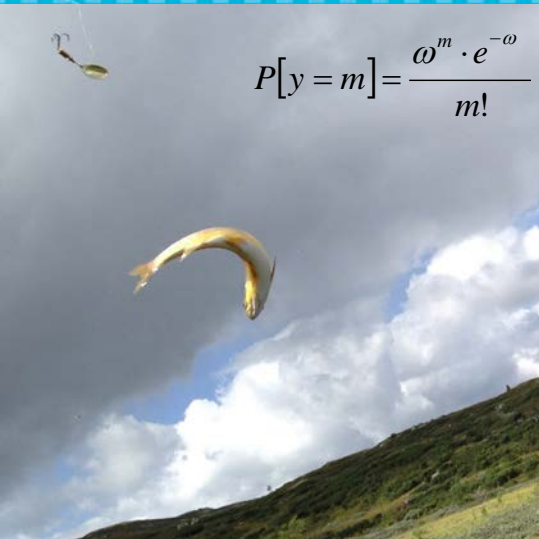
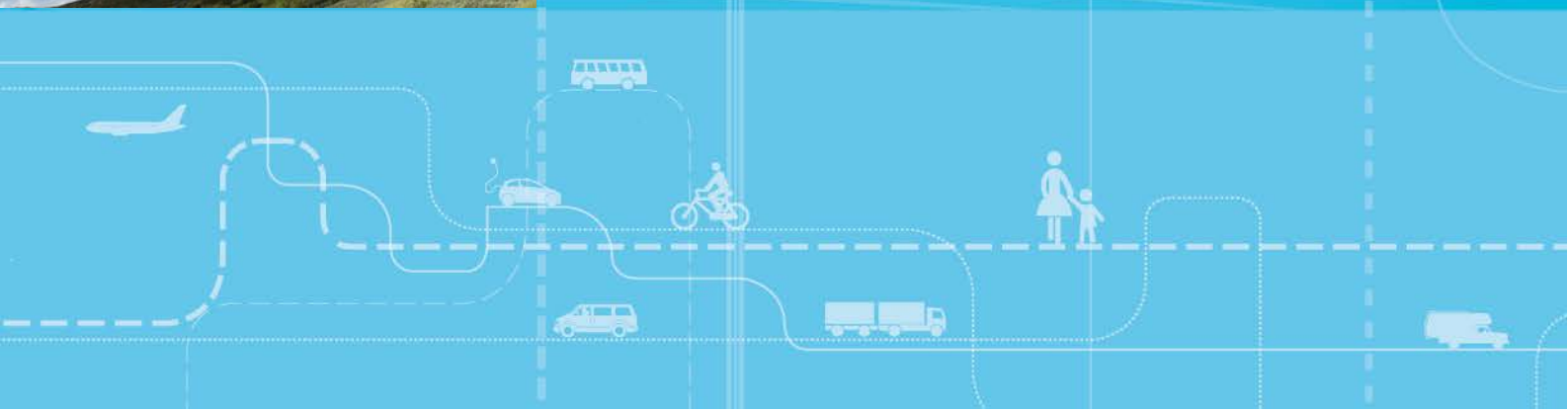
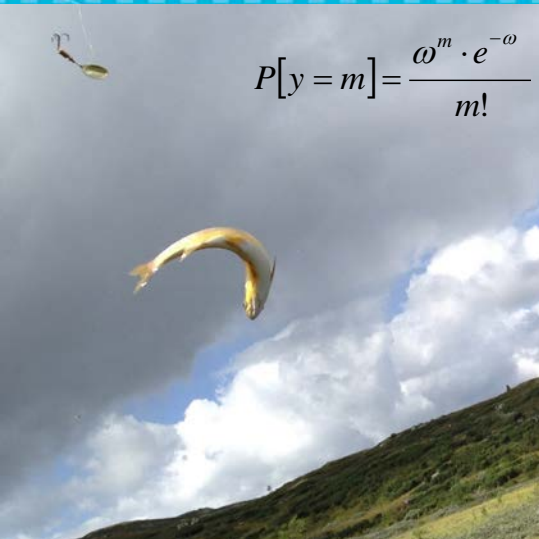


# Disaggregate Accident Frequency and Risk Modelling

A Rough Guide


$$P[y = m] = \frac{\omega^m \cdot e^{-\omega}}{m!}$$





# **Disaggregate Accident Frequency and Risk Modelling**

A Rough Guide

Lasse Fridstrøm

This report is covered by the terms and conditions specified by the Norwegian Copyright Act. Contents of the report may be used for referencing or as a source of information. Quotations or references must be attributed to the Institute of Transport Economics (TØI) as the source with specific mention made to the author and report number. For other use, advance permission must be provided by TØI.

ISSN 0808-1190

ISBN 978-82-480-1168-2 Electronic version

Oslo, March 2015

---

**Title:** Disaggregate accident frequency and risk modelling.  
A rough guide

**Tittel:** Statistisk analyse av færdselsuheld.  
En uformel vejleder

**Author(s):** Lasse Fridstrøm

**Forfattere:** Lasse Fridstrøm

**Date:** 03.2015

**Dato:** 03.2015

**TØI report:** 1403/2015

**TØI rapport:** 1403/2015

**Pages** 35

**Sider** 35

**ISBN Electronic:** 978-82-480-1168-2

**ISBN Elektronisk:** 978-82-480-1168-2

**ISSN** 0808-1190

**ISSN** 0808-1190

**Financed by:** Danish Council for Strategic Research

**Finansieringskilde:** Det Strategiske Forskningsråd

**Project:** 3685 - Road safety models for Denmark

**Prosjekt:** 3685 - Road safety models for Denmark

**Quality manager:** Rune Elvik

**Kvalitetsansvarlig:** Rune Elvik

**Key words:** Estimation  
Poisson model  
Road safety

**Emneord:** Estimering  
Poisson-modell  
Trafikkrisiko

**Summary:**

Accident count data are non-negative integers. A large part of the variation in such data is due to sheer and unexplainable randomness. There are strong reasons to believe that accident data are, at least approximately, Poisson distributed. Acknowledging this fact opens the door to an arsenal of quite efficient inference methods. We explain how these opportunities can be exploited – or missed.

**Sammendrag:**

Uheldstal er nær ved å følge den statistiske Poisson-fordeling. Denne kendsgerning kan udnyttes i analysen med sigte på beregning af risiko. Vi forklarer hvorledes, idet vi også giver råd og advarsler om hvilke feil man ikke skal gjøre, hvad enten man studerer små eller store uheldstal.

Language of report: English

---

*This report is available only in electronic version.*

*Rapporten utgis kun i elektronisk utgave.*

---

*Institute of Transport Economics  
Gaustadalleen 21, 0349 Oslo, Norway  
Telefon 22 57 38 00 - [www.toi.no](http://www.toi.no)*

*Transportøkonomisk Institutt  
Gaustadalleen 21, 0349 Oslo  
Telefon 22 57 38 00 - [www.toi.no](http://www.toi.no)*

# Preface

The IMPROSA project for the Danish Council for Strategic Research aims at **Improving Road Safety – Developing a Basis for Socio-Economic Prioritising of Road Safety Measures**. The Institute of Transport Economics (TØI) has participated in the project as an advisor on statistical and econometric modelling.

The present report is a slightly amended version of TØI working paper 50268, dated 11 December 2012. Its aim is to gather and present relevant advice for the econometric practitioner working with disaggregate or meso-aggregate accident data.

The appendices are all excerpts from the author's doctoral dissertation (TØI report 457/1999).

Thanks are due to Marc Gaudry for his insightful comments on a draft version of the report, and to Michael W. J. Sørensen for his Danish language laundering.

The project manager at TØI has been Senior Research Economist Lasse Fridstrøm. He has also authored the report. The quality assurance at TØI has been assumed by Chief Research Officer Rune Elvik. The final editing of this report was done by Trude Rømming. The front-page photo, entitled 'Méthode poisson-klingenbergeoise', was taken by Sissel H. Klingenberg.

Oslo, March 2015

Institute of Transport Economics

*Gunnar Lindberg*  
Managing Director

*Rune Elvik*  
Chief Research Officer



# Contents

## Summary

### Resumé

1	Introduction and overview .....	1
2	Keep track of your unit of analysis.....	2
3	Exploit the information inherent in the inner logic of casualty counts .....	3
4	Use econometrically efficient methods.....	5
5	Make the analysis as general as possible.....	7
6	Measure exposure .....	8
7	Preserve high levels of measurement.....	9
8	Measure size only once. Make smart decompositions. ....	10
9	Start with simple models, then add detail and sophistication.....	12
10	Avoid endogenous right-hand side variables. Draw path diagram .....	13
11	Leave left-hand side intact. Do all transformations on right-hand side...	15
12	Beware of accident underreporting .....	16
13	Don't worry about multicollinearity.....	17
14	Compute the maximal goodness-of-fit .....	18
15	Apply casualty subset tests .....	19
16	Concentrate on the systematic part of the variation .....	20
	Literature.....	21
	Appendix A. The Poisson process .....	23
	Appendix B. The generalized Poisson distribution .....	27
	Appendix C. Specialized goodness-of-fit measures for accident models .....	30
	Appendix D. Casualty subset tests .....	32





**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 *independent* 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 counter-measures 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.

## Resumé:

# Statistisk analyse af færdselsuheld

## En uformel vejleder

TØI rapport 1403/2015  
Forfatter: Lasse Fridstrøm  
Oslo 2015 35 sider

*Uheldsanalysen er en fascinerende gesjæft. Uheldstallenes iboende natur giver ophav til et ualmindelig rigt og træfsikkert arsenal af statistiske metoder. Det gælder bare at udnytte dem.*

### Uheldstallenes iboende natur

Antal færdselsuheld er med nødvendighed et ikke-negativt heltal. Det gælder hvad enten vi tæller op uheldene i hele kongeriget i løbet af et helt år, eller kun angiver, hvor mange uheld én bestemt person var udsat for i sidste uge.

Når vi skal analysere forekomsten af færdselsuheld, ved vi altså, at vores model for dette ikke skal give rum for udfald, der ikke er heltal (0, 1, 2, 3, ...). Den skal heller ikke kunne give negative udfald.

Det betyder i realiteten, at sammenhængen mellem uheldstal og forklarende faktorer ikke kan have form af en sum. Det forventede uheldstal skal være et *produkt* af positive faktorer.

Sagt på en anden måde, skal regressionsmodellen ikke være lineær i variableerne, men log-lineær. De uafhængige variabler skal som hovedregel være målt på en logaritmisk skala.

Det enkelte færdselsuheld rammer tilfældigt og uforudsigeligt. Om uheldet havde været forudsagt, med nøjagtig sted, tid og involverede personer, skulle det slet ikke have sket. Således er det logisk umuligt at forudsige det enkelte uheld. Uheldstallene er behæftet med en statistisk tilfældighed lige så fundamental som kvantefysikkens elementærpartikler. Niels Bohr skulle nok have nikket genkendende.

