Road Crashes in Milan and Severity Prediction

Matteo Silvio Mario Mello Grand, Riccardo Scibetta

Abstract
Crashes are a non-negligible problem in metropolitan cities: they can cause traffic, damages to the vehicles, or in the worst case even injuries and deaths. As citizens and drivers in the city of Milan, we decided to investigate which are the most common characteristics of crashes in our city and which factors influence their severity. As severity, we mean the number of injured people and the fatality of the crash. The aim of the project is exploring the distribution of crashes and creating two models: one that predicts whether a crash is going to be fatal, and another that predicts the number of injured people.

Dataset

To study road crashes in the city of Milan we looked for the most complete dataset available. We wrote to ISTAT asking for a dataset with all the details of the crashes that occurred. ISTAT replied that some data are protected by Italian privacy Law No. 675/1996 and GDPR Article 9, since information like exact date, exact age and exact coordinates could lead to identification of the people involved in the accident. Nonetheless, they have pointed us to a dataset with all legally available information on "Road crashes with injuries to people" in the year of 2022 on the Italian national territory. The dataset that was initially provided included 165.889 observations (road crashes) of 120 variables. It must be noted that all the observations in the dataset are crashes which caused at least one injury.

0.1 Data Cleaning

First, we filtered for the crashes that took place in the territory of the municipality of Milan, getting 7783 crashes. Nonetheless, 120 variables were still too many for the kind of analysis we envisioned, so we decided to keep only the relevant ones. The variables of the reduced dataset can be divided in the following groups:

- **Temporal Information**: Variables "day", "hour", and "trimester" capture when the accident occurred.
- **Accident Context**: Variables "accident_location", "type_of_street", "flooring", "intersection_type", "road_surface_conditions", and "meteorological_conditions" describe the location and the external conditions of the crash, while "nature_of_the_accident" classifies its type.
- **Vehicle Details and Circumstances**: Includes the type ("type_vehicle_a" and same for B and C), the registration year ("vehicle_a_registration_year" and same for B and C), the release year ("vehicle_a_release_year" and same for B and C), the cubic capacity ("cubic_capacity_a" and same for B and C), and the circumstances ("vehicle_a_circumstances_1", "vehicle_a_circumstances_2" and same for B) of the vehicles.
- **Driver Information**: Includes the age, sex, license status, and outcome for drivers of vehicles (e.g., "vehicle_a_driver_age", "vehicle_a_driver_sex", "vehicle_a_driver_licence", "vehicle_a_driver_outcome" and corresponding variables for vehicles B and C).
- **Casualties and Fatalities**: Includes "injured_pedestrians", "pedestrians_fatality", "tot_deaths_within_30_days", and "tot_injured", providing details on the crash's severity.

Most of the variables are categorical, with multiple levels. The meaning of each possible value of these variables is specified in the metadata file that came together with the dataset. Due to the great number of possible values for each one we do no explain them thoroughly here but will simply refer to them and explain them when it is needed. Many variables are not complete, in the sense that some observations might be missing. How we handled missing data will be addressed later in the analysis. Furthermore, for our analysis we created some new variables:

- **Age of the drivers**: originally a categorical variable, we created a numerical second version where

each datapoint is the mean of the extremes of its level, rounded by excess. For example the range [20, 40] becomes 30.

- **Year of registration**: created with the same method of age of drivers.
- **Fatality**: binary variable with value 1 in case the crash was fatal within 30 days for at least one individual.
- **"cos_hour" and "sin_hour"**: two continuous variables. It is necessary to create these two variables since directly looking at the hour as a numerical variable would misrepresent its periodical nature, as 0 and 23 would be very further apart when in reality they are adjacent. By using sin_hour = $sin((2\pi * hour)/24)$ and cos_hour = $cos((2\pi(hour))/24)$, we correctly treat the hour as periodical, enabling us to represent each hour data as if they were uniformly distributed on a circle, like a 24 hours watch.

