \tag{11}$$

where $\beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$. This can be written more compactly as

$$\log(\mathbb{E}(Y \mid X)) = \beta^T X \tag{12}$$

where $X$ is now a $(p+1)$-dimensional vector consisting of $p$ covariates concatenated with number one, and $\beta$ is now concatenated with $\beta_0$.
If $y_i$ are independent observations with corresponding values $x_i$ of the predictor variables, then $\beta$ is estimated by maximum likelihood.

A characteristic of the Poisson distribution is that its mean is equal to its variance. In many circumstances it is found that the observed variance is greater than the mean; this is known as overdispersion and indicates that the model is not appropriate. In this case a less restrictive model called Negative Binomial regression is more appropriate [3].

## 6.2 Negative Binomial regression

The Poisson regression model can be generalized by introducing an unobserved heterogeneity term for observation $i$. Thus, the individuals are assumed to differ randomly in a manner that is not fully accounted for by the observed covariates. This is formulated as

$$\mathbb{E}(y_i \mid x_i, \tau_i) = \mu_i \tau_i = e^{x_i^T \beta + \epsilon_i} \tag{13}$$

where the unobserved heterogeneity term $\tau_i = e^{\epsilon_i}$ is independent of the vector of covariates $x_i$.
The most common implementation of the Negative binomial model is called NB2 [2]. It is derived from the assumption that the unobserved heterogeneity $\tau_i$ has a one-parameter gamma density $g(t;\theta) = t^{\theta-1}e^{-t\theta}\theta^\theta/\Gamma(\theta)$ with $\mathbb{E}\tau_i = 1$ and $\mathbb{V}(\tau_i) = 1/\theta = \alpha$. Averaging over $\tau_i$ one has

$$\mathbb{E}(y_i \mid x_i) = e^{x_i^T \beta}, \tag{14}$$

$$\mathbb{V}(y_i \mid x_i) = \mu_i(1 + \mu_i/\theta) = \mu_i(1 + \alpha\mu_i) > \mathbb{E}(y_i \mid \mathbf{x}_i). \tag{15}$$

The conditional variance of the negative binomial distribution exceeds the conditional mean. The overdispersion results from the neglected unobserved heterogeneity.
The Poisson distribution is a special case of the Negative Binomial where $\alpha = 0$.

## 6.3 K-Fold Cross Validation

[4] Cross-validation (CV) is a way to asses the prediction capacity of a model by splitting the data set into several fold). It can be naturally used to choose between several candidate models (choosing the one with

the best prediction capacity). A universal approach to do cross validation is the K-fold CV.
The data set is first randomly split into $K \in \mathbb{N}$ subsets (folds) of approximately equal size. That is $J_k \subset \{1, \ldots, N\}$ for $k = 1, \ldots, K$ such that $J_k \cap J_{k'} = \emptyset$ for $k \neq k'$ and $\bigcup_{k=1}^{K} J_k = \{1, \ldots, n\}$ are created by dividing a random permutation of the data indices $\{1, \ldots, n\}$.

For a model $m$ it becomes:

$$CV_K(m) = K^{-1} \sum_{k=1}^{K} |J_k|^{-1} \sum_{n \in J_k} \left(Y_n - m^{(-J_k)}(X_n)\right)^2, \tag{16}$$

where $m^{(-J_k)}$ is the model fitted without the data in the $k$-th fold $J_k$. The model with the smallest $CV_K(m)$ is then chosen.

Computationally, $K$-fold CV requires every candidate model to be fit $K$-times. Since $K$ is usually taken small, such as $K = 5, 10$, this mostly poses no issues.

## 6.4 Violin Plot

A violin plot is a method of plotting numeric data. Violin plots are similar to box plots, except that they also show the probability density of the data at different values, usually smoothed by a kernel density estimator (k). The violin plot include all the summary's data that is in a box plot: a marker for the median of the data; a box or marker indicating the inter-quartile range.
A violin plot is more informative than a plain box plot. While a box plot only shows summary statistics such as mean/median and interquartile ranges, the violin plot shows the full distribution of the data.
In case of counting data, as ours, is needed a $cut = 0$, since the values begin at 0.