Teoretiske udlægninger så vel som erfaring har gjort det tydeligt, at uheldstal som hovedregel følger den statistiske Poisson-fordelingen, opkaldt efter den franske matematiker Siméon Denis Poisson. At denne fordeling har praktisk anvendelse i uheldsanalysen, blev åbenbart med den polsk-russiske matematiker Ladislaus Bortkiewicz' bog af 1898, 'De små tals lov', hvor han fastslog, at antal soldater i den preussiske hær som i et givent år bliver dræbt af hestespark, netop følger Poisson-fordelingen.

Denne sandsynlighedsfordeling har den enestående egenskab, at variansen er lig forventningsværdien (middelværdien). Så snart vi har estimeret middelværdien, ved vi altså også, hvor meget variation vi skal forvente *omkring* denne værdi.

Den som hånd i handske specialtilpassede metode for uheldsanalyse er altså *Poisson-regressionsmodellen*. Efter at vi har specificeret vores uheldsfunktion, som et produkt af en række uafhængige faktorer, estimerer vores software nemt alle de koefficienter vi

er interesserede i, gennem såkaldt *sandsynlighedsmaksimering*, eller tilsvarende. Metoderne håndterer lige så let datamaterialer bestående af store, aggregerede tal, som datasæt hvor de fleste subjekter har nul uheld, nogle få har ét, og kun nogle yderst få har mere end ét uheld. Sådanne *disaggregerede* datasæt kan let løbe op i mange hundred tusind observationer – personer, husstande, køretøj, vejstrækninger eller vejkryds.

## Kend din observationsenhed

Det værste mareridt en forsker kan opleve, er måske det, at hun i slutfasen af sit projekt bliver i tvivl om, hvad der udgør hendes observationsenheder, eller fra hvilken population disse enheder er samlet. Fortolkningen af ethvert resultat kan i et sådant tilfælde være uigenkaldeligt kompromitteret.

Det er en fælde, som uheldsforskeren let kan falde i. Der er nemlig så mange muligheder at vælge blandt. Enheden i et uheldsmateriale kan for eksempel være:

- a. Et uheld
- b. Et personskadeuheld
- c. En tilskadekommen
- d. En person gennem en bestemt periode (år/måned/uge/dag/time)
- e. En person involveret i et uheld (tilskadekommet eller ej)
- f. En chauffør involveret i et (personskade)uheld
- g. Et køretøj
- h. En køretøjkilometer
- i. Et køretøj involveret i et (personskade)uheld
- j. En familie/husstand gennem en bestemt periode (år/måned/uge/dag/time)
- k. Et geografisk område gennem en bestemt periode (år/måned/uge/dag/time)
- l. En vejstrækning gennem en bestemt periode (år/måned/uge/dag/time)
- m. Et vejkryds gennem en bestemt periode (år/måned/uge/dag/time)

Enhederne af type a til j er *disaggregerede* – de udgøres af enkelthændelser eller enkeltsubjekter. Enhederne k til m er *aggregerede* – de består af optællinger eller gennemsnit indenfor en gruppe.

Et meget vigtigt point er, at nogle af disse enheder *forudsætter*, at et uheld har indtruffet. Det gælder enhederne a, b, c, e, f og i. Sådanne enheder kan bruges til at undersøge, hvorledes *skadegraden* afhænger af de respektive forklarende variabler.

Men i studier af *risiko* eller *uheldshyppighed* er de nær ved at være ubrugelige. I sådanne tilfælde er det nemlig lige så vigtigt at have information om alle de tilfælde, der *ikke* leder til uheld, som at vide noget om de, der gør. Det er jo ved at sammenligne de to slags tilfælde, at vi kan lære noget om, hvad der kendetegner det ene, men ikke det andet – med andre ord; hvad der forårsager uheld.

## Eksposering – vores vigtigste forklarende variabel

Med *eksposering* menes omfanget af den aktivitet, der genererer uheld. I trafiksikkerhedsanalysen bruger man ofte antal køretøjkilometer som mål på eksposeringen. *Risikoen* defineres som det forventede antal uheld (af en vis type) per enhed eksposering, f. eks. dødsulykker per køretøjkilometer.

Ingen statistisk uheldsmodel vil fungere godt uden, at vi har inkluderet et mål på eksponeringen. Det er vigtigt at måle denne så nøjagtigt som muligt.

Eksponeringen kan være flerdimensionel. Uheldene opstår som funktion af hvor lang distance som tilbagelægges af henholdsvis personbiler, busser, lastbiler, cyklister, motorcyklister og fodgængere.

## Kolinearitet – et indbildt problem

Mange forskere bilder sig ind, at man i en statistisk regressionsmodel for al del skal undgå, at variablerne er kolineære. Det er en gedigen vildfarelse.

Når man arbejder med ikke-eksperimentelle data, er (multi-)kolinearitet mere reglen end undtagelsen. I virkelighedens verden er alle variabler mere eller mindre korrelerede (kolineære). Det samme gælder også det lille udsnit af virkeligheden, som udgøres af vores datasæt.

Regressionsmodellens funktion er netop at efterligne denne virkelighed, og alligevel deducere partielle, parvise sammenhænger i en verden, der alt hænger sammen med alt. Modellen udskiller – destillerer, så at sige – virkningen af hver enkel variabel, på den hypotetiske betingelse, at de øvrige sidder i ro, til trods for at alle i virkeligheden har bevæget sig sammen.

Det har altså slet ingen mening at kræve, at variablerne ikke må være kolineære. Når det er sagt, skal det indrømmes, at når to eller flere variabler er stærkt korrelerede, er det svært at beregne effekten af hver enkelt. Vores estimater bliver let upræcise. Men problemet er ikke større, end hvad der fremgår af computerudskriften. Når to variabler er stærkt korrelerede, vil standardfejlen i hver koefficient blive høj, og p-værdien ligeså. Estimeringsprogrammet giver altså besked om, hvor dårlig præcisionen er.

## Vi skal nøjes med kun ét mål på størrelsen

Der er heldigvis nogle tricks, man kan anvende for at reducere kolineariteten i modellens variabler, og samtidig gøre fortolkningen nemmere.

Lad os antage, at vores datasæt består af individuelle køretøjer. Vi vil måske gerne beregne, hvorledes risikoen varierer med køretøjets vægt, længde, bredde, højde, motorkraft, acceleration, antal sæder, antal døre, etc. Disse variabler vil selvfølgelig være mere eller mindre kolineære.

Vores råd er at inkludere ét og kun ét mål på størrelse. Alle de øvrige variabler måles så i forhold til denne størrelse, eller i forhold til hinanden. I det ovenfor givne eksempel kan vi tænke os at specificere uheldshyppigheden som afhængig af;

- (a)  $m^2$  grundflade (længde gange bredde),
- (b) vægt (kg) per  $m^2$  grundflade,
- (c) motorkraft (kW) per kg vægt,
- (d) antal sæder per  $m^2$  og
- (e) antal døre per sæde.

Om alle variabler er målt på logaritmisk skala, har vi her lavet os en *multiplikativ dekomposition*, hvor den samlede effekt er brudt ned i fem bestanddele, og hvor de fem koefficienter har hver deres meningsfulde fortolkning. Koefficient (a) måler effekten af størrelse, når vægt, motorkraft, antal sæder og antal døre ændrer sig

proportionalt med bilens grundflade. Koefficient (b) måler effekten af, at bilen bliver én procent tungere, uden at blive større. Koefficient (c) måler effekten af, at samme bil får en stærkere motor – og dermed hurtigere acceleration. Koefficient (d) måler, hvorvidt tosædede biler har lavere uheldshyppighed end lige så store og lige så kraftige familievogne, mens koefficient (e) angiver, om risikoen øger eller synker når, groft regnet, bagsædepassagerene får deres egne døre.

Også summene og differencerne mellem koefficienterne har i vores eksempel interessante fortolkninger. Koefficient (a) minus (d) måler virkningen af én procent større grundflade, vægt og motorydelse, mens antal sæder og døre ikke ændres. Koefficient (b) plus (c) angiver effekten af, at vægten går op med én procent og motorydelsen med to.

Bemærk også, at i den model vi her har sat op, er kolineariteten stærkt reduceret. Om det er høj korrelation mellem grundflade og vægt, vil den være ringe mellem grundflade og vægt *per m<sup>2</sup>*. Vores multiplikative dekomposition gør variablerne tilnærmet ortogonale, det vil sige ukorrelerede med hinanden.

## Tegn stidiagram for at undgå endogenitet

Man siger gerne, at hver af koefficienterne i en regressionsmodel måler den isolerede effekt af den tilhørende variabel, *ceteris paribus*, det vil sige 'alt andet lige'. Det er dog ikke helt sandt, at *alt andet* skal være 'lige': Faktisk er forudsætningen kun, at de øvrige forklarende variabler *som vi har taget med i regressionen*, ikke ændrer sig.

Det indebærer, at man i regressionsmodellen ikke skal medtage to uafhængige variable X og Z, der hænger sådan sammen, at X altid ændrer sig, når Z gør det. Vi siger da, at X er *endogen* i forhold til Z, og det har ringe mening at spørge, hvad effekten bliver af en ændring i X, for konstant Z, eller omvendt.

Om for eksempel X er fartgrænsen og Z er hastigheden, så giver det ikke mening at tage dem begge med i deres oprindelige form. Selve hensigten med fartgrænsen er jo at få hastigheden ned. Men her spørger vi, hvad er effekten af fartgrænsen, for givet hastighed!

En mulig løsning kunne være at inkludere Z/X i tillæg til X, som i en multiplikativ dekomposition. Z/X måler da den relative fartoverskridelsen.

Hvorledes ved man så, i et givet fald, om X er endogen i forhold til Z? Til dette findes der intet facit svar. Det er et spørgsmål om faglig intuition og om teoretisk og empirisk indsigt.

Ét kneb er alligevel i denne situation til stor hjælp: Tegn et stidiagram, med bokse og årsagspile, der du, så godt du formår, beskriver hvorledes du formoder, at de forskellige variabler afhænger af hinanden. Da ser du nemmere, hvilke variabler du skal have med, og hvilke du skal lade falde bort.

## Ødelæg ikke den afhængige variabel

Til forskel fra nær sagt al anden økonometrisk analyse, ved vi, i tilfældet med færdselsuheld, noget meget vigtigt om vores restled: De følger, i den perfekt specificerede model, Poisson-fordelingen.



Om vi imidlertid transformerer vores afhængige variable – uheldstallene, kan vi ikke længere vide, hvilken sandsynlighedsfordeling som gælder. Vi skal derfor bevare uheldstallene i deres oprindelige, absolutte form og bruge dem som afhængige variable sådan som de står.

Om vi i stedet for uheldstallet bruger en form for rate, for eksempel uheld per år eller per køretøjkilometer, ja så ved vi øjeblikkelig mindre om modellens fordelings-egenskaber. Om vi omdanner uheldstallene til logaritmisk skala, så kender vi hverken fordelingen eller dens varians. Faktisk er variansen i dette tilfælde uendelig.