1. Dataset Inspection

We first perform a visual analysis of how crashes are distributed inside the week. As it can be seen from the following heatmap, in the work days we can observe more clearly a concentration of the crashes around common commuting hours and in the hours in between, with the absolute maximum taking place at 8 am on Wednesdays.

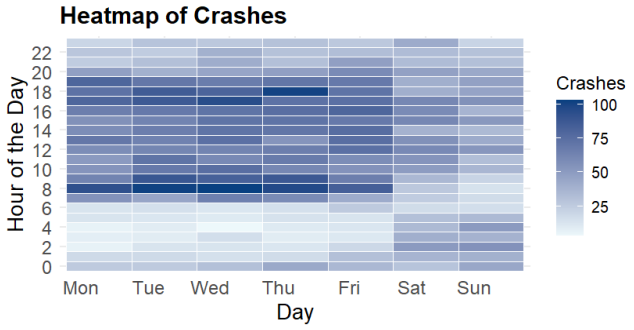

Figure 1. Heatmap of crashes across each day and hour, with color intensity representing the frequency.

It is important to emphasize that we lack information on the number of vehicles driving in the city at specific hours. Without this data, we cannot conclude that driving between 8-9 am or 5-8 pm is inherently more dangerous. Indeed, to properly assess the likelihood of a crash, we need to account for both frequency of crashes and the total number of vehicles on the road in a given time frame. Ignoring this crucial factor exemplifies the **base rate fallacy**, a cognitive bias where people focus on situational information while neglecting the broader statistical context, such as the overall base rate of events in the population.

For what regards personal information on the drivers, a first important aspect regards sex: out of 13236 drivers whose sex is known, 10384 were male and 2852 were female. Furthermore, excluding the 734 drivers for which age information is not present, we can also investigate the relative proportion of age ranges for both sexes.

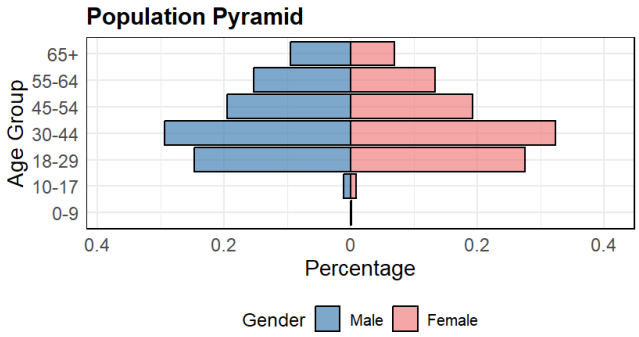

Figure 2. Population pyramid illustrating the relative distribution of crashes in each sex across different age groups.

Lastly, we also went to see which vehicles are most often involved in a crash. We see that the **type of vehicle** by far more involved in crashes are automobiles, representing 53,4% of all involved vehicles, followed by motorcycles and mopeds (25,3%), bicycles (8,1%), trucks and heavy good vehicles (4,8%) and electric scooters, which recently sparked some public debate, at 3,9%. Correlated to the vehicles, we also plotted the distribution of the cubic capacities of the vehicles involved, discovering for example that the most frequent range of cubic capacity is 100-200 and that generally it is very rare to find a vehicle with more than 2000 cc in an accident.

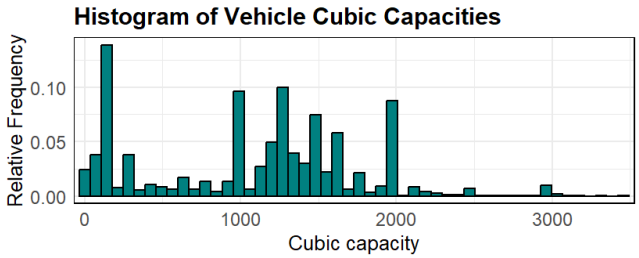

Figure 3. Histogram of cubic capacities for 8,241 vehicles, excluding 80 outliers within the range [3,500–99,000].

