Stockholms universitet

Att dö eller att inte dö? Analys av trafikolyckors dödsutfall i Sverige

Pavel Lukashin

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2017:22
Matematisk statistik
Juni 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2017:22**
http://www.math.su.se/matstat

Att dö eller att inte dö? Analys av trafikolyckors dödsutfall i Sverige

Pavel Lukashin*

Juni 2017

Sammanfattning

Den här uppsatsen har till syfte att förklara dödligt utfall av trafikolyckor. I denna studie används datamaterial för trafikolyckor i Sverige mellan åren 2005-2016. Genom logistisk regression har vi skattat sannolikheten att en trafikolycka leder till ett dödligt utfall givet olika förutsättningar som väg- och väderförhållandena, geografiska koordinater samt årstiden och inblandade typer av fordon. Då en del observationer saknar värden för vissa av de förklarande variabler används imputering för att estimera och ersätta dessa värden. För att testa hur väl denna modell kan prediktera utfallet av framtida trafikolyckor används metoden korsvalidering, där man anpassar modellen på en del av datan och predikterar på en annan.

Analysen har visat att dödligheten i trafikolyckorna beror på inblandningen av tunga fordon, fotgängare, cyklister och övrigt trafikelement (se Figur 8), men att det även finns geografisk och temporal inverkan. Av de väder- och vägförhållanden som fanns med i den givna datamängden var det sikten som influerade responsutfallet mest.

Som prediktionsmodell visade det sig att modellen var överanpassad på vår datamängd. För att hantera överanpassning hade man kunnat använda regularisering, det vill säga olika sätt att hantera skattningar av de parametrar som modellen använder.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: p.lukashin@gmail.com. Handledare: Tom Britton, Benjamin Allévius.

ABSTRACT

The purpose of this study is to explain the deadly outcome of traffic accidents. In this thesis we use data material for traffic accidents in Sweden during the years of 2005-2016. With the use of logistic regression we are going to estimate a model that explains the deadly outcome of traffic accidents with the help of a set of covariates. Since some observations contain unknown values we perform imputation to estimate and replace these observations. Furthermore to test how well our model can predict the outcome of traffic accidents we use cross-validation, where we estimate residuals of predictions by dividing data into separate blocks.

Our analysis have confirmed that the death rate in traffic accidents mostly depends on heavy vehicles, pedestrians, cyclists and vehicles of a special category (see Figure 8), but that there is also geographical and temporal influence present. Among the road and weather effects, vision indicators were the ones that affected the outcome most.

We have confirmed that the model was overfitted to our data and therefore could not predict the outcome of future accidents well. To control overfitting regularization could have been used, which is a method to change estimation of the parameters that the model uses.

Förord

Först och främst vill jag framföra ett stor tack till mina handlerade Tom Britton och Benjamin Allévius som har hjälpt mig från början till slut och agerat bollplank under hela arbetsgången. Jag vill även tacka mina studiekamrater Henri Jäderberg och Marika Lisinski som har bidragit till intressanta diskussioner kring olika frågeställningar. Detta examensarbete utgör 15 högskolepoäng vid Matematiska Institutionen vid Stockholms Universitet och resulterar i en kandidatexamen inom matematisk statistik.

Innehåll

1 Introduktion

Bilolyckor är ett fenomen vars början dateras redan till år 1869, då Mary Ward var den första som dog i en trafikolycka när hon blev överkörd av en ångdriven bil, trots att det knappt fanns några bilar på den tiden[19].

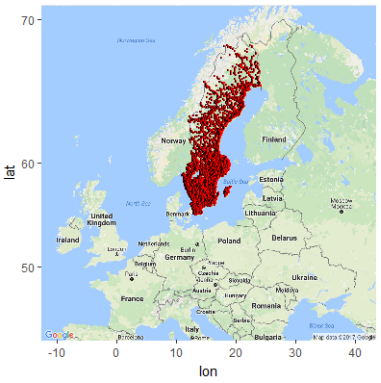
Sverige började föra statistik över vägtrafikolyckor år 1935, och med hjälp av den kunde man bygga sig en bättre uppfattning om antalet omkomna i olyckor. Antalet bilar ökade och det gjorde även antalet dödsfall, och man började inse att det finns ett starkt samband mellan hastighet, väglag, antalet trafikolyckor och dödlighet[4].