## Beregn den maksimale tilpasning

Så længe vi ved, hvorledes restleddene ser ud i den perfekte model, ja så kan vi også regne ud, hvor god tilpasningen til data i bedste fald kan blive. Selv i den bedste og mest fuldstændige model vil der forblive en vis mængde uforklarlig variation. Ved at sammenligne vores uheldsmodel mod den maksimalt opnåelige tilpasning kan vi få at vide, hvor langt vi er fra at forklare alt, der kan forklares.

Det findes altså i uheldsanalysen en form for facit for den optimale tilpasningen, som man i øvrige økonometriske anvendelser kun kan drømme om.

## Tjek for spuriøse sammenhænge gennem delmængdetesten

Som i enhver anden økonometrisk model risikerer man at estimere såkaldte spuriøse (uægte) sammenhænge: Jo flere storke, des flere børn kommer til verden.

Igen har man, i uheldsanalysen, en speciel mulighed for kontrol. Man kan nemlig gøre en *delmængdetest*.

Vi har måske opdaget, at køretøj tilhørende husstande med mandlige bilførere i alderen 18-20 år har øget skadefrekvens. Vi mistænker at dette skyldes de unge mænds kørefærdighed, overmod og/eller holdning til risiko.

Om så var, skulle der være en endnu kraftigere sammenhæng mellem forekomsten af unge mænd i husstanden og den *delmængde* af uheldene som involverer – netop – mænd i alderen 18-20 år. Om vi bruger samme set af uafhængige variable til at forklare, ikke alle typer uheld, men kun denne bestemte delmængde, så skulle vores koefficient-estimat blive væsentlig større. I modsat fald skal vi konkludere at sammenhængen er – i det mindste delvis – spuriøs.

Som en dobbelttjek kan vi køre samme regressionsmodel på en komplementær delmængde: Der skal ikke være signifikant sammenhæng mellem andelen unge mandlige førere i husstanden og antallet uheld som *ikke* involverer unge mænd. Er der det, så er sammenhængen spuriøs, og vi skal tro, at der også i vores oprindelige, mere generelle model er nogle ugler i mosen.

## Fokuser på den systematiske variation

Der er en omfangsrig litteratur om, hvorledes man kan specificere de tilfældige restleddenes simultane sandsynlighedsfordeling i en ligning, der forklarer færdselsuheld. Men man skal slet ikke bruge tid på dette.

I uheldsanalysen ved vi allerede i udgangspunktet en hel del om restleddsfordelingen – mere end på nogen anden arena for statistisk analyse. Vi skal derfor bruge vores tid og intellekt, ikke på den uinteressante, tilfældige variation, men på at forstå den langt vigtigere, *systematiske* variation – den som vi specificerer, når vi vælger vores forklarende variable.

Det er ud fra denne systematiske variation vi i bedste fald kan trække slutninger om, hvad der forårsager uheld.

# 1 Introduction and overview

The use of econometric models to analyze non-experimental data has been common practice in economics for half a century. There are, however, several reasons why this method would be at least as well suited for accident analysis as it is for economics.

Road accidents occur as a result of a potentially very large number of (causal) factors exerting their influence at the same location and time. To separate out and estimate the partial influence of any one factor, multivariate statistical methods are obviously called for.

Accidents are unwanted events, frequently even very traumatizing ones. To a large extent, this fact serves to preclude the use of perfectly controlled experiments as a means of gaining insights into the causal relationships behind the accident generating process.

There is, however, an abundance of non-experimental data available, in the form of road accident statistics and data sets covering a large number of different geographic or socio-economic units. The inner logic of accident data is such as to allow for unusually fruitful and efficient statistical methods. These opportunities should be exploited. We explain how.

<b>In essence, our advice consists of the following:</b>		<b>Page</b>
2	Keep track of your unit of analysis .....	2
3	Exploit the information inherent in the inner logic of casualty counts.....	3
4	Use econometrically efficient methods .....	5
5	Make the analysis as general as possible.....	7
6	Measure exposure.....	8
7	Preserve high levels of measurement .....	9
8	Measure size only once. Make smart decompositions. ....	10
9	Start with simple models, then add detail and sophistication .....	12
10	Avoid endogenous right-hand side variables. Draw path diagram .....	13
11	Leave left-hand side intact. Do all transformations on right-hand side .....	15
12	Beware of accident underreporting.....	16
13	Don't worry about multicollinearity .....	17
14	Compute the maximal goodness-of-fit.....	18
15	Apply casualty subset tests.....	19
16	Concentrate on the systematic part of the variation .....	20

## 2 Keep track of your unit of analysis

Just about the worst nightmare that can occur to a researcher, is when she, as she is approaching the end of the project, discovers that the data set which she has been using, consists of differing, i. e. incommensurable, units of analysis. Or, even worse, that she doesn't really know what exactly defines her unit of analysis, nor from which population the sample has been drawn.

To fix ideas, let us list a few units of analysis possible.

- a. An accident
- b. An injury accident
- c. An accident victim
- d. A person (in a given year/month/day)
- e. A person involved in an (injury) accident
- f. A driver involved in an (injury) accident
- g. A vehicle
- h. A vehicle kilometre
- i. A vehicle involved in an (injury) accident
- j. A household/family (in a given year/month/day)
- k. A geographic area (in a given year/month/day)
- l. A road link (in a given year/month/day)
- m. An intersection (in a given year/month/day)

While units a through j are disaggregate, i. e. consist of single events or decision makers, units k, l and m are (meso)aggregate, i. e. they are typically characterized by *counts* or *averages* of single events, individuals, households, vehicles, vehicle kilometres, or similar.

At his point it is worth noting that some of these units already presuppose that an accident has happened. This is true of units a, b, c, e, f and i. Such units aren't very useful in the study of accident frequency, for the simple reason that they are subject, if not to self-selection (since accidents are not chosen<sup>1</sup>), to a rather similar lopsidedness, in that all those 'cases' where no accident occurs, fail to enter the data set. However, if we want to study the causation of accidents or risk, it is just as important to have information on the 'no accident' cases as on those cases which do lead to accidents. It is precisely by studying the differences between these two types of cases that we can hope to draw inferences about accident causation.

---

<sup>1</sup> If the event results from a deliberate action, most countries would classify it as suicide or crime, not as an accident.

### 3 Exploit the information inherent in the inner logic of casualty counts

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) along the road.

What are the implications? They 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. There must be something wrong with the mathematical structure of that model. More precisely, the fact that casualty counts are non-negative numbers suggests a log-linear rather than a linear model structure. The antilog (exponential function) of zero is 1, while the antilog of minus infinity is zero. Thus, by taking the antilog we convert a range between minus infinity and infinity into a range between zero and infinity. Just what we need.

‘Log-linear’ essentially means ‘multiplicative’. Taking the antilog converts addition to multiplication. *The risk function is a product*. Risk factors work multiplicatively, not additively.

In mathematical notation, the ‘canonical’ form of an accident equation would be

$$(1) \quad \omega_{tr} = e^{\sum_i \beta_i x_{tri}} = \prod_i e^{\beta_i x_{tri}}$$

or, equivalently,

$$(2) \quad \ln(\omega_{tr}) = \sum_i \beta_i x_{tri}$$

Here,  $\omega_{tr}$  could be interpreted as the *expected* number of accidents or victims for unit  $(t,r)$ , where, in general, the index  $t$  could stand for time, while  $r$  represents the cross-sectional dimension (individual, vehicle, road link, region, or similar). The  $x$ 's are independent variables and the  $\beta$ 's are parameters to be estimated.

Second, there is no such thing as 1.2 accidents. Or 3.14 victims. Or 345.73. Any accident *count* placed *between* the integers is logically impossible. Thus, the set of possible casualty counts is much, 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 (smaller<sup>2</sup>) casualty counts should preferably be analyzed by statistical methods explicitly developed for count data, i. e. for non-negative, integer-valued dependent variables.

---

<sup>2</sup> When working with large accident counts, the distinction between integer and real-valued numbers becomes less important.

But make no mistake. All this does not mean that the *expected value* ( $\omega_r$ ) is also an integer. The accident model makes use of integer valued *observations* in order to estimate expected values that could be *any positive real number* (confer Figure 1 in the next section).

An analogy is the logit probability model, whereby one typically estimates *probabilities* between 0 and 1, using *data* consisting of only binary observations (0 or 1).

## 4 Use econometrically efficient methods

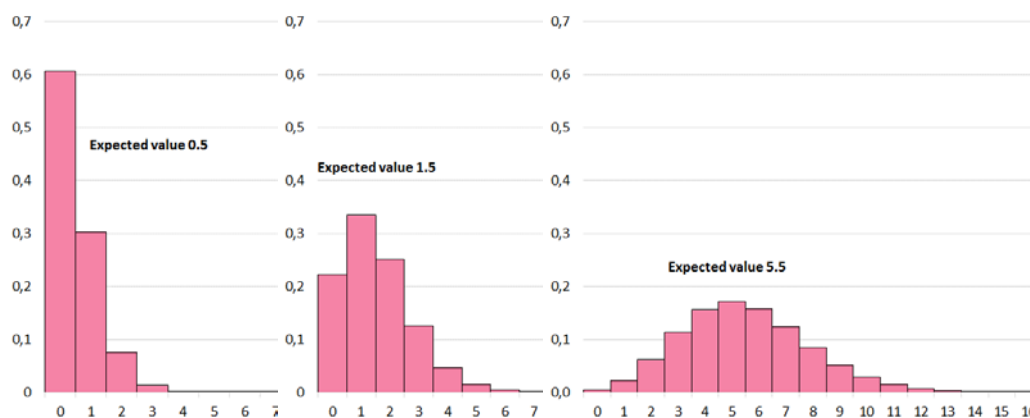
There is, luckily, an array of count data methods available. Even more luckily, it can be argued, empirically as well as theoretically, that these methods place the econometrician working with accident data in a uniquely privileged position. Unlike almost any other field of application, the econometrician working with accident counts will have excellent *a priori* information on the shape of his error distribution. This is, again, a mass of information which must be safeguarded and exploited for the purpose of a maximally efficient analysis.

The core model for count data analysis is the Poisson model<sup>3</sup>. According to this model, the probability that there will be exactly  $m$  events in unit  $(t,r)$  can be written

$$(3) \quad P[y_{tr} = m] = \frac{\omega_{tr}^m \cdot e^{-\omega_{tr}}}{m!},$$

where the random variable  $y_{tr}$  is the observed number of events, with mean  $\omega_{tr}$ .

In Figure 1 we show three different Poisson distributions, with expected values 0.5, 1.5 and 5.5, respectively.



**Figure 1. Histograms of the Poisson distributions with means 0.5, 1.5 and 5.5.**

The distribution is more skewed (asymmetric) the lower is the mean. In the left-most case, where the mean is one half, the most frequent outcome is 0 (zero).

