Transportøkonomisk institutt
Stiftelsen Norsk senter for samferdselsforskning

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 variablerne, 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 estimered 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 samplet. 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.

Eksponering – vores vigtigste forklarende variabel

Med *eksponering* menes omfanget af den aktivitet, der genererer uheld. I trafiksikkerhedsanalysen bruger man ofte antal køretøjkilometer som mål på eksponeringen. *Risikoen* defineres som det forventede antal uheld (af en vis type) per enhed eksponering, 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 meningsfyldte 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^2. 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 variablen, *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 facitsvar. 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 variablen

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 variabler 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) sammenhænger: 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 variabler 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 del-mængde: 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 rest-leddenes 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 variabler.

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