I dagens samhälle är vägarna mycket mer trafikerade och farligare. Bil- och järnvägar byggs på uppdrag av staten som får sin information av myndigheter som Transportstyrelsen. Eftersom verksamheten är skattefinansierad samt är en viktig del av samhällsutvecklingen, är det viktigt att kunna dra korrekta slutsatser från insamlade uppgifter om trafikolyckor och utveckla infrastrukturen så att den gynnar samhället mest[4].

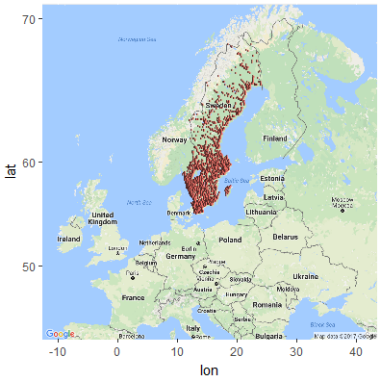
Med hjälp av statistisk modellering tar vi fram en logistisk regressionsmodell som förklarar vad är det som orsakar att en trafikolycka resulterar i ett dödligt utfall. Eftersom vi använder en logistisk modell används metoden för generaliserade linjära modeller för att skatta parametrarna. De variabler som saknas inom vissa observationer ersätts med imputering för att få en mer fullständig modell. Slutligen används metoden korsvalidering för att uppskatta modellens prediktionsförmåga, där vi skattar medelkvadratfelet. Arbetet genomförs med hjälp av programmet R(R Core Team, 2016) samt datamaterial från Transportstyrelsen om trafikolyckor daterat från år 2005-2016[5].

Genom tillgång till datamaterialet 'Trafikolyckor' som är en samling av 41 015 olyckor är målet att ta fram en modell som kan förklara om trafikolyckor slutade med dödsfall baserat på ett antal kovariater. De förklaringsvariabler som finns tillgängliga i vårt dataset är bland annat år, månad, geografiska koordinater, förhållanden på vägen samt information om olyckan såsom typen, platsen och de trafikelement som var inblandade. Totalt har vi tillgång till 22 olika förklaringsvariabler i vår datamängd. I modellen används både kategoriska och kontinuerliga förklaringsvariabler. Under 11 år har det skett 4 588 trafikolyckor som resulterade i dödsfall. Fördelning över samtliga olyckor,

samt dödliga olyckor representeras i Figur 1 och 2.



Figur 1:
Trafikolyckor i Sverige 2005-2016



Figur 2:
Dödliga trafikolyckor i Sverige
2005-2016

I detta arbete konstrueras en modell som förklarar dödsutfallet i trafikolyckor. Eftersom det finns ett stort antal av prediktorer där flera är kategoriska med flera nivåer, förenklas problemet först genom att kategorisera nivåer och variablerna till en mer kompakt format. Den slutliga modellen innehöll både prediktorer som fanns med i andra studier och några som var oväntade. Prediktionsanalysen visade på överanpassning och att regularisering krävs för att åtgärda detta[16].

I Avsnitt 2 tar vi upp den statistiska teorin som använts för modelleringen av trafikolyckorna. Avsnitt 3 handlar om hur modelleringsprocessen går till för att ta fram den förklarande modellen samt analys av prediktionsförmågan. Avsnitt 4 innehåller en sammanfattning av resulatet. Slutligen i Avsnitt 5 diskuteras resultatet, vilka slutsatser som kan dras och hur man eventuellt kan förbättra modellen. Avsnitt 6 innehåller bilagor med härledningar och tabeller med skattningar av olika modeller.

1.1 Data

I den här undersökningen har vi tillgång till ett dataset med över 40 000 observationer av trafikolyckor i Sverige under åren 2005-2016 som har erhållits från Transportstyrelsens databas. Datamaterialet består av förklaringsvariablerna

år, månad, län, kommun, gata, platstyp, olyckstyp, svårighetsgrad, väglag, väderlek, ljusförhållande, latitud, longtitud samt trafikelement 1-10. Vi inför förkortning på samtliga variabler som illustreras i Tabell 1.

Tabell 1: Förkortning för förklaringsvariabler.

Förklaringsvariabel	Förkortning
År	Å
Månad	M
Län	L
Kommun	K
Gata	G
Platstyp	P
Olyckstyp	O
Svårighetsgrad	S
Väglag	Väg
Väderlek	Väd
Ljusförhållande	Ljus
Latitud	Lat
Longitud	Lon
Trafikelement 1-10	T 1-10

Antalet observationer i förklaringsvariabeln trafikelement varierar för varje observation beroende på hur många trafikelement som var inblandade i en olycka. Svårighetsgrad är en binär variabel som kan anta värdena 'Dödligt' eller 'Svårt skadad'. Information om övriga skador finns inte med i det givna registret. I modellen kodar vi om 'Dödligt' till det numeriska värdet 1 och 'Svårt skadad' till värdet 0. Detta innebär att modellen skattar sannolikheten för att en olycka har ett dödlig utfall.