<sup>3</sup> Developed by the French mathematician Siméon Denis Poisson (1781-1840), the Poisson model was long regarded as a mostly theoretical exercise in mathematical probability. It was only with the discovery made by the Russian-Polish researcher Ladislaus Bortkiewicz, a few generations later, that its enormous empirical applicability came to light. Bortkiewicz found that the number of soldiers in a Prussian cavalry corps that are killed by horse kick during a given year followed the Poisson distribution almost exactly. He published his results in his 1898 book 'The law of small numbers' – a provocative title since almost all statistical analysis up until then had been based on Bernoulli's celebrated law of *large* numbers and on the large sample theory emanating from the equally prestigious *central limit theorem* of Gauss/Laplace.

As the mean increases, the Poisson converges fast to the normal (Gaussian) distribution. One notes that, already when the mean is 5.5, the histogram is almost perfectly bell-shaped.

The Poisson distribution has the remarkable property that

$$\omega_{tr} = E[y_{tr}] = \text{var}[y_{tr}].$$

The variance equals the mean. Hence, once we have estimated the expected value, we also know what to expect in terms of variation *around* that mean. We can compute the minimum expectable random noise or, by subtracting it from the total variation, the *maximally explicable variation* (see section 14 below).

Strictly speaking, this is true only if the expected value is known. But even the simplest method of estimation will usually provide a sufficiently accurate (set of) estimate(s) for all practical purposes.

Various software packages are available that would estimate the parameters of equation (1) by means of maximum likelihood methods or similar. These methods are efficient<sup>4</sup>, in small as well as in large samples, and allow us to test a wide range of simple and/or composite hypotheses on the model parameters, using all the available information. The likelihood ratio test is particularly useful.

It gets even better....

---

<sup>4</sup> I. e., their precision cannot be improved.



## 5 Make the analysis as general as possible

The Poisson model, as derived by Poisson (1837, 1841), results from an apparently restrictive set of assumptions concerning the process leading up to an accident (Appendix A). In essence, it is assumed that accidents result from a random process governed by a constant *intensity*, and that all events are probabilistically independent, i. e. the probability of another event does not depend on previous occurrences. History does not matter.

Fortunately, this apparently restrictive model can be generalized in a number of useful directions:

- a. The intensity may vary continuously over time.
- b. The sum of independent Poisson variates is itself Poisson distributed, meaning that we may aggregate arbitrarily small time intervals (milliseconds) into one larger time unit (hour, day, month, year) without violating the Poisson assumption.
- c. The log of the Poisson parameter can be specified as a linear regression (equation 2), so as to vary between units.
- d. ‘Linear’ means ‘linear-in-parameters’. There is nothing preventing us from including non-linear *variables*, in the form, e. g., of roots, logarithms, variable products (interaction terms), power functions, polynomials, or combinations thereof.
- e. Exposure<sup>5</sup> could be one of the independent variables – or several, if exposure is multidimensional. If the exposure parameter is set to one, we are essentially estimating a risk function. A more general model is estimated if the exposure parameter is allowed to vary. Risk can be computed as the expected number of casualties divided by exposure.
- f. To account for heterogeneity or lack of independence, the Poisson parameter may itself be specified as random, leading to a *compound Poisson* distribution. If we imagine that  $\omega_r$  is drawn from a gamma distribution, the resulting compound distribution is the *negative binomial* (also referred to as generalized Poisson – see Appendix B).

Taken together, this means that we are perfectly justified in assuming that accident counts follow the (generalized) Poisson distribution, even when the counts are too aggregate to assume a constant accident intensity.

Moreover, the negative binomial generalization allows us to treat events that are probably not independent, like accident *victims*. Here, we must expect *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.

---

<sup>5</sup> See Section 6.

## 6 Measure exposure

By *exposure*, we mean the amount of activities that expose certain subjects to risk. Exposure is likely to be the most important explanatory variable in any accident model. Without a measure of exposure, the fit will be bad, the noise component large, and the coefficient estimates of other explanatory variable rather imprecise or biased.

Exposure could be measured in various ways, depending on the unit of analysis. If the unit is an individual or a household, we ideally would want to include the number of kilometres travelled, preferably by mode. If the unit is a vehicle, we want to use vehicle kilometres. If the unit is a region or a road link in a certain period, again we would use vehicle kilometres or AADT (average annual daily traffic), the latter possibly multiplied by the length of the road link.

To the extent that the analyst can choose between different data sets, with varying units of observation, access to exposure measures becomes a crucial argument.

For instance, a data set of vehicles whose odometers have been recorded at regular intervals could prove superior to data sets consisting of individuals or households with unknown travel history. An analysis based on vehicle data could provide highly interesting information on the risk associated with different vehicle models, vehicle age, and/or model years.

Moreover, if the vehicle records can be linked to owner or household characteristics, this approach may even shed light on the risk associated with the households members' personal characteristics. One set of independent variables could be like this:

- i. log of increase in odometer reading since year
- ii. log of (vehicle age (months) + 1)
- iii. vehicle age
- iv. vehicle age squared
- v. log of vehicle weight
- vi. log of engine effect per ton of vehicle weight
- vii. household size (number of persons above 18)
- viii. driver's licenses per household member above 18
- ix. average driver's license seniority
- x. share of license holders below 25
- xi. share of which are below 20
- xii. number of children below 18
- xiii. owners' gender (dummy)
- xiv. gender of youngest license holder (dummy)
- xv. number of household members with a criminal record
- xvi. log of household income
- xvii. highest level of education in household
- xviii. residential degree of urbanization
- xix. regional dummies
- xx. type of insurance (collision coverage? Bonus/malus?)
- xxi. number of cars in the household
- xxii. any other variable of interest....

## 7 Preserve high levels of measurement

Variables may be measured at the nominal, ordinal, interval or ratio level. There is more information in a ratio level measurement than at the interval level, which in turn is better than the ordinal level, and so on. Dummies are nominal level variables. Do not throw away the information contained in a ratio or interval level measurement, by converting the variable into a set of dummies!

Take the example of age (of a person, vehicle, license, etc). Many researchers make the curious choice of entering dummies for various age groups rather age itself. True, in this way one does not have to assume a particular curvature for the age effect. But by entering, e. g., a 5<sup>th</sup> degree polynomial in age, the age profile can take almost any shape, while requiring no more degrees of freedom than six age groups, and preserving all the information contained in the ratio level measurement!

In the specification suggested above, age (items ii through iv) is entered as a 3<sup>rd</sup> degree polynomial, since  $\log(x)$  is essentially equivalent to an exponent of 0.<sup>6</sup> Even exponents like  $-1$  or  $1/2$  can be used, if deemed necessary in order to capture the true curvature of a relationship.

---

<sup>6</sup> The Box-Cox transformation (Box and Cox 1964) is defined by

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \quad (x > 0) \\ \ln(x) & \text{if } \lambda = 0. \end{cases}$$

The parameter  $\lambda$  is generally referred to as the *Box-Cox parameter*. Different values of this parameter correspond to different curvatures or functional forms for the transformation. For instance,  $\lambda = 1$  yields a linear relation,  $\lambda = 0.5$  a square root law,  $\lambda = 2$  a quadratic function, and  $\lambda = 3$  a cubic function, while  $\lambda = 0$  and  $\lambda = -1$  correspond to the logarithmic and reciprocal (hyperbolic) functional forms, respectively. A most remarkable property of the Box-Cox transformation is the fact that it is continuous and differentiable even at  $\lambda = 0$ . It is, however, undefined for negative  $x$ .

## 8 Measure size only once. Make smart decompositions.

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 experiments, but must rely on non-experimental data.

Yet, there are some tricks available to keep related groups of variables from killing each other, thereby enhancing the ease of interpretation.

Take the example of vehicle size. One might consider entering weight, length, height, engine effect, fuel consumption, 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. What is, e. g., the meaning of the length coefficient, given weight, height, engine effect, etc.? The effect of making a car skinnier?

The solution is this: Enter only one variable describing vehicle size, and measure all other variables in relation to this one variable. For instance, enter  $\log^7$  of weight ( $w$ ), log of engine power ( $p$ ) per ton, log of fuel consumption ( $f$ ) per horsepower, etc. In this way, all three variables are entered in the form of a multiplicative decomposition:

$$(4) \quad \ln(\omega) = \beta_w \ln(w) + \beta_p \ln(p/w) + \beta_f \ln(f/p) + \dots,$$

The weight variable will capture the effect of size per se, the horsepower variable captures the effect of putting a stronger engine *into the same vehicle*, while the fuel variable captures the effect of a less energy efficient, *but otherwise equivalent* engine.

The sums and differences between these coefficients will also have interesting interpretations. For instance, the effect of increased weight, for a vehicle with given engine power, is measured by  $\beta_w - \beta_p$ , and the effect of a one per cent weight increase combined with a two per cent power increase and a two per cent higher fuel consumption is given by  $\beta_w + \beta_p$ .

A most important multiplicative decomposition is this: accidents = risk X exposure.

Another one is this: fatalities = injuries X deaths per injury (mortality) = accidents X injuries per accident (morbidity) X mortality = exposure X risk X morbidity X mortality.

To be able to disentangle all of these effects, one needs data on accidents of all degrees of severity, including property damage only (PDO) accidents. This is so because, typically, the same accident countermeasures that reduce fatalities, also serve to diminish the frequency of injury accidents, shifting some of these into the PDO category. Increased seat belt use, for instance, while obviously reducing the number of fatalities, shifting some of these cases into injury accidents, also has the effect of

---

<sup>7</sup> Confer end of section 11 below.

reducing many injury accidents to material damage accidents. The latter effect is likely to be much larger than the former, as measured by the *absolute* number of cases, and it is impossible to tell *a priori* which effect will be larger in *relative* terms. To avoid misinterpretations of a change in injury accident frequency, data on PDO accidents would be essential.

## 9 Start with simple models, then add detail and sophistication

The simplest possible accident model is this:

$$(5) \quad \ln(\omega_{tr}) = \beta_0 + \ln(x_{tr1}),$$

or

$$(6) \quad \omega_{tr} = e^{\beta_0} x_{tr1},$$

where  $x_{tr1}$  is some measure of exposure. The expected number of accidents is just proportional to the exposure,  $e^{\beta_0}$  being the proportionality constant. Here, trivially, the risk is given by

$$(7) \quad \frac{\omega_{tr}}{x_{tr1}} = e^{\beta_0}.$$

Now, if we want to relax the assumption of strict proportionality, which, by (8), is tantamount to constant risk, we estimate

$$(8) \quad \ln(\omega_{tr}) = \beta_0 + \beta_1 \ln(x_{tr1}),$$

corresponding to

$$(9) \quad \omega_{tr} = e^{\beta_0} x_{tr1}^{\beta_1}.$$

Here,  $\beta_1$  is the *elasticity of accident frequency with respect to exposure*.