## 6.5 Heatmaps of dataframes

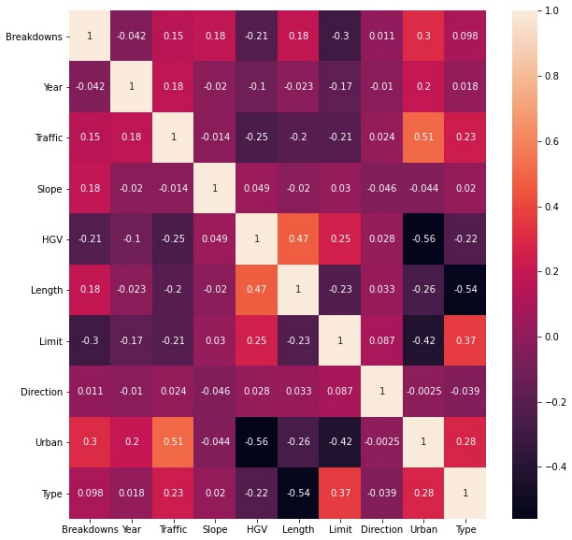


Figure 12: Heatmap of variables' linear dependencies in the dataframe of Breakdowns

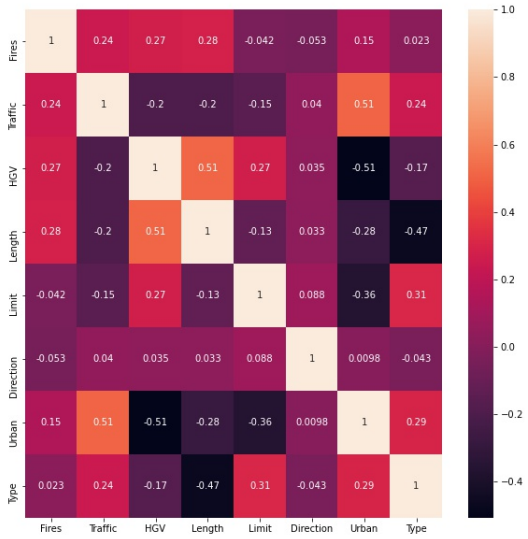


Figure 13: Heatmap of variables' linear dependencies in the dataframe of Fires

## 6.6 Violin Plots

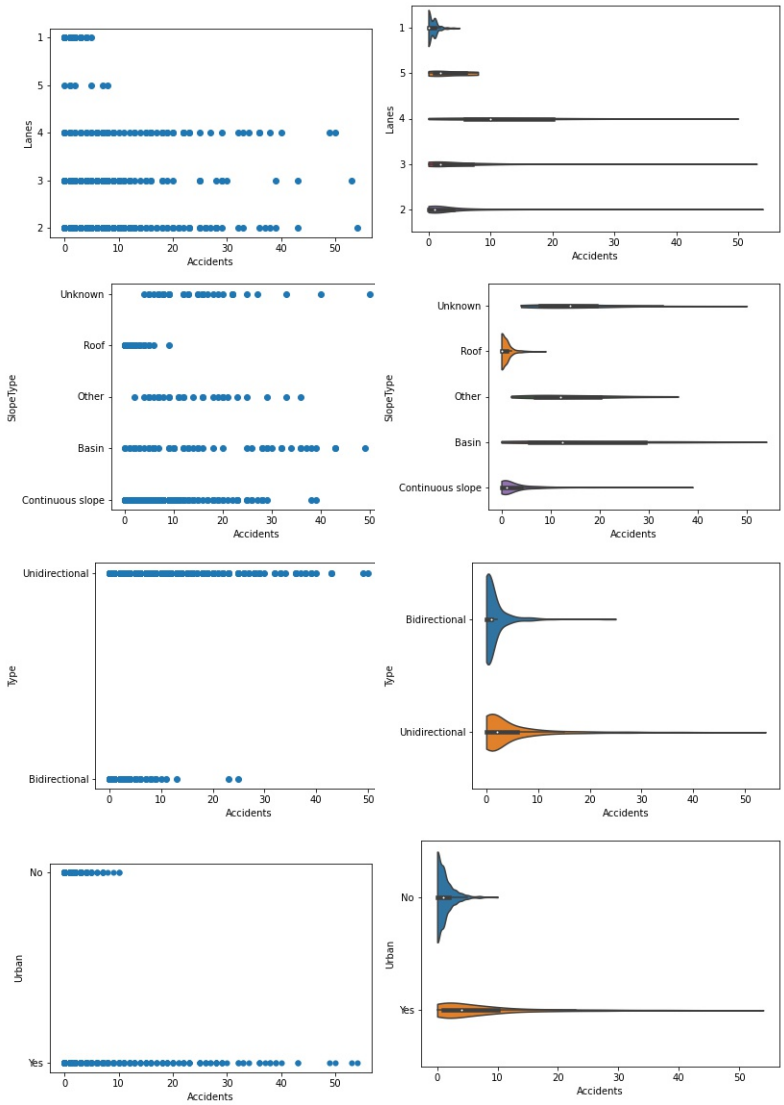


Figure 14: Scatterplots and Violinplots of Accidents against categorical variables