2. Multiple Linear Regression

Next we want to build a model that is able to predict the number of injured people in a crash given its data and characteristics. Specifically, we use linear regression to investigate whether there is a linear correlation between the variables in the dataset and the number of injured people. We use as dependent variable Y the total number of injured people, called "tot_injured". In order to apply linear regression, we must assume that given X = $(x_1, x_2, \ldots, x_n)$, Y is normally distributed and a continuous random variable. A problem in using linear regression for estimating the number of injured people is that "tot_injured" is a discrete variable, therefore it is not continuous. This could alter the normality of the residuals and their variance, breaking the linear regression assumptions. The solution we decided to adopt is to try to approximate Y as a continuous random variable, and see the results of the regression. If problems persist, then linear regression would not be the best model to use and in a further analysis we could adopt other methods, such as the Poisson regression or the Negative Binomial regression.

2.1 Model Selection

Building a linear model that contains all the variables of the dataset yields only 80 observations due to missing data. Therefore we decided to restrict the dataset to a subset of variables that are present in the majority of observations, so that we get more observations usable in the regression. In Particular, we decided to exclude all the variables related to vehicle C due to the rare presence of a third vehicle in the crash. The restriction of the dataset leads to a new dataset with 1362 observations without missing values instead of just 80. Another problem is that most of the variables inside our dataset are categorical with more than two values. Therefore an important step is transforming them into dummy variables. After converting the categorical variables into dummy ones and after excluding the dummy variables with only values 0, we end up with a model containing 58 variables. In order to select the best model, we apply the step-down and step-up methods.

Step up method We start with a model with no predictors and iteratively add the variable with the lowest p-value smaller than 0.05. After 49 iterations, the method stops with a model of 9 covariates.

Step down method We start with a model containing all the predictors, and iteratively remove the covariate with the greatest p-value greater than 0.05. The result is a model with 7 covariates.

The step down model has an R^2 of 0.1404 and an adjusted R^2 of 0.1359. The step up model has instead the following values: $R^2 = 0.1456$, adjusted $R^2 = 0.1399$. Since the R^2 and the adj R^2 of the models do not differ significantly, in order to prevent overfitting we choose the step down model which is smaller.

2.2 Analysis of results

The selected model contains the following 7 variables: "vehicle_a_driver_sex", "more_than_two_carriageways", "vehicle_a_scooter", "vehicle_b_scooter", "Sunday", "Saturday", "cos_hour". All of them are positively correlated with the number of injured people in an accident with the exception of vehicle_a_scooter and vehicle_b_scooter, which are negatively correlated. The reason for the negative correlation could be that a scooter can bring at most two people, so also the number of possible victims of the accident is lower than the one of an accident involving two cars.

2.3 Regression Diagnostics

Finally, we checked whether normality of the residuals and homoscedasticity are verified. In order to verify the normality of the residuals, we used a Shapiro-Wilk test, which rejected the null hypothesis that the residuals are normally distributed. However, since the Shapiro-Wilk test with large amounts of data detects even small deviations from normality, we paired it with a QQ-plot and an histogram. From them, we notice that the residuals follow partially a normal distribution. Particularly, for $e > 1$ the observed values strongly deviates from a normal distribution.

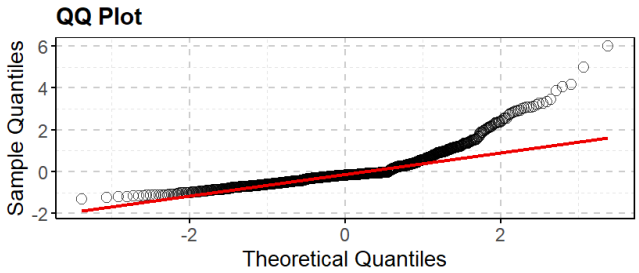

Figure 4. QQ plot of regression residuals compared to theoretical normal quantiles.