In this model, the risk is given by

$$(10) \quad \frac{\omega_{tr}}{x_{tr1}} = e^{\beta_0} x_{tr1}^{\beta_1-1}.$$

In most models, the accident frequency elasticity will come out somewhere between 0.5 and 1, implying a *risk elasticity* between  $-0.5$  and  $0$ :

$$(11) \quad -0.5 < \beta_1 - 1 < 0.$$

If the accident frequency is less than proportional to exposure, it means that risk is a decreasing function of exposure.

Having gotten this far, we are ready to start adding more variables of interest. But...

## 10 Avoid endogenous right-hand side variables. Draw path diagram

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

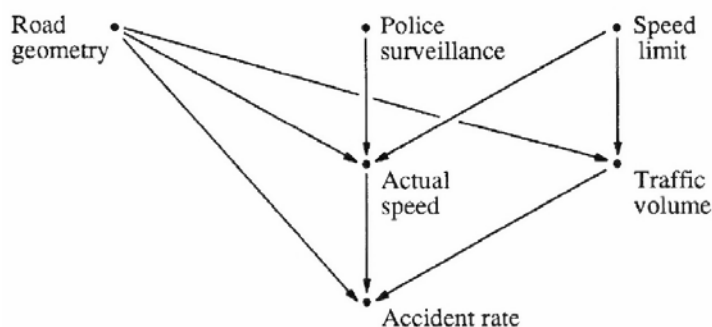
Take the example of education level (variable xvii listed in section 6 above). Obviously, if income (variable xvi) is also entered into the model, the education coefficient will measure the effect of higher education, given that it does not affect income! How interesting is that? A major motivation for, and effect of, higher education is to obtain better paid jobs. Here, income is clearly endogenous with respect to the education variable.

To avoid this problem, one would either have to drop one of the two variables, or define the income variable as 'income relative to mean income among persons with a similar (level of) education', much like the idea behind multiplicative decompositions.

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.

As a second and simpler example, assume that we want to estimate how lowering the speed limit on a given road link would affect the monthly injury accident frequency – the *accident rate*, for short. We imagine that the accident rate depends on the road geometry, the degree of police surveillance, the traffic flow, the average actual speed, and the speed limit. When we set out to draw a path diagram, we quickly discover that at least two of these variables – actual speed and traffic volume – are to some degree endogenous. It is reasonable to postulate that there will be arrows into these two 'independent' variables from the remaining three (Figure 2).

Obviously, it would make no sense to include actual speed in the model, as this would amount to asking: What are the effects of the speed limit, the police surveillance and the road geometry, given that the actual speed is kept constant?



**Figure 2. Partial causal path diagram for the effect of speed limits on the monthly accident frequency on a given road link. Source: Fridstrøm (1992).**

How to handle the traffic flow variable is, in this example, a bit more tricky. In a network where motorists have a route choice, the traffic flow will be positively related to the speed limits, since drivers tend to prefer faster roads to slower ones. This makes the traffic flow variable endogenous and speaks in favour of dropping it from our one-equation model.

But the traffic flow will depend on a host of other variables as well – economic, demographic and geographic, factors which one might want to control for. The speed limit may, in this context, be thought to have only a marginal effect. To take full account of these relationships, a richer and more complex, multi-equation econometric model may be called for.



## 11 Leave left-hand side intact. Do all transformations on right-hand side

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.

In short, transformed accident data present all kinds of challenges related to heteroskedasticity and estimation efficiency. These problems are avoided if we use the casualties themselves as dependent variable. True, raw accident data are heteroskedastic, too – but in a known way. The straightforward maximum likelihood method as applied to a (generalized) Poisson model implicitly accounts for heteroskedasticity in an optimal manner. We need not worry about it.

So our strong advice is this: Do not transform the left-hand side variable in an accident equation. It 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 *independent* variables. Then proceed to estimate by generalized Poisson maximum likelihood.

When you do specify the right-hand side of your equation, be aware that most software packages for maximum likelihood (generalized) Poisson estimation will implicitly assume that data are entered in a form compatible with

$$(2) \quad \ln(\omega_{tr}) = \sum_i \beta_i x_{tri}$$

i.e. the right-hand side is a *linear* combination. This means that if we want exposure (or any other independent variable) to be multiplicatively related to the casualty count, like in equations (7) or (10), we must measure the independent variable on a logarithmic scale. The implication of entering a variable  $x_{tri}$  measured on a linear scale is to assume that casualties are exponentially related to  $x_{tri}$ , i. e. proportional to  $e^{\beta_i x_{tri}}$ .

## 12 Beware of accident underreporting

Official accident statistics, generally based on police records, usually do *not* include property-damage-only (PDO) accidents, and even for injury accidents subject to mandatory police reporting, the coverage is notoriously incomplete. More seriously, underreporting varies systematically with a number of interesting factors, such as accident severity, travel mode, road type, age and gender (Høye et al. 2012: 14-17). Accidents involving bicyclists are, as a case in point, massively underreported.

Meticulous juxtaposition of police and medical records may allow for estimation of a reporting incidence function using the so-called capture-recapture method (Janstrup et al. 2013a, 2013b, 2014).

For reasons stated in section 11 above, it is, however, not advisable to ‘correct’ the accident counts prior to model estimation, even if we happen to have relevant and reliable estimates of reporting incidence at hand. A preferable solution would be to include these factors as explanatory variables in the equation.

Underreporting may represent a particularly difficult hurdle to accident *severity* analyses. This is so because severity countermeasures may be expected to affect, not only the number of fatal and serious injuries, but also the number of injury accidents altogether, shifting some of these into the PDO category, and hence beyond the scope of official accident statistics.

Increased seat belt use, for instance, while obviously reducing the number of fatalities, shifting some of these cases into the ‘serious injury’ or perhaps even into the ‘slight injury’ category, also has the effect of reducing many injury accidents to PDO accidents. The latter effect is likely to be much larger than the former, as measured by the *absolute* number of cases, and it is impossible to tell *a priori* which effect will be larger in *relative* terms.

To minimize these problems, it is strongly preferable to use data sources including accidents of all degrees of severity, down to the PDO level. While police and hospital records generally do not include such data, insurance records do. The analyst able to access individual insurance policy records, or a convenient aggregation thereof, will be in a privileged position, although – obviously – not even these data will be 100 per cent complete.

## 13 Don't worry about multicollinearity

Many practitioners, and even some renowned econometricians, believe that multicollinearity must or should be avoided. Don't listen to them.

As mentioned in section 8 above, the very idea of regression analysis is to separate out – distill, so to speak – the respective partial effects of a host of independent variables, which combine to produce an end result – the dependent variable – through a complex and non-transparent process. Non-experimental data are notoriously interrelated or at least correlated, i. e. collinear. In fact, collinearity is the very reason why we need multiple regression analysis to understand what is going on<sup>8</sup>. It makes no sense at all 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-tests, the p-values, and so on! The regression program will tell us all there is to say about this. The problem is only as big as your reported standard errors.

Many practitioners find it tempting, when faced with multicollinearity, to throw out some of the variables causing 'trouble'. Be aware that when you do that, every single parameter in the model acquires a new and different interpretation.

Ideally, the set of independent variables in the model should be determined by the *hypothetical experiments* that you imagine doing: by the variables whose effect you want to estimate, and by the variables you want to control for in these experiments (Haavelmo 1943, 1944; Pearl 2014). If some of these variables happen to be collinear, too bad: Then we cannot extract sufficient information without a more controlled, *real* experiment.

There is one case where multicollinearity cannot be ignored, namely when collinearity is 'perfect', meaning that the matrix of independent variables is (near-)singular and cannot be inverted. In this case, one of the independent variables can be expressed as an exact linear combination of the others and is, hence, redundant in the proper meaning of the word.

---

<sup>8</sup> Had all the independent variables been orthogonal to each other, i. e. with zero collinearity, we would be able to read off the effect of each variable by simple bivariate correlation or cross-tabulation.

## 14 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 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 average, 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. Any accident model exhibiting smaller residuals than this will be overfitted, i. e. beset by spurious correlation.

The suggestion is therefore, for any accident model, to compute the maximally obtainable fit ( $P^2$ , say) and relate the standard goodness-of-fit measure to this, i. e. to compute  $R^2/P^2$ . It will give the analyst a realistic picture of how much systematic variation is explained by his model. If  $R^2/P^2 = 1$ , there is no more variation to explain than what is already captured by the model.

This exercise is particularly useful when working with very small accident counts, as in a disaggregate model of individual persons of vehicles. In such models,  $R^2$  and  $P^2$  will be quite small numbers. But we can use  $P^2$  as a benchmark to judge just how badly (or well) the model fits.<sup>9</sup>

---

<sup>9</sup> See Appendix C or Fridstrøm et al. (1993, 1995).

## 15 Apply casualty subset tests

Casualty sets may be subdivided into subsets. For instance, the set of road traffic injury victims in Denmark in 2014 consists of

1. Males aged under 18
2. Males 18-25
3. Males above 25
4. Females aged under 18
5. Females 18-25
6. Females above 25

or of

1. Car drivers
2. Car passengers
3. Bus passengers
4. Truck drivers
5. Pedestrians
6. Bicyclist
7. Motorcyclists
8. Others

or of any cross-tabulation between these two.

In many cases, accident countermeasures work because they affect the behaviour of one particular road user group. Or some risk factor is relevant only for a particular group of people. To the extent, e. g., that vehicle age is important, it should affect car drivers and car passengers more than the average road user, while we do not expect pedestrian or bicyclist risk to be equally sensitive to vehicle technology.

It is possible to run accident regression models on just about any subset of casualties. This seems particularly useful if the data set consists of vehicles rather than persons or household.

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, run the exact same model on a smaller subset of casualties, such as 'males 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 model.

The general principles of casualty subset tests are described in Appendix D.

## **16 Concentrate on the systematic part of the variation**

Many researchers spend a lot of time inventing fanciful and sophisticated ways to specify the structure of the random disturbance term in the accident equation. There is a wide literature on this. Please 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 into account in the 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.