## 6.7 Numerical encoding of Company and SlopeType

| Accidents | | Breakdowns | |
|---|---|---|---|
| Company name | Associated value | Company name | Associated value |
| CANIDAE | 0.3 | CASTORIDAE | 2.062 |
| BALAENOPTERIDAE | 0.5 | HYSTRICIDAE | 3.166 |
| PHYSETERIDAE | 0.5 | CANIDAE | 3.625 |
| DIPODIDAE | 0.622 | DIPODIDAE | 4.777 |
| CASTORIDAE | 0.625 | PHYSETERIDAE | 4.812 |
| CRICETIDAE | 0.65 | DELPHINIDAE | 5.266 |
| FELIDAE | 0.82 | BALAENIDAE | 5.857 |
| HYSTRICIDAE | 0.833 | MONODONTIDAE | 6.79 |
| MONODONTIDAE | 1.014 | MOLOSSIDAE | 7.187 |
| CERCOPITHECIDAE | 1.05 | HERPESTIDAE | 8.203 |
| BALAENIDAE | 1.083 | FELIDAE | 9.5 |
| HERPESTIDAE | 1.233 | GLIRIDAE | 10.81 |
| ERINACEIDAE | 1.604 | MURIDAE | 11.35 |
| PHOCIDAE | 1.958 | BALAENOPTERIDAE | 11.944 |
| GLIRIDAE | 2.444 | CRICETIDAE | 14.4 |
| DELPHINIDAE | 2.833 | ERINACEIDAE | 17.062 |
| MOLOSSIDAE | 6.078 | LEPORIDAE | 19.906 |
| CERVIDAE | 7.523 | PHOCIDAE | 23.812 |
| ESCHRICHTIIDAE | 9.604 | MUSTELIDAE | 34.529 |
| MURIDAE | 10.0 | CERVIDAE | 37.690 |
| BOVIDAE | 11.476 | ESCHRICHTIIDAE | 38.0 |
| PHOCOENIDAE | 13.166 | CERCOPITHECIDAE | 41.0 |
| MUSTELIDAE | 14.312 | BOVIDAE | 42.076 |
| LEPORIDAE | 14.677 | ODOBENIDAE | 62.529 |
| ODOBENIDAE | 16.521 | PHOCOENIDAE | 89.3 |

Table 10: Company encoding

| Accidents | | Fires | |
|---|---|---|---|
| SlopeType | Associated value | SlopeType | Associated value |
| Roof | 0.837 | Roof | 0.476 |
| Continuous slope | 3.056 | Continuous slope | 0.521 |
| Other | 14.464 | Other | 1.5 |
| Unknown | 15.527 | Constant slope | 1.5 |
| Basin | 17.714 | Unknown | 1.875 |
| | | Basin | 1.923 |

Table 11: SlopeType encoding

## 6.8 Other figures

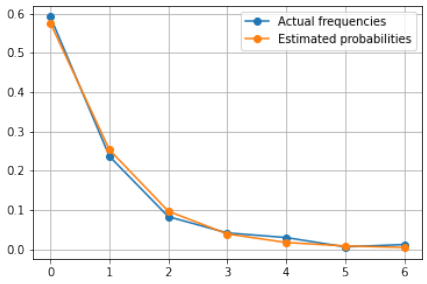


Figure 15: Estimated fire probabilities vs actual frequencies

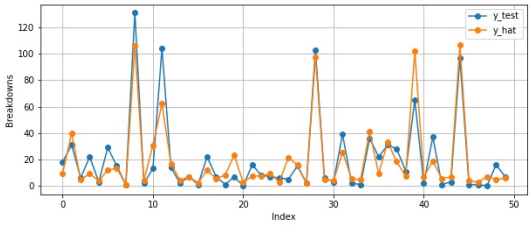


Figure 16: Breakdowns values on test subset vs predicted values

# References

[1] Cameron, A. C. and Trivedi, P. K. (2005). Microeconometrics: methods and applications. Cambridge university press.

[2] Cameron, A.C. and Trivedi, P.K. (1990). Regression-based Tests for Overdispersion in the Poisson Model. Journal of Econometrics, 46, 347–364.

[3] Davison A. (2005). Notes from Regression Methods' lecture. EPFL, pp. 159-173.

[4] Masak T. (2022). Notes from Statistical Computation and Visualization's lecture. EPFL, `https://github.com/TMasak/StatComp/blob/master/Notes/06_CV.Rmd`