Institute of Transport Economics
Norwegian Centre for Transport Research

Summary:

Disaggregate Accident Frequency and Risk Modelling

A Rough Guide

TØI Report 1403/2015
Author: Lasse Fridstrøm
Oslo 2015, 35 pages English language

The analyst working with accident count data is fortunate. The inner logic of the data is such as to allow for unusually fruitful and efficient statistical methods. These opportunities should be exploited. We explain how.

The nature of accident data

Accident counts are non-negative integers. And so are victim counts.

This means that even before we start looking at our data, we know something about them. This *a priori* information is potentially quite valuable, and we should make sure that we do not lose it or forget it.

The implications are twofold.

First, since accident or victim counts (*casualty* counts, for short) cannot be negative, any model able to predict a negative number of casualties is necessarily leaving something to be desired. More precisely, the fact that casualty counts are non-negative numbers suggests a log-linear rather than a linear model structure. 'Log-linear' essentially means 'multiplicative'. Risk factors work multiplicatively, not additively. The risk function is a *product* of positive factors.

Second, any accident number placed *between* the integers is logically impossible. Thus, the set of possible casualty counts is much smaller than the set of real numbers. Any model that does not implicitly take account of this, is in a sense more general than necessary – in other words more vague, less precise.

This suggests that casualty counts be analyzed by statistical methods explicitly developed for count data, i. e. for non-negative, integer-valued dependent variables.

The core model for count data analysis is the Poisson regression model. The Poisson distribution has the remarkable property that the variance equals the mean. That means that, once we have estimated the expected value, we also know what to expect in terms of variation *around* the mean.

The Poisson distribution can be *generalized* into the *negative binomial* distribution. In this distribution, the data are subject to *overdispersion* compared to the pure Poisson case, i. e. a variance that exceeds the mean. Most software packages will put out the overdispersion *parameter* as part of its estimation, so you can *test* whether the pure Poisson model holds, or if your data suggest a more general formulation, such as the negative binomial distribution.

Measure exposure

By *exposure*, we mean the amount of activities that expose certain subjects to risk. Exposure is likely to be the most crucial explanatory variable in any accident model.

In many cases, exposure should be modelled as multi-dimensional. The expected number of injury accidents may, e. g., depend on passenger car miles travelled, freight vehicle miles travelled, bus passenger miles travelled, bicyclist mileage, and pedestrian mileage.

Don't worry about multicollinearity

In a regression model, independent variables are always to a smaller or larger degree collinear (correlated). Just like in real life. Indeed, the regression model is our tool to mimic this reality, in a situation where we cannot perform controlled experiments, but must rely on non-experimental data.

Yet many practitioners and econometricians believe that multicollinearity must or should be avoided. Don't listen to them.

In fact, collinearity is the very reason why we need multiple regression analysis to understand what is going on. It makes absolutely no sense to require that collinearity be avoided.

That said, it is a sad fact that when several relevant variables are collinear, it is hard to estimate their respective partial effects. The estimates will be imprecise. But this will be reflected in the estimated standard errors, the t-statistics, the p-values, etc. The regression output will tell us all there is to say about this. The problem is only as big as your reported standard errors.

Measure size only once. Make smart decompositions.

There are, fortunately, some tricks available to keep related groups of variables from obliterating each other, while also enhancing the ease of interpretation.

Take the example of vehicle size. In a data set consisting of individual accidents or vehicles, one might consider entering vehicle weight, length, height, engine effect, number of seats/doors/wheels, etc. They will all be highly correlated. More importantly, their coefficients will be hard to interpret, since they all express partial effects, conditional on all other variables being held constant.

The solution is this: enter only one variable related to vehicle size, and measure all other variables in relation to this *one* size variable. For instance, enter the log of *weight*, log of engine power *per tonne*, log of fuel consumption *per horsepower*, etc. In this way, all three variables are entered in the form of a multiplicative decomposition. All coefficients, as well as their sums and differences, will have interesting subject-matter interpretations.

Draw path diagram to avoid endogeneity

Bear in mind that the interpretation of any one coefficient in the regression model is the partial effect of changes in the corresponding variable, *conditional on all other explanatory variables being held constant.*

Therefore, it does not serve the purpose to enter two independent variables, of which one (X, say) always changes in response to another (Z). In such a case, we say that X is endogenous with respect to Z. It does not make sense to measure the partial effect of Z, given X, or vice versa, as one would actually do in a model including both variables.

To fix ideas, and keep track of any possible endogeneity present in the model, it is highly recommended to draw a causal path diagram before specifying the model, or in parallel with it.

Don't mess with your dependent variable

Accident counts have a known distribution: the (generalized) Poisson. This extremely valuable piece of information must be safeguarded and exploited.

Transformed accident counts do not, however, necessarily obey any known statistical law. When, e. g., we take the log of an accident count, we no longer know its distribution, or variance. In fact, its variance is not even finite (since the log of zero is minus infinity).

Similarly, if we use accident *rates* (casualties divided by exposure) rather than crude accident counts as the dependent variable, we no longer know the distribution or variance of the error term.

Use the crude casualty count as your dependent variable. Do not transform it, as this amounts to throwing away valuable statistical information. If you want to constrain the accident generating function in a particular way, do all your transformation on the right-hand side, i. e. on the *in*dependent variables. Then proceed to estimate by generalized Poisson maximum likelihood.

Compute the maximal goodness-of-fit

Accidents are truly random events, logically unpredictable at the disaggregate level. Think of it: if the single accident were predictable, in terms of its exact time, place, and persons involved, then it would not happen. The individual accident is as random and unpredictable as the movement of the elementary particle in quantum physics.

Thus, the bad news is that in a statistical accident model, there will always be a minimum, inevitable amount of random noise. The random noise component will be larger, relative to the systematic part, the smaller is the mean expected number of accidents.

The good news is that, since accidents counts are known to follow the Poisson distribution, we can compute the maximally obtainable fit, or the minimal amount of unexplained variation. This confers more meaning to the well-known coefficient of determination R^2 than in any other econometric application. We can tell how far we are from explaining all the variation explainable.

Apply casualty subset tests

Casualty sets may be subdivided into subsets. In many cases, accident countermeasures work because they affect one particular subset of accidents or victims. Or some risk factor is relevant only for a particular subset of subjects.

Suppose, e. g., we find that vehicles belonging to households with a male license holder aged 18-25 exhibit an increased injury accident frequency. We naturally interpret this as the effect of higher risk among male, young drivers.

To check whether this interpretation is tenable, we may run the exact same model on a smaller subset of casualties, such as ‘male car drivers aged 18-25 involved in an injury accident’. If our interpretation is correct, the effect on this casualty subset should come out stronger than in the more general (main) model. If not, one must conclude that at least part of the relationship found in the main model is spurious.

Concentrate on the systematic part of the variation

There is a wide literature on how to specify fanciful and sophisticated structures for the random disturbance term of the accident equation. Ignore it.

Working with accident counts we already have plenty information on the structure of the random error. We know that the error terms behave more or less like Poisson or negative binomial residuals. This knowledge is automatically taken account of in standard maximum likelihood estimation software.

The juice of an accident equation is in the *systematic part*, i. e. in the linear combination of coefficients and independent variables. It is the systematic part that will tell us something about accident causation. You should spend your intellect on specifying this combination rather than on the uninformative random error.