## Literature

- Ben-Akiva M and Lerman S R (1985): *Discrete choice analysis: theory and application to travel demand*. MIT Press, Cambridge, Mass.
- Bickel P J and Doksum K A (1977): *Mathematical statistics: basic ideas and selected topics*. San Francisco, Holden-Day.
- Bortkewitsch L von (1898): *Das Gesetz der kleinen Zahlen*. B G Teubner, Leipzig.
- Box G E P and Cox D R (1964): An analysis of transformations. *Journal of the Royal Statistical Society B* **26**: 211-243.
- Cameron A C and Trivedi P K (1998): *Regression analysis of count data*. Econometric Society Monographs no 30. Cambridge University Press, Cambridge.
- Eggenberger F and Pólya G (1923): Über die Statistik verketteter Vorgänge. *Zeitschrift für angewandte Mathematik und Mechanik* **1**: 279-289.
- Feller W (1943): On a general class of ‘contagious’ distributions. *Annals of mathematical Statistics* **14**: 389-400.
- Fridstrøm L (1992): Causality – is it all in your mind? Pp. 102-133 in Ljones O, Moen B and Østby L (eds.): *Mennesker og modeller – livsløp og kryssløp*. [Sosiale og økonomiske studier](#) **78**. Statistisk sentralbyrå, Oslo/Kongsvinger.
- Fridstrøm L (1999): *Econometric models of road use, accidents and road investment decisions. Volume II*. Report 457, Institute of Transport Economics, Oslo
- Fridstrøm L, Ifver J, Ingebrigtsen S, Kulmala R and Thomsen L K (1993): *Explaining the variation in road accident counts*. Report **Nord 1993:35**, Nordic Council of Ministers, Copenhagen/Oslo.
- Fridstrøm L, Ifver J, Ingebrigtsen S, Kulmala R and Thomsen L K (1995): Measuring the contribution of randomness, exposure, weather and daylight to the variation in road accident counts. *Accident Analysis & Prevention* **27**: 1-20.
- Gourieroux C, Monfort A and Trognon A (1984a): Pseudo maximum likelihood methods: theory. *Econometrica* **52**: 681-700.
- Gourieroux C, Monfort A and Trognon A (1984b): Pseudo maximum likelihood methods: application to Poisson models. *Econometrica* **52**: 701-720.
- Greenwood M and Yule G U (1920): An enquiry into the nature of frequency distributions to multiple happenings, with particular reference to the occurrence of multiple attacks of disease or repeated accidents. *Journal of the Royal Statistical Society A* **83**: 255-279.
- Haavelmo T (1943): The statistical implications of a system of simultaneous equations. *Econometrica* **11**: 1–12. Reprinted in Hendry and Morgan (1995: 477-490).
- Haavelmo T (1944): The probability approach in econometrics . Supplement to *Econometrica* **12**. Partially reprinted in Hendry and Morgan (1995: 440-453).

- Haight F A (1967): *Handbook of the Poisson distribution*. Wiley, New York.
- Hausman J, Hall B H and Griliches Z (1984): Econometric models for count data with an application to the patents-R&D relationship. *Econometrica* **52**(4): 909-938.
- Hendry D F and Morgan M S (eds) (1995): *The Foundations of Econometric Analysis*. Cambridge University Press, New York.
- Hoel P G, Port S C and Stone C J (1971): *Introduction to probability theory*. Houghton Mifflin, Boston.
- Høye A, Elvik R, Sørensen M W J and Vaa T (2012): *Trafikksikkerhetsboken*. Transportøkonomisk institutt, Oslo.
- Janstrup K, Kaplan S and Hels T (2013a): The under-reporting of traffic accidents: A logistic regression for calculating the probability of a traffic victim appearing in data set for police- or emergency room registered traffic accidents. Young Researchers Seminar (YRS), Lyon, France, 5-7 June 2013.
- Janstrup K H, Kaplan S and Hels T (2013b): Estimating the number of road accidents in Denmark: An application of the capture-recapture method. Strategic research in transport and infrastructure, Technical University of Denmark, 11-12 June 2013.
- Janstrup K H, Hels T, Kaplan S, Sommer H M (2014): Understanding traffic crash under-reporting: linking police and medical records to individual and crash characteristics. Transport Research Arena, 14-17 April 2014, Paris La Défense.
- Pearl J (2014): Trygve Haavelmo and the emergence of causal calculus. [\*Econometric Theory, Special Issue on Haavelmo Centennial\*](#).
- Peltzman S C (1975): The effects of automobile safety regulation. *Journal of Political Economy* **83**: 677-725.
- Poisson S D (1837): *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités*. Bachelier, Paris.
- Poisson S D (1841): *Lehrbuch der Wahrscheinlichkeitsrechnung und deren wichtigsten Anwendungen*. Meyer, Braunschweig.
- Ross S M (1970): *Applied probability models with optimization applications*. Holden-Day, San Francisco.
- Salmon W C (1984): *Scientific explanation and the causal structure of the world*. Princeton University Press, Princeton.



## Appendix A. The Poisson process<sup>10</sup>

At first, we shall need a few definitions.

A *stochastic process*  $\{Y(t), t \in T\}$  is a family of random variables. For each  $t$  contained in the index set  $T$ ,  $Y(t)$  is a random variable. The index  $t$  is often interpreted as time, in which case  $Y(t)$  represents the *state* of the process at time  $t$ . The set of possible values of  $Y(t)$  is called the *state space* of the process.

A continuous time stochastic process is said to have *independent increments* if, for all choices  $t_0 < t_1 < t_2 < \dots < t_n$ , the random variables

$$Y(t_1) - Y(t_0), Y(t_2) - Y(t_1), \dots, Y(t_n) - Y(t_{n-1})$$

are mutually independent. The process is said to have *stationary independent increments* if, for all  $t_1, t_2 \in T$  and  $s > 0$ , the variables  $Y(t_2 + s) - Y(t_1 + s)$  and  $Y(t_2) - Y(t_1)$  have the same distribution.

The *stochastic process*  $\{Y(t), t \geq 0\}$  is said to be a *counting process* if  $Y(t)$  represents the total number of events which have occurred up to time  $t$ .

A particularly important counting process is the *Poisson process*, defined by

$$(13.a) \quad Y(0) = 0,$$

$$(13.b) \quad \{Y(t), t \geq 0\} \text{ has stationary independent increments,}$$

$$(13.c) \quad P[Y(t) \geq 2] = o(t), \text{ and}$$

$$(13.d) \quad P[Y(t) = 1] = \lambda t + o(t),$$

where we have made use of the notation  $o(t)$  defined as follows: A function  $f$  is said to be  $o(t)$  if

$$(14) \quad \lim_{t \rightarrow 0} \frac{f(t)}{t} = 0.$$

Assumption (13.a) can be seen as an innocuous normalization rule. Assumptions (13.b-d) may, in plain language, be interpreted as follows:

- i. The time of recurrence of an event is unaffected by past occurrences.
- ii. The distribution of the number of events depends only on the length of the time for which we observe the process. For time intervals  $(s)$  of identical lengths, the event counts have identical distributions.

---

<sup>10</sup> This exposition is taken from Fridström (1999), which in turn relies on Ross (1970), Bickel and Doksum (1977), and Haight (1967).

- iii. The probability of exactly one event, divided by the length of the time period, tends towards a stable, positive parameter  $\lambda$ , which is called the *rate* or *intensity of the process*.
- iv. The chance of any occurrence in a given period goes to 0 as the period shrinks, and having only one occurrence becomes far more likely than multiple occurrences. For this reason, the Poisson process has been referred to as the *law of rare events*.

It can be shown (see, e. g., Ross 1970) that, for any process fulfilling these conditions, the number of events occurring during any interval of length  $t$  (say) has a *Poisson distribution*<sup>11</sup> with mean  $\lambda t$ . That is, for all  $s, t \geq 0$

$$(15) \quad P[Y(t+s) - Y(s) = m] = \frac{(\lambda t)^m \cdot e^{-\lambda t}}{m!}, \quad m = 0, 1, 2, \dots$$

It follows that

$$(16) \quad E[Y(t)] = \lambda t,$$

- i. e. the expected number of events is proportional to the length of the time period and to the rate of the process  $\lambda$ .

A bit simplified, one might say that, for any stationary counting process characterized by rare, mutually independent events, the number of events occurring during a time period of arbitrary length  $t$  follows the Poisson distribution with parameter  $\omega = \lambda t$ ,  $\lambda$  being the rate of the process.

This property is, of course, the reason why the process characterized by assumptions (13.a-d) is called a Poisson process.

Note, however, that the Poisson distribution is in no way part of these same assumptions. It is a remarkable, non-trivial mathematical fact that the Poisson distribution *follows* from these assumptions<sup>12</sup>.

A well-known example of a process fitting this description is the disintegration of *radioactive isotopes*. The atom decays by emitting neutrons at a given rate. The number of atoms disintegrating during a certain period is Poisson distributed.

It is impossible to tell when a specified atom will decay, but since all atoms are equal and the rate of decay is stable, we can predict with fairly large accuracy *how many* atoms will decay during a specified period. This is an example of what Salmon (1984) has referred to as an «irreducibly statistical law» – a causal law that includes an inevitable, objectively random component. No matter how much we learn about the radioactive substance, we would never be able to predict the behaviour of each elementary particle. Only their aggregate behaviour is knowable, and this only up to a certain (statistical) margin of error.

Another example of a process fitting the above description is – and this should come as no surprise – *accident counts*.

<sup>11</sup> Named after Poisson (1837, 1841).

<sup>12</sup> Alternatively, one might have taken (13.a-b) and (15) as the set of assumptions and derived (13.c-d) as implications. The latter relations are, in other words, both necessary and sufficient conditions for a Poisson process (Ross 1970: 13-14).

By striking analogy to the decaying radioactive isotope, accidents are also random and unpredictable at the micro level. Had the accident been anticipated, it would not have happened. Each single accident is, therefore, in a sense unpredictable by definition. Thus, even accident counts may seem to be governed by an «irreducibly statistical law», according to which single events occur at random intervals, but with an almost constant overall frequency in the long run. Although the single event is all but impossible to predict, the collection of such events behaves in a perfectly predictable way, amenable to description by means of precise mathematical-statistical relationships. There is reason to think that this principle applies to traffic accidents as it does to quantum physics, or to the (repeated) toss of a die.

Now, road users and road conditions are not, like the atoms of an isotope, all equal. At first sight, therefore, the stationarity part of condition (13.b) above may seem like a rather unrealistic assumption as applied to accidents, since it requires that the accident rate be constant over time. Even this condition is, however, for all practical purposes, an innocuous one. This is so on account of the convenient *invariance-under-summation property* of the Poisson distribution: any sum of independent Poisson variates is itself Poisson distributed, with parameter equal to the sum of the underlying, individual parameters (Hoel et al. 1971: 75-76). Thus all we need to assume is that, through some very short time interval (say, a minute, second, or fraction thereof), the accident rate can be considered constant, and that events occurring during disjoint time intervals are probabilistically independent. In such a case the number of events occurring during a given period  $t$  (week, month, or year) will, indeed, be Poisson distributed.

In fact, the conditions (13.a-d) may be generalized so as to describe the *non-homogeneous Poisson process*, in which the rate of the process may vary continuously over time, yet giving rise to Poisson distributed event counts. In the non-homogeneous Poisson process, the intensity is a function of time ( $t$ ), and the mean of the resulting Poisson variable is found by integration over the range of the intensity function:

$$(17) \quad E[Y(t)] = \int_0^t \lambda(s) ds .$$

The crucial condition left to be fulfilled, in order for the Poisson distribution to apply, is the independence part of criterion (13.b). Even this condition is, however, less restrictive than it may seem. It does *not* mean that accident counts should not be autocorrelated over time. If the underlying accident intensity  $\lambda(t)$  depends on *systematic* explanatory factors showing some degree of stability across consecutive time periods (a rather plausible assumption), changes in  $\lambda(t)$  will occur slowly and gradually, and this «inertia» will be reflected in the observed accident counts as well. Only the *random* part of the observed variation is, according to the Poisson process, uncorrelated across time.