At the same time, in order to verify whether the variance of the residuals is constant, we plotted the values of the residuals with respect to the fitted values.

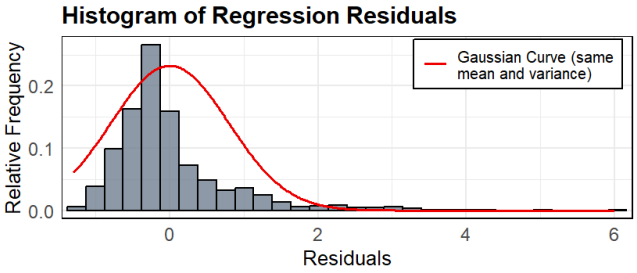

Figure 5. Histogram of regression residuals showing relative frequencies, and an overlayed Gaussian with same mean and variance as the residuals.

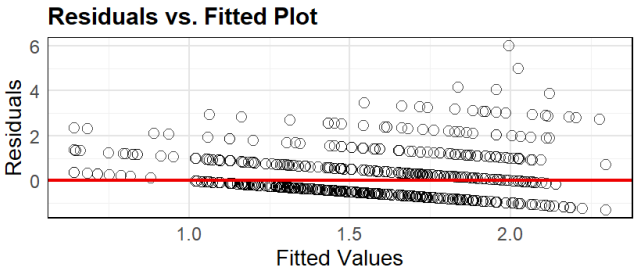

Figure 6. Scatterplot of residuals versus fitted values.

It is possible to note that the variance is not constant. We can therefore conclude that our model does not have normality of the residuals nor homoscedasticity. A reason for this result could be that, as we initially introduced, the dependent variable is not continuous but discrete. Particularly the violation of both normality of the residuals and of homoscedasticity and the very low R^2 of the model suggests to us that a linear model is not the best solution.

3. Does speeding lead to more injuries?

Now let's test a common thought hypothesis: is it true that exceeding speed limits leads to more injured people during an accident? Let n = 805 be the number of crashes in which a vehicle exceed limits of speed, and let m = 6957 be the number of crashes in which no speeding is registered. Since n and m are big enough, we can use a one-sided asymptotic t-test. It is preferred to the use of a two-sample t-test since it does not require normality of the data or that the variances of the two samples are the same. We test whether the mean number of injured people in speeding-related crashes is greater than in non-speeding crashes. The null hypothesis and alternative hypothesis are:

H_0 Crashes in which speed limits are violated do not lead to higher levels of injured people.

H_1 Crashes in which speed limits are violated lead to higher levels of injured people.

By performing the test, we get a p-value of 0.002181, which is smaller than the conventional size of a test 0.05. Therefore we conclude that we can reject the null hypothesis.

4. Logistic Regression

We now try to build a classifier to determine which road crashes will lead to fatalities, the most tragic of the possible consequences of a crash. In order to build the model, we decided to use logistic regression. Let Y = 1 when the crash is fatal and Y = 0 otherwise. Firstly, we want to select the most relevant variables for the model. A problem arises due to the fact that our dataset is quite unbalanced with regards to the crashes that result in fatalities, as they represent the 0,56% of all instances. When faced with imbalanced datasets, the model tends to be biased towards the majority class, as arbitrarily predicting all observations as non-fatal would yield a +99% accuracy. As a result, it may have lower sensitivity for the minority class, which we do not want since we are particularly interested in predicting which ones will result in fatalities. There are different techniques to handle this type of problem:

- **Undersampling the minority class**: Randomly remove samples from the majority class to balance the dataset.
- **Oversampling the minority class**: randomly duplicating samples from the minority class or generating synthetic data using SMOTE.
- **Using class weights**: Assign higher weights to the minority class during model training to penalize misclassification.

Since undersampling inevitably leads to not using most of our data, and oversampling is quite prone to overfitting, we decided to go with the weighting technique. Its drawbacks are possible model instability and difficulty in tuning the weights, which could thus lead to overfitting. We initially choose weights inversely proportional to class frequency, a standard choice.