Innan modelleringen påbörjas måste datan filtreras. I vår datamängd ges uppgifter om var olyckan har skett med prediktorer som län, kommun, gata samt geografiska koordinater. Samtliga variabler har hög korrelation och vi bör därför välja endast en av dessa. Vi väljer att behålla longtitud och lattitud för vår modell för att på så sätt få kontinuerliga förklaringsvariabler snarare än kategoriska kovariater med ett stort antal nivåer i detta fall. Vidare ska vi reducera trafikelement 1-10 eftersom det finns över 34 olika typer av fordon, djur och övriga trafikelement som kan anta dessa värden (se Figur

7). För att besvara de hypoteser som är av intresse kommer vi att göra en kategorisering av dessa trafikelement ned till sex stycken kategorier. Dessa är: personbil, person, tungfordon, djur, motorcykel och övrigt med respektive beteckning i Tabell 2.

Tabell 2: Förkortning för trafikelement

Trafikelement	Förkortning
Personbil	Pbil
Person	Per
Tungfordon	Tung
Djur	D
Motorcykel	Moto
Övrigt	Ö

Liknande procedur genomförs med variablerna 'Platstyp' och 'Olyckstyp', där typen av plats och olycka generaliseras till vad som kan anses vara samma kategori. Genomförd kategorisering underlättar tolkning av den slutliga modellen och ger en klarare bild av de faktorer som har signifikant effekt på responsutfallet.

Eftersom variation av trafikolyckor kan även bero på tid på året används en periodisk komponent

$$a\cos(\tfrac{2\pi \mathrm{M}}{12}) + b\sin(\tfrac{2\pi \mathrm{M}}{12}),$$

där a och b skall skattas och M kan anta värden från 1 till 12[3].

Värden för latitud och longitud ges i referensramen SWEREF99 som är en svensk omvandling av det Europeiska tredimensionella systemet ETRS89 och motsvarar systemet WGS84, som är det system som används mest runt om i världen för navigering. SWEREF99 motsvarar WGS84 med några decimeters noggranhet och vi väljer att byta ut dessa mot varandra[7]. Vidare för att erhålla mer precision i modellen omvandlas koordinaterna i systemet till meter, där mitten av Sverige används som startpunkt. På det viset kan vi enklare förstå effekten av prediktorerna 'Lon' och 'Lat' på vår responsvariabeln.

2 Teori

Eftersom responsutfallet Svårighetsgrad i vårt dataset är kategorisk med två nivåer: 'Dödligt' och 'Svår skada', är det passande att använda sig av logistisk regression istället för linjär. För detta ändamål använder vi GLM, Generaliserade Linjära Modeller, som tillåter att responsvariabeln kommer från en exponentialfamilj[1].

2.1 Komponenter

En generaliserad linjär modell består av tre komponenter: en systematisk komponent, en slumpmässig komponent och en länkfunktion[1]. Nedan följer en kort beskrivning av samtliga komponenter.

2.1.1 Slumpmässig komponent

En slumpmässig komponent i den generaliserade linjära modellen består av en responsvariabel med oberoende observationer. Den naturliga exponentialfamiljen är en grupp av sannolikhetsfördelningar som är ett specialfall av en exponentialfamilj, där varje medlem innehåller en momentgenererande funktion. Det vill säga, om en täthetsfunktion kan skrivas på formen

$$f(y_i;\theta_i) = A(\theta_i)B(y_i)\exp[y_iQ(\theta_i)],\ i = 1, ..., N,$$

där $A(\theta)$, $B(y)$ och $Q(\theta)$ är funktioner som beror på respektive parameter och $i = 1, ..., N$ motsvarar rader i vår designmatris, så tillhör den exponentialfamiljen. I vårt fall har responsen Svårighetsgrad binomialfördelning med nivåer 'Dödligt' och 'Svår skada', där $Y_i = 1$ motsvarar dödliga trafikolyckor. $Y_i \sim$ Bernoulli$(\pi(\mathbf{x}_i))$ och $\pi(\mathbf{x}_i) = \theta_i = \mu_i$ är den naturliga parametern, det vill säga sannolikheten för en olycka att sluta i dödsfall med förklaringsvariablerna $\mathbf{x}_i$, där varje rad i $\mathbf{x}_i$ innehåller information från respektive olycka. Bernoullifördelningen är ett specialfall av binomialfördelningen, som i sin tur tillhör den naturliga exponentialfamiljen, vilket vi visar senare i avsnitt 2.2.1.