The fact that an accident has just taken place does not increase the probability of another one occurring within the next few seconds, minutes, hours, or days. Nor does it reduce it. It may, however, well be that the systematic factors influencing  $\lambda(t)$  in period  $t_0 < t < t_1$ , take on similar values in the next period  $t_1 < t < t_2$ , thus increasing the accident probability in both periods. Such a phenomenon will manifest itself in the form of autocorrelated empirical accident counts. It does *not* contradict the assumption of probabilistically independent<sup>13</sup> accident counts or events<sup>14</sup>.

---

<sup>13</sup> We use the term *probabilistically* independent precisely to avoid confusion with respect to the two other meanings of the term «independent», that of *functional* independence (a uniformly zero partial derivative between two variables) and that of independent (exogenous) *variables* in a regression model.

<sup>14</sup> It might be argued that in certain cases, one cannot rule out the possibility that accident events may be probabilistically dependent. This occurs, e. g., (i) when the decision makers (the road users, the road authorities, the car manufacturers etc) learn from an accident and change their behaviour so as to avoid repetitions, or (ii) when an accident disrupts the traffic flow and thereby increases the risk of another one. In the statistical literature, this case is sometimes referred to as «*true contagion*». Unless, however, we are working with very disaggregate accident counts – pertaining to, say, individual drivers, vehicles, road links, or intersections – it is unlikely that such effects would represent more than an almost negligible deviation from the independence assumption. Moreover, to the extent that behaviour is changed in ways affecting risk, this would be reflected in the intensity of the Poisson process and – ideally – captured by the systematic factors included in the model.

## Appendix B. The generalized Poisson distribution

There are thus rather compelling arguments in favour of treating accident counts as a sample generated by the *Poisson* probability law, given by the formula

$$(3) \quad P[y_{tr} = m] = \frac{\omega_{tr}^m \cdot e^{-\omega_{tr}}}{m!}$$

where  $\omega_{tr}$  denotes the expected number of accidents during period  $t$  in area  $r$ , while  $y_{tr}$  is the corresponding, actual number of accidents.

In terms of analysis, the Poisson assumption has a number of useful and interesting implications (Fridstrøm et al. 1995). Most importantly, the variance of a Poisson variable equals its expected value, both being equal to the Poisson parameter –  $\omega_{tr}$ . Having estimated the expected value – relying, e. g., on a regression specification like (2) above – one also knows how much random variation is to be expected *around* that expected value.

Assume, for the sake of the argument, that we have somehow acquired complete and correct knowledge of all the factors  $x_{tri}$  causing systematic variation, and of all their coefficients  $\beta_i$ . In other words, the expected number of accidents – i. e., *all there is to know* about the accident generating process – *is known*. Could we then predict the accident number with certainty? The answer is no: there would still be an unavoidable amount of purely random variation left, the variance of which would be given – precisely – by  $\omega_{tr}$ . The residual variation should never be smaller than this, or else one must conclude that part of the purely random variation has been misinterpreted as systematic, and erroneously attributed to one or more causal factors<sup>15</sup>.

In practice one is seldom in the fortunate situation that all risk factors have been correctly identified and their coefficients most accurately estimated, so that the expected number of accidents is virtually known. A generalization of the Poisson probability model, and a sometimes more realistic regression model, is obtained when one assumes that the Poisson parameter  $\omega_{tr}$  is itself random, and drawn from

---

<sup>15</sup> American planners, politicians and scientists deliberately seek to avoid the term «accidents», replacing it by «crashes», on the grounds that the former tends to evoke the connotation of sheer randomness or bad luck, thereby neglecting the role of responsible, causal agents. In our view, however, the connotation of randomness is an entirely relevant one, as there is hardly, within the realm of social science, any phenomenon coming closer than road accidents to being truly (objectively) random in character. Moreover, randomness does not in any way contradict causation. As should be clear from the above discussion, random and systematic (causal) influences coexist. Being aware of the random component and of the need to separate it from the systematic part adds to our understanding, to our analytical opportunities, and hence ultimately to our knowledge on efficient accident countermeasures. We shall therefore continue to use the term «accidents», though definitely *without* implying that no one or nothing is to blame for them.

a *gamma* distribution with shape parameter  $\xi$  (say). In this case the observed number of accidents can be shown (Greenwood and Yule 1920, Eggenberger and Pólya 1923, Gourieroux et al. 1984 a, b) to follow a *negative binomial* distribution with expected value  $E[\omega_{tr}] = \varpi_{tr}$  (say) and variance

$$(18) \quad \sigma_{tr}^2 = \varpi_{tr} \cdot [1 + \theta \varpi_{tr}],$$

where  $\theta = 1/\xi$ .

In the negative binomial distribution, the variance thus generally exceeds the mean. In the special case  $\theta = 0$ , the gamma distribution is degenerate, and we are back to the simple Poisson distribution, in which the variance equals the mean. We shall refer to  $\theta$  as the «*overdispersion parameters*», and to models in which  $\theta > 0$  as «*overdispersed*». In such a model, the amount of unexplained variation is larger than the normal amount of random disturbance in a perfectly specified Poisson model, meaning, in fact, that not all the noise is purely random. The model does not explain all the systematic variation, but lumps part of it together with the random disturbance term.

The above line of arguments constitutes what could be termed the *epistemic* (subjective) reason for overdispersion. We recognize our lack of (complete) knowledge and specify the model accordingly, as when utility is treated as «*observationally random*», i. e. as random *as seen from the viewpoint of the analyst* (Ben-Akiva and Lerman 1985: 55-57).

More fundamentally, *ontic*<sup>16</sup> (objective) overdispersion may exist if the events are not probabilistically independent, such as accident *victims*, of which there may be several in a single accident. This fact tends to inflate the variance more than the expected value. In victim count models one should therefore never expect zero overdispersion.

This distinction between epistemic and ontic overdispersion is reflected in the two alternative derivations first offered for the negative binomial distribution. As noted by Feller (1943), quoted by Cameron and Trivedi (1998), these differed in a rather interesting way.

Greenwood and Yule (1920) based their derivation on an assumption of *unobserved population heterogeneity*, adjusting the statistical procedure so as to take explicit account of the analyst's less than perfect knowledge of the true expected values. This rationale is clearly epistemic: one does not question the underlying probability model, only our ability to learn about it.

Eggenberger and Pólya (1923), on the other hand, derived the very same distribution from an assumption of «*true contagion*», meaning that the occurrence of one event tends to increase the probability of another, as when counting disease cases during an epidemic. In this case, one relaxes the independence assumption of the underlying stochastic process, based on a belief that such independence is *inconsistent with reality*. This rationale is ontic in nature.

---

<sup>16</sup> Ontology is the theory of what really exists, i. e. of how the world really *is*, as opposed to what it looks like. Epistemology is the theory of knowledge, i. e. of how and whether we can *learn* or *know* about the real world. While ontic laws are, in a sense, true by definition, epistemic laws are just expressions of what we presently believe. The ontic law may exist even if its epistemic counterpart does not (the case of ignorance), or vice versa (the case of false theories).

As applied to accident victims, the «true contagion» assumption is obviously more realistic than the independence assumption. The fact that there is one victim increases the probability of another one.

When considering certain subsets of victims, however, deviations from the independence assumption may in some cases be so small as to be practically negligible. For instance, very few accidents involve more than one *pedestrian* or *bicyclist*. Hence, a good model for pedestrian and/or bicyclist injury victims should normally exhibit very little overdispersion. *Bus* or *car* accidents, in contrast, often involve more than one injury victim. Models explaining bus or car occupant injuries will therefore inevitably be overdispersed.

## Appendix C. Specialized goodness-of-fit measures for accident models

Fridstrøm et al. (1993, 1995) demonstrate how one can construct goodness-of-fit measures for accident models, which take account of the fact that casualty counts inevitably contain a certain amount of purely random, unexplainable variation.

Consider the well-known (squared) multiple correlation coefficient

$$(19) \quad R^2 = 1 - \frac{\sum_t \sum_r \hat{u}_{tr}^2}{\sum_t \sum_r (y_{tr} - \bar{y})^2} = \frac{\sum_t \sum_r (y_{tr} - \bar{y})^2 - \sum_t \sum_r \hat{u}_{tr}^2}{\sum_t \sum_r (y_{tr} - \bar{y})^2}$$

where  $\hat{u}_{tr}$  are the residuals and  $\bar{y}$  is the sample average of all casualty counts  $y_{tr}$ .

If  $y_{tr}$  is Poisson distributed with mean (and variance)  $\omega_{tr}$  (say), conditional on the independent variables, then the expected value of  $u_{tr}^2$  is equal to the variance of  $y_{tr}$ , which in turn equals  $\omega_{tr}$  (assuming no specification error). Thus, the total squared residual variation will have an expected value, correcting for the degrees of freedom, given by

$$E\left[\sum_t \sum_r \hat{u}_{tr}^2\right] = \frac{n-k}{n} \sum_t \sum_r \omega_{tr},$$

where  $n$  is the sample size and  $k$  is the number of estimated parameters.

A consistent (and usually very precise) estimate of  $\sum_t \sum_r \omega_{tr}$  is the sum of the fitted values  $\sum_t \sum_r \hat{y}_{tr}$ . This means that even in the perfectly specified accident model (in which all relevant variables have been included and all parameters have been estimated virtually without error), an *observable* upper bound on the coefficient of determination  $R^2$  is given by

$$(20) \quad P^2 = 1 - \frac{\frac{n-k}{n} \sum_t \sum_r \hat{y}_{tr}}{\sum_t \sum_r (y_{tr} - \bar{y})^2}.$$

Given this bound<sup>17</sup>, an intuitively appealing procedure would be to always judge the explanatory power of an accident model in relation to the maximally obtainable goodness-of-fit, by computing

---

<sup>17</sup> We refer to  $P^2$  as a «bound» not in the strict mathematical sense, but in the sense of an optimal (target) value – a prescriptive benchmark, so to speak. As noted below, an overfitted model would exhibit  $R^2 > P^2$ .



$$(21) \quad R_p^2 = \frac{R^2}{P^2} = \frac{\sum_t \sum_r (y_{tr} - \bar{y})^2 - \sum_t \sum_r \hat{u}_{tr}^2}{\sum_t \sum_r (y_{tr} - \bar{y})^2 - \frac{n-k}{n} \sum_t \sum_r \hat{y}_{tr}}.$$

Note that this measure differs from (19) only in that the normal amount of pure random variation has been subtracted from the total sample variation appearing in the denominator. To obtain a relevant measure of the model's power to explain systematic variation, we «purge» the overall sample variance of its inevitable random component.