Metric Now, we want to perform a step up selection process on the independent variables. Due to the highly imbalanced dataset we were not able to properly evaluate the model with standard metrics, like p-value and

AIC/BIC, as they do not prioritize the minority class. Instead, we decided to employ F1 score, which is the harmonic mean of precision and recall, thus being more sensitive to the model's performance on the minority class. Optimizing for this value still yielded many false positives, so we decided to opt for Fβ score, with $\beta < 1$, a variation of the F1 score that enhances emphasis on precision.

$$F\beta = (1+\beta^2) \times \frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}} \quad (1)$$

Evaluating all the models with just one independent variable, the highest score was achieved by "accidents_location". No further addition of variables enhances model performance relative to the Fβ score. The drawback of this metric is that, being the minority class extremely rare, the selected metric can be unstable or sensitive to minor changes in predictions, possibly reducing its reliability.

Finally, we tried to improve model performance by **optimizing the weights** of the two classes of the fatality variable. To do so we created a function to perform a 5-fold cross validation on the model with custom set weights, and return the mean Fβ score. We then run the function on many permutations of weights for the classes of the model, achieving for certain combinations a maximum of circa 0.036. This result shows that further weight optimization can improve model performance. In particular, all the combinations of weights obtaining this optimal result approximately lie on the straight line $y = 0.01415 * x$, indicating that the proportion between class weights is the critical factor in determining the quality of the weights.

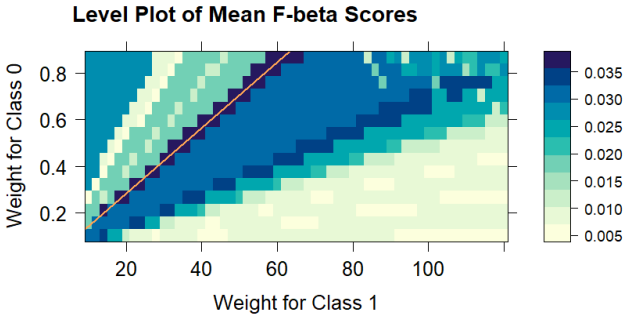

Figure 7. Level plot of Mean Fβ scores for class weight combinations, an orange line $y = 0.01415x$ intersects optimal results.

4.1 Comments on the model

In conclusion, the performance of our model was not satisfactory, as the maximum Fβ score we obtained was 0.036 when the score ranges from 0 (worse) to 1 (perfect). This might be due to multiple causes: two of them are the highly imbalanced nature of the dataset and the many missing observations for the independent variables, which meant we had actually little information of the minority class we were trying to predict.

5. Conclusion and possible other research questions

The analysis we conducted was ambitious and the dataset quite complex. As a result, the outcomes leave room for improvement. Summarizing the work done in this project, we created a linear model and a logistic model. The multiple linear regression did not work as we hoped: indeed the very low R^2 and the violation of homoscedasticity and normality of the residuals are proof that linear regression is not the best regression method for counting variables. In a further analysis, it could be interesting trying to predict the number of injured people using regression methods specifically designed for counting variables, such as the Poisson regression or the Negative Binomial regression.

Also the logistic regression did not give interesting results: indeed, the very low Fβ suggests that the model is not really precise in predicting the fatality of a crash. Possible improvements could be the use of oversampling using SMOTE in alternative or complementary to class weights for training, and the use of different models for classification such as Random Forest and Gradient Boosting.

We could also continue the study by testing other common beliefs about crashes such as "Is a crash while driving a scooter more fatal than in a car?".

Generally, the dataset could be improved by adding information about traffic, about alcohol and drug use and about the specific location of the vehicles in the city. Some of this data can be legally accessed by being an authorized researcher in a university, so it would not be difficult to deepen the study.

Lastly, another possible expansion of the project could be to redo the study for each of the last ten years and identify possible trends of the crashes in time.