Technical University of Denmark

42186 - Model Based Machine Learning

# GTD - Global Terrorism Dataset

Betlem Aguado Pérez - s182245
Ignacio Agudo Rodrigo - s185227
Philip Nielsen Butenko - s182824

Our dataset is The Global Terrorism Database (GTD). Even though is open-source, it has been downloaded from Kaggle. The GTD per se, includes information of more than 180.000 terrorist attacks that have been taken place all around the world from 1970 to 2017. The consideration of this dataset is due to its officiality, the data quality, its format and size (182k-x-135).

The research question of this project is:

**Number of possible wounded people in a terrorist attack, to provide a good, fast and efficient medical service and mobilization.**

The final attributes selected for our data-input are: *year*, *region*, *type of attack*, *type of weapon used*, *number of killed people* and finally *number of wounded people* as our target attribute.

## 1 Models Roadmapping

The next figure shows the roadmap followed for the elaboration of the project:

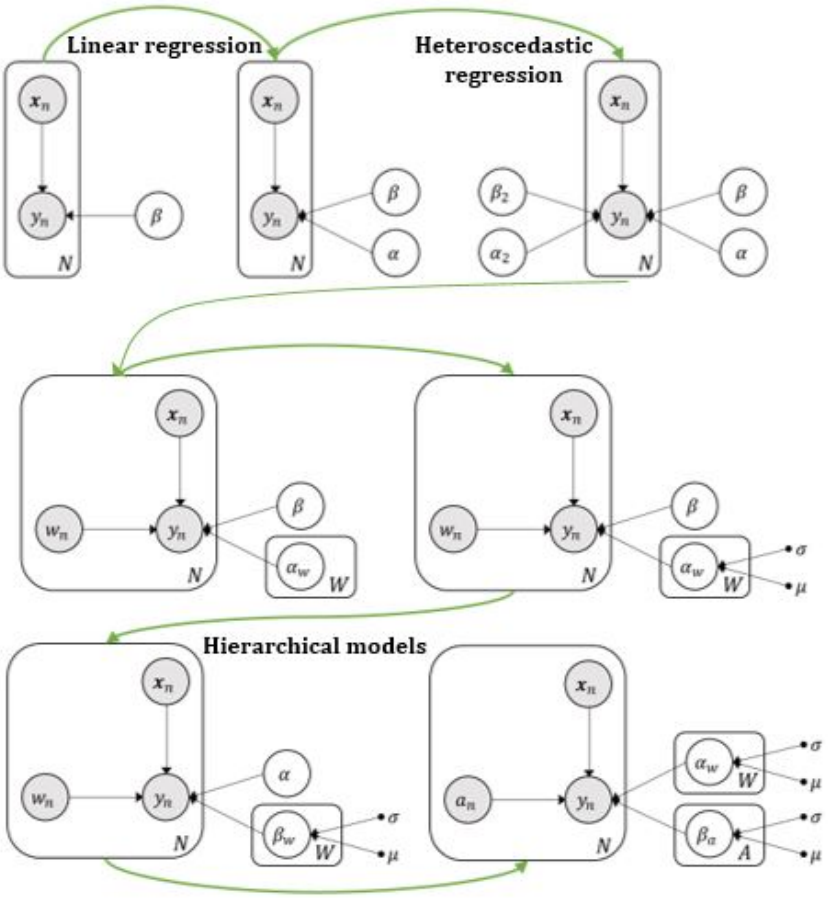


Figure 1.1: Models Roadmap

As indicated in the figure, different implementations of the model have been used for trying to improve the results. The obtained results are the ones indicated in the table:

| Linear regression | STAN alpha | STAN alpha & beta | STAN Poisson | STAN Hesteroscedastic |
|---|---|---|---|---|
| | 1min 22s | 1min 23s | 11min 32s (5000 iter.) | 1h 42min 46s |
| CorrCoef: 0.366 | CorrCoef: 0.366 | CorrCoef: 0.369 | CorrCoef: 0.017 | CorrCoef: -0.077 |
| MAE: 4.128 | MAE: 4.145 | MAE: 4.253 | MAE: 3.941 | MAE: 12.133 |
| RMSE: 16.628 | RMSE: 16.639 | RMSE: 16.572 | RMSE: 18.312 | RMSE: 21.711 |
| R2: 0.114 | R2: 0.113 | R2: 0.120 | R2: 0.000 | R2: 0.000 |

| STAN hierarchical by weapon type | | | |
|---|---|---|---|
| **Alpha** | **Alpha with priors** | **Beta with priors** | **Alpha & beta with priors** |
| 4min 14s | 3min 53s | 3min 45s | 3min 21s |
| CorrCoef: 0.255 | CorrCoef: 0.169 | CorrCoef: 0.736 | CorrCoef: -0.032 |
| MAE: 15.667 | MAE: 10.337 | MAE: 14.057 | MAE: 11.555 |
| RMSE: 26.910 | RMSE: 20.204 | RMSE: 25.219 | RMSE: 24.352 |
| R2: 0.000 | R2: 0.000 | R2: 0.000 | R2: 0.000 |

| STAN hierarchical by attack type | | | |
|---|---|---|---|
| **Alpha** | **Alpha with priors** | **Beta with priors** | **Alpha & beta with priors** |
| 3min 12s | 3min 51s | 3min 34s | 2min 56s |
| CorrCoef: 0.071 | CorrCoef: 0.181 | CorrCoef: 0.611 | CorrCoef: 0.965 |
| MAE: 20.016 | MAE: 9.607 | MAE: 10.166 | MAE: 24.495 |
| RMSE: 26.977 | RMSE: 21.892 | RMSE: 20.557 | RMSE: 32.717 |
| R2: 0.000 | R2: 0.000 | R2: 0.000 | R2: 0.000 |

Figure 1.2: Tenor IR swap table (Prepared by the authors)

Throughout all the process, it is not possible to appreciate the improvements of the different models. However, there are some aspects that can be highlighted:

- The Poisson regression obtains bad results as the events in the dataset does not occur with a known-constant rate.
- The variance of the Heteroscedastic model is really big, something that can influence the effectiveness. According with other results obtained with a more restricted variance, shows better results even though they are not good ones.
- Referring to the hierarchical models, using *attack type* as the hierarchic attribute gets better results than *weapon type*.
- Tuning the distribution of the priors has not given significative differences because of dataset. This is the main reason why its analysis is not included in the notebook.
- The best possible model cannot be determined as each trial with X_train gives very different outputs, as the relationship between the attacks is almost negligible.
- It has been confirmed, that using a bigger dataset(f.e. including the attribute *target type*) the results are much better. On the other hand, the computing time is so incredibly bigger (20 more columns to analyze) that was impossible to implement due to our computing capacity.
- In spite of being something absolutely motivational, the use of real-life datasets leads to these complicated problem. Real-life information is full of noise and difficult parameters to analyze. The proof is not even improving the models, the calculated errors did not get better.

Our first intention, was to implement a model capable of deducting the main players when an attack occurs. Looking at the selected parameters, they seemed relevant according to the target (wounded people) as for example, everybody would think that if the attack implies a bomb, the result would be much different than with a knife. Unfortunately, the results were not the expected, the noise and complexity of the dataset did not allow to obtain the predictions for materialize the improvements as is shown above.

## Workload

| **Section** | **Betlem Aguado Pérez (s182245)** | **Ignacio Agudo (s185227)** | **Philip Nielsen Butenko (s182824)** |
|---|---|---|---|
| Data preparation | X | X | X |
| Statistical Analysis | X | X | X |
| Models development | X | X | X |
| STAN | X | X | |
| Report | X | X | |