2.1.2 Systematisk komponent

Den systematiska komponenten specifierar kopplingen mellan förklarande variabler $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k)^T$ och slumpmässiga komponenter. Den kan skrivas som

$$\eta_i = \sum_j \beta_j x_{ij} = \mathbf{x}_i^T \beta$$

eller på matrisform $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, där $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^T$ och $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_N)$.

2.1.3 Länkfunktionen

Den tredje komponenten i en GLM modell är en länkfunktion, som kopplar samman den systematiska och slumpmässiga komponenten. Denna funktion används för att projicera väntevärdet för den slumpmässiga komponenten till den systematiska. Det vill säga om $\mu_i = E[Y_i]$ så kopplas μ_i till η_i som $\eta_i = g(\mu_i)$, där funktionen g är monoton och deriverbar. Så funktionen g länkar $E[Y_i]$ till förklaringsvariablerna genom

$$g(\mu_i) = \sum_j \beta_j x_{ij}.$$

Länkfunktionen som projicerar medelvärdet till den naturliga parametern η_i kallas för den kanoniska länkfunktionen, där $g(\mu_i) = Q(\theta_i)$ och $Q(\theta_i) = \sum_j \beta_j x_{ij}$.

2.2 Logistisk regression

Den kanoniska länken för den binära responsvariabeln är logit-länk. Det man egentligen modellerar är att sannolikheten för en observation är 1, $P(Y = 1) = \pi$. Den linjära komponenten av modellen innehåller designmatrisen samt en parametervektor som ska estimeras. Designmatrisen $\mathbf{X}$, som innehåller förklaringsvariablerna, består av N rader och K+1 kolumner, där K är antalet kovariater i modellen. Första kolumnen i designmatrisen består endast av 1:or, som motsvarar interceptet β_0. Parametervektorn $\boldsymbol{\beta}$ är en kolumnvektor med längden K+1. Logit-länken ser ut som följande

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{i=1}^{K} \beta_i x_i. \tag{1}$$

Om vi tar inversen av logfunktionen får vi att

$$\pi(x) = \frac{\exp(\beta_0 + \sum_i \beta_i x_i)}{1+\exp(\beta_0 + \sum_i \beta_i x_i)}$$

som är den logistiska regressionsmodellen[16].

2.2.1 Fördelningsantagande

När man jobbar med statistisk modellering finns ett antal antaganden som måste vara uppfyllda för att man ska kunna dra giltiga slutsatser om resultatet. Responsvariabeln som vi kommer att jobba med är $Y_i \sim \text{Bernoulli}(p_i)$, där p_i står för varje sannolikhet för dödligt utfall i vår modell. Binomialfördelningen är medlem av exponentialfamiljen och vi kan därför använda GLM för att ta fram modellen.

$$\begin{aligned} f(x; n, p) &= \binom{1}{x} p^x (1-p)^{1-x} && (2)\\ &= \exp\left(\log\left(\binom{1}{x}\right) p^x (1-p)^{1-x}\right) && (3)\\ &= \binom{1}{x} \exp\left(\log(p) + (1-x)\log(1-p)\right) && (4)\\ &= \binom{1}{x} \exp\left(x \log\left(\frac{p}{1-p}\right) + \log(1-p)\right) && (5) \end{aligned}$$

med den systematiska komponenten $\eta(p) = \log(\frac{p}{1-p})$ och $x \in \{0, 1\}$. Det betyder att länkfunktionen måste projicera från $(0, 1) \to (-\infty, \infty)$. Det lämpligaste valet är

$$g(\mu_i) = \text{logit}(\mu_i) = \log(\tfrac{\mu_i}{1-\mu_i}).$$

Ett annat villkor som krävs för att genomföra korrekt logistisk regression är att feltermer måste vara oberoende. Modellen bör också ha lite eller ingen

multikollineäritet, det vill säga att förklaringsvariablerna ska vara oberoende av varandra. Logistisk regression kräver även stora stickprov eftersom maximum likelihood-skattningar inte är lika kraftfulla som minsta kvadratmetoden. I vårt fall har vi tillgång till över 40 000 observationer och maximum likelihood-metoden kan betraktas som giltig[8].

