Institute of Transport Economics
Norwegian Centre for Transport Research

Summary:

Development of crash prediction models for national and county roads in Norway

TØI Report 1323/2014
Author: Alena Høye
Oslo 2014, 45 pages Norwegian language

Crash models for national and county roads in Norway were developed in order to calculate predicted numbers of injury crashes, slightly injured, seriously injured, fatalities and the total number of killed or seriously injured, as a function of traffic volume, segment length, road category, speed limit, number of lanes, number of intersections and other road characteristics. Models were calculated as generalized negative binomial models (negative binomial models with a variable overdispersion parameter). Results can be used in evaluations of road safety measures with the empirical Bayes method which is regarded as state of the art with respect to control for regression to the mean. The models can also be applied in conjunction with several tools of the Norwegian Public Roads Administration for road safety analyses and estimation of expected effects of road safety measures.

Crash models were developed on behalf of the Norwegian Public Roads Administration based on crash data in the national roads data base from the years 2006-2011. The models can be used to calculate predicted numbers of injury crashes, slightly injured, seriously injured, fatalities and the total number of killed or seriously injured (KSI) on national and county roads as a function of traffic volume, segment length, road category, speed limit, number of lanes, number of intersections and other road characteristics. Models were calculated as generalized negative binomial models (negative binomial models with a variable overdispersion parameter) which allows the calculation of expected crash numbers as a function of the model predictions, crash counts and an overdispersion parameter. Expected crash numbers refer to the number of crashes that can be expected on a road segment in the long run, based on general road characteristics and the specific crash history of the road segment in question. The statistical weights for model predictions and crash counts are calculated as functions of the overdispersion parameter. The overdispersion parameter varies as a function of traffic volume and segment length. The statistical weight for the crash counts increases with increasing segment length and traffic volume.

The expected numbers of crashes will always be between the model prediction and the observed numbers of crashes. In before-after studies of the safety effects of measures at high-crash locations, regression to the mean is less likely to affect the results when the analysis is based on expected crash numbers in the before period instead of observed crash counts. Regression to the mean occurs when the observed number of crashes in the before period was exceptionally high. One would then expect the number of crashes to decrease in the after period even without any (effective) safety measures.

Model and predictors

The models were developed as generalized negative binomial (NB) models in which the overdispersion parameter is estimated as a function of traffic volume, segment length and number of years. The model form is as follows:

$$E(n) = e^{\Sigma_i Predictor_i * Coeff_i}$$

E(n) is the predicted number of crashes (i.e. number of injury crashes / injuries / fatalities), predictors are traffic volume and a number of road characteristics, and i is the subscript for the predictors.

A generalized NB model was chosen because it takes into account that crash counts usually are overdispersed, and that overdispersion is not a constant but depending on traffic volumes, segment length and the number of years. The model results can be used in before-after evaluations of road safety measures with the empirical Bayes (EB) method. The EB method controls for regression to the mean by comparing the observed number of crashes in the after period with the expected number of crashes. The expected number of crashes is a function of the actual crash count and the model prediction of the number of crashes for the same road section and period of time. Crash counts and model predictions are weighted with a function of the overdispersion parameter. The expected number of crashes is always between the crash count and the model prediction.

The crash models for all dependent variables (injury crashes and numbers of injuries / fatalities) are based on the following predictors:

- *Segment length and number of years: Ln(segment length) and Ln(number of years).* The coefficients for the natural logarithm of segment length and number of years are set equal to one, thus normal crash numbers increase proportionally with segment length and number of years. The number of years is a predictor because segments with substantial changes (e.g. speed limit reductions) are represented with crash data only from after the change was made in the data the models are based on.
- *Traffic volume: Ln(AADT) and Ln(AADT)².* Volume predictors are the natural logarithm of AADT (annual average daily traffic) and the squared natural logarithm of AADT. Predicted crash numbers increase with increasing traffic volume, but at a decreasing pace.