One might therefore refer to  $R_p^2$  as the *coefficient of determination (R square) for systematic variation*. A model explaining virtually all systematic variation should have an  $R_p^2$  approaching one. In an overfitted model, we would have  $R_p^2 > 1$ .

## Appendix D. Casualty subset tests

Omitted variable bias is an important source of error in any econometric study. Whenever a regressor is correlated with the collection of explanatory variables *not* included in the model, the effect due to the excluded variables tends to be ascribed to the included one, inflating (or deflating) the coefficient of the latter. Any statistically significant effect found may thus, in principle, be due either (i) to a true causal relationship or (ii) to some kind of spurious correlation, or, indeed, to a combination of the two.

The number of factors influencing casualty counts is notoriously quite large. It is inconceivable that any econometric model would encompass all of them. Some factors are quite general, potentially influencing the frequency of (virtually) all types of accidents or victims, while other factors may be assumed to affect only certain subsets of casualties. To exploit our *a priori* knowledge of such relationships we introduce the following:

*Definition 1: Casualty subset tests.* Let  $A$ ,  $B$ ,  $C$  and  $D$  denote four sets of casualties (accidents or victims) such that

$$(22) \quad B \cap C = B \cap D = C \cap D = \emptyset \quad \text{and} \quad B \cup C \cup D = A,$$

i. e.  $B$ ,  $C$  and  $D$  are disjoint, exhaustive subsets of  $A$ , not all of them necessarily non-empty. Let

$$(23) \quad Y_{Ax} \equiv E(y_A | \mathbf{x}), \quad Y_{Bx} \equiv E(y_B | \mathbf{x}), \quad Y_{Cx} \equiv E(y_C | \mathbf{x}) \quad \text{and} \quad Y_{Dx} \equiv E(y_D | \mathbf{x})$$

denote the expected number of each type of casualties, conditional on a set of independent variables  $\mathbf{x} = [x_1 \ x_2 \ \dots ]'$ . Also, denote by

$$(24) \quad \varepsilon_{Ai} \equiv \frac{\partial Y_{Ax}}{\partial x_i} \frac{x_i}{Y_{Ax}}, \quad \varepsilon_{Bi} \equiv \frac{\partial Y_{Bx}}{\partial x_i} \frac{x_i}{Y_{Bx}}, \quad \varepsilon_{Ci} \equiv \frac{\partial Y_{Cx}}{\partial x_i} \frac{x_i}{Y_{Cx}} \quad \text{and} \quad \varepsilon_{Di} \equiv \frac{\partial Y_{Dx}}{\partial x_i} \frac{x_i}{Y_{Dx}}$$

the partial elasticities of  $Y_{Ax}$ ,  $Y_{Bx}$ ,  $Y_{Cx}$ , and  $Y_{Dx}$  with respect to some element  $x_i$  of  $\mathbf{x}$ . Note that, by definition,

$$(25) \quad \varepsilon_{Ai} = \varepsilon_{Bi} s_{Bx} + \varepsilon_{Ci} s_{Cx} + \varepsilon_{Di} s_{Dx},$$

where

$$(26) \quad s_{Bx} \equiv \frac{Y_{Bx}}{Y_{Ax}} \geq 0, \quad s_{Cx} \equiv \frac{Y_{Cx}}{Y_{Ax}} \geq 0 \quad \text{and} \quad s_{Dx} \equiv \frac{Y_{Dx}}{Y_{Ax}} \geq 0$$

denote the share of casualties belonging to subsets  $B$ ,  $C$ , and  $D$ , respectively.

Suppose that  $D = \emptyset$  and that we want to test a hypothesis of the form

$$(27) \quad H_1^+ : \varepsilon_{Bi} > \varepsilon_{Ai} > 0 = \varepsilon_{Ci}$$

or

$$(28) \quad H_1^- : \varepsilon_{Bi} < \varepsilon_{Ai} < 0 = \varepsilon_{Ci}$$

in other words that  $x_i$  has a larger positive (negative) effect on the number of casualties within subset  $B$ , a smaller positive (negative) effect on the total number of casualties (set  $A$ ), and a zero effect on casualties of type  $C$ .

Let  $\hat{\varepsilon}_{Ai}$ ,  $\hat{\varepsilon}_{Bi}$ ,  $\hat{\varepsilon}_{Ci}$ , and  $\hat{\varepsilon}_{Di}$  denote empirical sample estimates corresponding to the theoretical elasticities  $\varepsilon_{Ai}$ ,  $\varepsilon_{Bi}$ ,  $\varepsilon_{Ci}$ , and  $\varepsilon_{Di}$ , respectively.

*The hypothesis  $H_1^+$  (or  $H_1^-$ ) is said to **pass the affirmative casualty subset test as applied to  $B$  versus  $A$**  if and only if*

$$(29) \quad \hat{\varepsilon}_{Bi} > \hat{\varepsilon}_{Ai} > 0 \text{ (in case } H_1^+) \quad \text{or} \quad \hat{\varepsilon}_{Bi} < \hat{\varepsilon}_{Ai} < 0 \text{ (in case } H_1^-).$$

*It is said to **pass the complement casualty subset test as applied to  $B$  versus  $C$**  if and only if*

$$(30) \quad \hat{\varepsilon}_{Bi} > \hat{\varepsilon}_{Ci} \approx 0 \text{ (in case } H_1^+) \quad \text{or} \quad \hat{\varepsilon}_{Bi} < \hat{\varepsilon}_{Ci} \approx 0 \text{ (in case } H_1^-).$$

Alternatively, assume that  $C = \emptyset$  and consider the hypotheses

$$(31) \quad H_2^+ : \varepsilon_{Bi} > 0 > \varepsilon_{Di}$$

or

$$(32) \quad H_2^- : \varepsilon_{Bi} < 0 < \varepsilon_{Di}$$

*Hypothesis  $H_2^+$  (or  $H_2^-$ ) is said to **pass the converse (opposite) casualty subset test as applied to  $B$  versus  $D$**  if and only if*

$$(33) \quad \hat{\varepsilon}_{Bi} > 0 > \hat{\varepsilon}_{Di} \text{ (in case } H_2^+) \quad \text{or} \quad \hat{\varepsilon}_{Bi} < 0 < \hat{\varepsilon}_{Di} \text{ (in case } H_2^-).$$

The logic of these tests is illustrated by the following examples.

*Example 1:* Let  $A$  denote the set of *all road users* injured,  $B$  the set of *car occupants* injured,  $C$  the set of *non-occupants* injured.  $D$  is an empty subset. Also, let  $x_i$  denote the rate of seat belt *non-use*. Clearly, in this case one expects hypothesis  $H_1^+$  to hold. If the total number of road victims goes up as a result of reduced seat belt use (increased non-use), one should – *ceteris paribus* – be able to observe a stronger (relative) effect on car occupants ( $B$ ) than on road injuries in general ( $A$ ). This is the *affirmative* casualty subset test, confirming the impact of the safety measure by narrowing in on its target group.

One should, however, not see any effect of seat belt (non-)use on bicyclist and pedestrian injuries ( $C$ ) – unless, of course, car drivers adapt in the way maintained by Peltzman (1975), exposing non-occupants to higher risk. This is the *complement* casualty subset test, comparing the effect on the target group to the effect on its complement subset.

*Example 2:* Let  $A$  denote the set of *car occupants* injured,  $B$  the set of car occupants injured *while wearing a seat belt*, and  $D$  the set of car occupants injured while *not* wearing a seat belt.  $C$  is empty. As in the previous example, let  $x_i$  denote the rate of seat belt *non-use*. In this case one expects hypothesis  $H_2^-$  to hold: increased seat belt non-use should be positively related to the number of non-users injured, but negatively related to the number of seat belt users injured, simply because of the exposure effects. This is the *converse* (or *opposite*) casualty subset test, checking if the risk factor in question has the expected converse (opposite) effect on a suitably defined subset of the casualties. More seat belt use should – *ceteris paribus* – mean more seat belt users injured, even if the injury risk is much lower than in the non-user group.

At this stage the reader may want to ask what is the point of «testing» such entirely trivial relationships. It is this:

If our seat belt variable does not pass the complement casualty subset test as applied to car occupants versus non-occupants, but shows, e. g., a clearly significant, *positive* partial elasticity of *non-occupant* injuries with respect to seat belt non-use, there is reason to suspect omitted variable bias, probably inflating the effect of the seat belt variable *on its target group* (car occupant injuries) as well.

An even stronger indication of such bias is conveyed if our hypothesis fails to pass the converse casualty subset test as applied to seat belt users versus non-users.

One may note that our casualty subset tests are not set up as formal statistical significance tests. Only point estimates are compared, and pragmatic conclusions are drawn on the basis of their relative magnitudes. This is so because in most practical applications, one would not possess the relevant covariance estimates needed to perform, e. g., the asymptotic Wald test. Nor would comparable likelihood statistics be available, since casualty subset tests are generally based on separate, identical regressions explaining different dependent variables.

Only when a single elasticity is to be tested against a zero (or constant) alternative will we have enough information to perform a significance test.

In some cases, however, the zero alternative (in the complement casualty subset test) must be regarded as only approximate, such as when a risk or safety factor has a diluted effect even outside its main «target group». This will rarely apply to severity reducing (or increasing) factors, but quite frequently to accident reducing (or increasing) variables, since the latter will have spillover effects to other road user groups involved in bipartite or multipartite accidents. For instance, measures to reduce the accident risk of young drivers have a primary effect (if any) on this particular age group, but presumably also a diluted effect on the average risk experienced by other road users. In this case, therefore, one should not expect the effect observable within the complement subset to be exactly zero.

## **Institute of Transport Economics (TØI) Norwegian Centre for Transport Research**

Established in 1964, the Institute of Transport Economics is an interdisciplinary, applied research centre with approximately 70 professionals. Its mission is to develop and disseminate transportation knowledge that has scientific quality and practical application.

A private, non-profit foundation, TØI receives basic funding from the Research Council of Norway. However, the greater part of its revenue is generated through contract research. An important part of its activity is international research cooperation, mostly in the form of projects under the Framework Programmes of the European Commission.

TØI participates in the Oslo Centre for Interdisciplinary Environmental and Social Research (CIENS) located near the University of Oslo. See [www.ciens.no](http://www.ciens.no)

TØI covers all modes of transport and virtually all topics in transportation, including road safety, public transport, climate change and the environment, travel behaviour, tourism, land use and urban planning, decision-making processes, freight and travel demand, as well as general transport economics.

Claiming copyright to its products, TØI acts independently of its clients in matters of scientific approach, professional judgment and evaluation. TØI reports are generally downloadable for free at [www.toi.no](http://www.toi.no).

**Visiting and postal address:**  
Institute of Transport Economics  
Gaustadalléen 21  
NO-0349 Oslo

+ 47 22 57 38 00  
[toi@toi.no](mailto:toi@toi.no)  
[www.toi.no](http://www.toi.no)