2.2.2 Maximum likelihood-skattningar

När man använder GLM metoden så skattas parametrarna med maximum likelihood istället för minsta kvadrat skattningen som vid linjär regression, och därför förlitar sig metoden på stora stickprov. Anledningen till det är att minsta kvadratmetoden inte kan producera skattningar med minsta varians för parametrarna om feltermerna inte är normalfördelade. Likelihoodfunktionen indikerar hur troligt det är att ett observerat stickprov är en funktion av möjliga parametervärden. Maximering av likelihoodfunktionen ger de skattningar som är mest troliga att generera vår observerade data. Fördelarna med att använda sig av maximum likelihood-metoden är att vi får mer precisa skattningar.

Målet med logistisk regression är att estimera de $K + 1$ okända parametrarna i ekvation (1). I GLM används maximum likelihood-metoden som utgår från att hitta en uppsättning av parametervärden för vilka sannolikhet av observerad datamängden är störst. Maximum likelihood-skattningar fås från sannolikhetsfunktionen för responsvariabeln. Eftersom responsvariabeln i vår modell är binomialfördelad har den sannolikhetsfunktionen

$$f(y|\boldsymbol{\beta}) = \prod_{i=1}^{N} \tfrac{n_i!}{y_i!(n_i-y_i)!}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}.$$

Likelihoodfunktionen har samma form som sannolikhtesfunktionen förutom att parametrarna av funktionen är omvända, det vill säga likelihoodfunktionen uttrycker $\boldsymbol{\beta}$ i termer av kända värden för y på följande vis,

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i-y_i)!}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i} \tag{6}$$

och maximum likelihood-skattningarna är de värden för β som maximerar (6).

2.2.3 Oddskvot

Oddskvoten är ett sätt att mäta storleken av en effekt som beskriver styrkan av association eller oberoende mellan två binära förklaringsvariabler. Oddset definieras som $\Omega = \frac{\pi}{1-\pi}$, där π här är sannolikheten att dö i en trafikolycka. Oddskvoten ges av $\Theta = \frac{\Omega_1}{\Omega_2}$, där Ω_1 och Ω_2 är oddset mellan två händelser, till exempel om ett tungt fordon var inblandat i olyckan eller ej.

Θ kan anta tre olika värden som är intressanta för oss

$$\Theta = \begin{cases} < 1, & \text{associationen mellan två variabler är negativ} \\ 1, & \text{det finns ingen association mellan två variabler} \\ > 1, & \text{associationen mellan två variabler är positiv} \end{cases}$$

Från den slutliga modellen kan vi avläsa från skattningarna vilken effekt varje förklaringsvariabel har på responsvariabeln. Viktigt att notera är att för de faktorvariabler som har flera nivåer (se Figur 7), är det de nivåerna som följer med i modellen som är signifikanta gentemot basnivån. Dessa basnivåer väljer vi själva för respektive faktor[1].

2.2.4 Konfidensintervall

Det finns vissa fördelar och nackdelar med att använda sig av p-värdet för att avgöra om en variabel är signifikant för vår modell. Eftersom vi i denna studie genomför logistisk regression är det mer passande att använda sig av konfidensintervallet för oddskvoten. Om vi observerar att intervallet för oddskvoten för någon av våra förklaringsvariabler innehåller värdet 1, så är det möjligt att variabeln inte är signifikant för vår modell.

De maximum likelihood-skattningarna som vi får från regressionen är asymptotisk normalfördelade och väntevärdesriktiga. Wald-konfidensintervallet kan härldeas från Wald-statistikan som är asymptotiskt normalfördelad och definieras som

$$\sqrt{I(\hat{\theta}_{ML})}(\hat{\theta}_{ML} - \theta) \sim N(0, 1),$$

där funktionen $I(\theta)$ står för kovariansmatrisen, även kallad för Fisher informatonsmatris, för de olika parametrar θ som skattas och $\hat{\theta}_{ML}$ är maximum

likelihood skattningar. Vi kan då transformera de skattade $\hat{\theta}_{ML}$ med hjälp av en inverterbar funktion h, så att

$$\hat{\theta}_{ML} \sim N(0, Var(h(\hat{\theta}_{ML})),$$

där variansen fås från Fisher informationsmatrisen. Utifrån ovanstående information kan vi konstruera ett 95% Wald-konfidensintervall som följande

$$\hat{\theta}_{ML} \pm \lambda_{0.025} I(\hat{\theta}_{ML})^{\frac{-1}{2}},$$

där $\lambda_{0.025}$ är 0.025 kvantil för normalfördelningen. Härledning av Wald-konfidensintervallet finns i bilagan[12][9].