- ***Speed limit: Dummy variables.*** For each speed limit a dummy variable is defined in order to take into account that crash numbers not necessarily are a monotonous (or other) function of speed limit because of general differences between roads with different speed limits (e.g. many roads with a 70 km/h speed limit had previously an 80 km/h speed limit that was reduced because of exceptionally high crash numbers).
- ***Number of lanes: Dummy variables.*** For each number of lanes a dummy variable is defined for the same reasons as for speed limit. Segments with one lane are a highly heterogeneous group of different kinds of roads and only a small proportion of all data. They are therefore omitted from the data.
- ***At-grade intersections, roundabouts and ramps (grade separated intersections): Ln(Number of … + 1).*** For four-armed intersections, three-armed intersections, roundabouts, off-ramps and on-ramps the natural logarithm of the number of intersections / roundabouts / ramps plus one is calculated (plus one in order to avoid taking the logarithm of zero on segments without any of these).
- ***Curves: Ln(Number of curves + 1) according to speed limit.*** The curve variable that was available was the number of parts of each road segment that is 50 m long and has a curve radius below 300 m. The natural logarithm of the number of curves according to this definition plus one is calculated and one such variable is defined for each speed limit (Ln(number of curves + 1) at the respective speed limit and zero at all other speed limits) in order to take into account different effects of curves on crash numbers at different speed limits. The available curve variable has only a weak relationship with the actual number of curves or other definitions of road curvature as it does not take into account turning points in road curvature.
- ***Grades: Ln(Number of grades + 1) according to speed limit.*** The grade variable that was available was the number of parts of each road segment that are 200 m long and that have a vertical grade of at least 4%. The natural logarithm of the number of grades according to this definition plus one is calculated and one such variable is defined for each speed limit (Ln(number of grades + 1) at the respective speed limit and zero at all other speed limits) in order to take into account different effects of grades on crash numbers at different speed limits. The available grade variable has only a weak relationship with the actual number of grades and contains no information about crest or sag curves.
- ***Road category: Five dummy variables.*** A dummy variable is defined for each of the following road categories: Motorway; two-lane road with grade separated intersections; TEN-T road (other than motorway or two-lane road with grade separated intersections); European or state highway (other than those previously mentioned); county road.
- ***Median and median guardrail: Four dummy variables.*** A dummy variable is defined for each of the following: Median with guardrail; median without guardrail; guardrail that separates opposing directions of traffic without median; neither guardrail nor median. Dummy variables for the presence vs. absence of either median or guardrail would not allow to detect interaction effects between guardrail and median presence, which is why four dummy variables for each possible combination were defined.

- *Center line rumple strips: Two dummy variables.* Narrow (below one meter) and wide (one meter or wider) center line rumble strips are represented in the model by dummy variables.
- *County: Dummy variables.* For each county a dummy variable is defined. These variables are meant to represent general differences between counties such as topography, weather and population density.
- *Constant:* A constant term is included in all models.

Additionally, all models contain coefficients for calculating the *overdispersion parameter* as a function of the natural logarithms of segment length, number of years and traffic volume. The overdispersion parameter decreases with increasing segment length, number of years and traffic volume.

How good are the models?

Several goodness-of-fit (GOF) indicators were calculated for each model: R^2, mean square prediction error, and Elvik-index. Additionally Cure plots were plotted. The number of injury crashes is the one with the smallest mean square prediction error.

Comparing models with different sets of predictor variables or with different functions of road characteristics (e.g. speed limit dummies vs. speed limit as a numerical variable) there are only small differences between the GOF indicators, although estimated normal crash numbers on individual road segments may differ considerably between different models. Only models with Ln(AADT) as the only predictor (in addition to segment length and number of years) are considerably weaker than models with most or all road characteristics as additional predictors.

Differences between model predictions and crash counts are for the most part only small and unsystematic. Only at high AADT (above ca. 10,000) the model predictions for numbers of injury crashes and slightly injured are systematically higher than the observed numbers. The total numbers of model predictions of crashes and injuries are at a maximum 2.5% higher (slightly injured) than the observed numbers.

Crash model spreadsheet

The attached spreadsheet Ulykkesmodeller.xlsx can be used to calculate:

- Model predictions for number of injury crashes, slightly injured, seriously injured, fatalities and KSI
- An overdispersion parameter for each dependent variable
- A statistical weight for each dependent variable that can be used to calculate the expected numbers of injury crashes, slightly injured, seriously injured, fatalities and KSI
- If observed numbers of crashes / injuries are entered, expected numbers of crashes / injuries are calculated as well.

It is also possible to convert the results to years between 1997 and 2020.