2.3 Modellval

För att välja den modell som förklarar datan bäst ska vi använda stegvis selektion, en metod som bygger på bakåt- och framåtselektion. Kortfattat fungerar algoritmen så att man utgår från en modell med endast intercept och stegvis lägger till variabler som genererar bästa anpassningen. Man fortsätter att addera variabler tills samtliga har använts och alla uppfyller gränsen för önskvärt p-värde. Givetvis kan någon av variablerna under proceduren ändra sin signifikans om en annan variabel läggs till och då elimineras den från den slutliga modellen[13].

2.3.1 Akaikes informationskriterium

En bra modell ska förklara mycket variation i data och det är inte alltid så enkelt att avgöra vilken modell som är bäst. Vissa modeller kan vara alldeles för komplexa att tolka om de erhåller för många variabler och interaktioner mellan termer. Andra kan vara alldeles för enkla och inte pålitliga. Ett mått för att kunna göra en avvägning mellan modellerna är Akaikes informationskriterium[1]. AIC är ett mått på hur väl modellen förklarar variationen i data korrigerad gentemot hur komplex modellen är och ges av

$$AIC = -2(\text{maximum likelihood - antal parametrar i modellen}).$$

När man jämför modellerna utgår man oftast från en mättad modell (modell med alla interaktioner och variabler) som jämförs med modeller med färre parametrar. Vi får dock ha i åtanke att det är en avvägningsfråga vilken

modell man ska välja och AIC är bara ett av kriterium som kan hjälpa oss på vägen. Det finns även andra faktorer som man bör väga in, till exempel vilka frågeställningar som ska besvaras av modellen.

2.4 Imputering

Ett av problemen i kvaliteten av datan är saknade värden. Att det finns flera observationen som innehåller okända värden kan bero på flera orsaker. Ett sätt att hantera det är imputering, som är benämningen på metoder som ersätter celler som saknas med rimliga värden. Fördelen med detta tillvägagångssätt är att behandligen av saknade värden är oberoende av skattningsmetoden som används. Vi kan alltså välja den metod som är mest lämplig beroende på situationen[2].

Imputeringsmetoden ersätter saknade värden med estimerade värden baserat på den tillgängliga informationen som finns i datamaterialet. Det finns flera alternativ att välja på: från en naiv metod som endast använder sig av medelvärdet till mer robusta metoder som utgår från relationer mellan observationer.

En annan viktig aspekt är hur imputering hanterar så kallade outliers, observationer som avviker markant från andra observationer, i kontinuerliga variabler. Modeller som tas fram med imputering kan påverkas starkt av avvikande observationer och imputerad data kan skilja sig markant från resten, och kan i sin tur bli outliers. Detta kan resultera i stor varians i de imputerade värdena.

För att hantera ovanstående problem använder vi k-nearest neighbour-metoden, som grundar sig på att låna information från närliggande observationer. kNN-metoden estimerar och ersätter observationer som saknas. Algoritmen som kNN-metoden använder antar att alla celler i datasetet motsvarar punkter i n-dimensionella rummet $\mathbb{R}^n$. De så kallade närmsta grannarna i fråga är definierade i termer av Euklidisk geometri, det vill säga avståndet mellan punkter i $\mathbb{R}^n$. Avståndet mellan två observationer är ett viktat medelvärde av bidrag från varje variabel, där vikten representerar hur viktig en variabel är. Vi definierar avståndet mellan observationer i och j som

$$d_{i,j} = \frac{\sum_{k=1}^{p} \omega_k \delta_{i,j,k}}{\sum_{k=1}^{p} \omega_k},$$

där ω_k är vikten och $\delta_{i,j,k}$ bidraget av variabel k. Bidraget δ för kontinuerliga variabler, lattitud och longtitud i vårt fall, ges av $\delta_{i,j,k} = \frac{|x_{i,k} - x_{j,k}|}{r_k}$, där $x_{i,k}$ är värdet av den k:te variabeln av den i:te observationen och $r_k = \max(x_{i,k}) - \min(x_{i,k})$ är räckvidden av den k:te variabeln. För variabler med nominal eller binär skala används distans 0 eller 1

$$\delta_{i,j,k} = \begin{cases} 0, & \text{om } x_{i,k} = x_{j,k} \\ 1, & \text{om } x_{i,k} \neq x_{j,k}. \end{cases}$$

Vi ser att alla $\delta_{i,j,k}$ ligger i intervallet $[0, 1]$, och som konsekvens följer då att alla $d_{i,j}$ mellan två observationer även ligger i detta intervall. Probabilistisk kan vi tolka ovanstående på följande vis: för givet datamängd och en okänd observation $x_{i,k}$ definieras en stokastisk variabeln Y med sannolikhetsfunktionen p, så att $Y \sim p$. Där för ett fix k, $p(y)$ definieras som andel av observationer $x_{i,k}$ i $N_k(x)$, så att $y_{i,k} = y$, där $N_k(x)$ är k närmaste grannar till $x_{i,k}$.

De främsta fördelarna med kNN-metoden är:

- Den kan prediktera både diskreta och kontinuerliga variabler.
- Algoritmen som används av kNN är robust för datamängden som innehåller mycket brus. Det medför stabilitet i algoritmen.
- Metoden är effektiv om datamängden är stor.

Den största nackdelen med kNN är att när den söker för de mest liknande förekomsterna av okända observationer, så används hela datamängden. Det kan vara väldigt krävande när man arbetar med stora datamängder. Denna begränsning kan vara mycket kritisk för KDD(Knowledge Discovery from Databases), eftersom algoritmen lär sig direkt av den givna datamängden. KDD är en process för att upptäcka användbart information från ett givet datamängd. KDD processen inkluderar förberedelse och selektion, datarensning m.m. Det medför även att algoritmen inte lär sig från 'träningsdatat' eftersom den använder hela datamängden. Ändring av antalet k-grannar kan även ge ett annat resultat vid körning av algoritmen.

2.5 Validering

Eftersom det är av intresse att undersöka hur väl den framtagna modellen kan prediktera utfallet av nya observationer är det passande att använda sig av validering. Metoden går ut på att man slumpmässigt delar upp givna observationer i två delar. En av dessa delar kommer att användas för att anpassa modellen och den andra för att validera. Den senare delen kallas även för hold-out set. Testfelet för modellen beräknas med hjälp av MSE(Mean Squared Error), där $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$ i de fall man arbetar med den kvantitativa responsvariabeln. Nedan följer algoritmen för valideringsprocedur

- Anpassa modellen på ett träningsset.
- Prediktera på ett valideringsset och beräkna felet som uppkommer i modellen.
- Justera modellen för bättre anpassning och upprepa de 2 föregående stegen.
- När felet är minsta möjliga, beräknar man felet på ett testset.

Felet på träningsdatan blir mindre så länge de predikterade värdena ligger nära de observerade och större om vissa predikterade och observerade observationer skiljer sig markant. Vi väljer modellen som har det lägsta testfelet.

Validering är en relativt enkel metod men för med sig två nackdelar:

- Uppskattningen av felfrekvenserna MSE kan variera mycket beroende på vilka observationer som hamnar i träningsmängden och vilka som ingår i valideringsuppsättningen.
- Valideringsmetoden använder sig endast av en delmängd av observationer för att anpassa modellen, nämligen de som ingår i träningsset. Eftersom statistiska metoder tenderar att prestera sämre när man tränar data med färre observationer kan det leda till att felet hos valideringssetet kan överestimera MSE för modellen som anpassas när man använder hela datamängden.

Genom att dela den givna datamängden i flera mindre dataset, där den ena delen kommer att användas för testning medan den andra används för träning, kan vi på så sätt skapa flera oberoende stickprov. Delning av da-

tamängden kommer att genomföras oberoende för att försäkra att alla dataset har samma fördelning. Denna metod ger oss en väntevärdesriktig skattning av testfelet men som kan ändå innehålla en del vit brus. Korsvalidering är en metod som löser detta problem. Vi tittar närmare på korsvalidering i nästa sektion[17].

2.5.1 k-faldig korsvalidering

Metoden involverar uppdelning av alla observationer i k-grupper av approximativt samma storlek. Vidare behandlas den första gruppen som valideringsset och resterande k-1 set anpassas. Man beräknar sedan MSE_1 för det dataset som användes för validering. Denna procedur repeteras k-gånger och varje gång använder man en annan grupp som valideringsset. Som resultat får man k stycken skattningar av MSE_1, MSE_2,..., MSE_k och k-faldig korsvalidering skattningar beräknas genom att ta medelvärdet av dessa värden $CV_k = \frac{1}{k}\sum_{i=1}^{k} MSE_i$.

När man genomför korsvalidering är målet att avgöra hur väl en statistisk inlärningsprocess kan förväntas att prestera på en oberoende datamängd.

Responsen i vår modell är kvalitativ och för att kvantifiera testfelet använder man antalet av felklassificerade observationer. Felet definieras då som $CV_k = \frac{1}{k}\sum_{i=1}^{k} Err_i$, där $Err_i = I(y_i \neq \hat{y}_i)$,. Vi kan betrakta det som felfrekvensen mellan de skattade och observerade värdena[17].

3 Modellering

I denna del av rapporten använder vi teorier och statistiska redskap från kapitel **2** för att välja modell som beskriver variationen av vår data bäst samt testar prediktionsförmågan. Avsnittet avslutas med en sammanfattning av de viktigaste slutsatserna.

3.1 Val och tolkning av modell

Modellen i vår logistiska regression väljs med hjälp av AIC-värdet, där vi har börjat med den mättade modellen som inkluderar alla förklaringsvariabler

och systematisk reducerat modellen med AIC-värdet som referens.

Efter bearbetning av datamängden börjar vi med att anpassa en modell inklusive alla parametrar. Notera att vi kommer att använda data som fortfarande innehåller okända värden. Modellen ur denna regression kommer att användas som utgångspunkt eftersom många variabler senare visade sig vara icke-signifikanta på 5 procent signifikansnivån (se Figur 9). Nästa steg är att genomföra imputering, som beskrevs i avsnitt 2.4, och ersätta de okända värdena med skattade samt anpassa en ny modell med samtliga förklaringsvariabler för att se om de imputerade värdena kan förbättra modellen. Modellerna visade sig ge liknande resultat och därför genomfördes bakåteliminering för att få bort de förklaringsvariabler som inte fångar upp variationen i datan. Den slutliga modellen kan skrivas som (se Figur 8)

$$\begin{aligned}\text{logit}(Y = 1|X) = {} & 35.56 - 0.018Tid - 0.059sin(2 * pi * M/12) \\ & - 0.065cos(2 * pi * M/12) - 0.383P : Korsning \\ & + 0.168P : Väg - 0.119O : Singelolycka \\ & + 0.068Ljus : Gryning/skymning + 0.254Ljus : Mörker \\ & - 0.002Lon + 0.0007Lat + 0.528Tung + 0.237Per + 0.544\ddot{O},\end{aligned}$$

där (:) indikerar nivån i respektive kategorisk variabel. Notera att nivån i variabeln Ljus:Gryning/skymning inte är signifikant. Problemet med att utesluta insignifikanta indikatorer är att vi kommer erhålla större p-värden för de andra nivåerna eftersom att vi flyttar på interceptet som motsvarar basnivån. Interceptet representerar medelvärdet av responsvariablen 'Svårighetsgrad:Dödligt' för nivå 'Dagsljus' i variabeln 'Ljus'. Eftersom vi endast utesluter nivå och inte fall för 'Gryning/Skymning' måste dessa fall ta vägen någon annanstans. Interceptet kommer att inkludera denna grupp för nivån 'Gryning/Skymning' och representera medelvärdet för responsen för nivå 'Dagsljus' och 'Gryninng/Skymning'. Att eliminera de osignifikanta nivåer leder till icke-väntevärdesriktiga skattningar och sämre p-värden, därför behåller vi modellen som den är.

För att kontrollera om det råder multikollineäritet bland förklaringsvariablerna kommer vi att använda VIF faktor, Variance Inflation Factor[13],

$$VIF_k = \frac{1}{1-R_k^2}.$$

VIF reflekterar alla andra faktorer som influerar osäkerhet i koefficienternas skattningar. Om vi observererar VIF > 10 så är kollineäriteten hög och vi bör överväga att utesluta vissa variabler. Det högsta värdet var ≈ 5 och antagandet om avsaknad av multikollineäritet är uppfyllt.

Eftersom antagandena för logistisk regression är uppfyllda kan vi dra giltiga slutsatser från modellen. Vi ska nu titta på skattningar i den slutliga modellen samt konfidensintervallet för samtliga variabler.

Vi tittar först på variabeln 'Tid' och de periodiska variablerna. Variabeln 'Tid' är definierad som Tid $=$ Å $+ \frac{\text{M}}{12}$. De periodiska komponenterna fås med hjälp av formeln $a\sin(x) + b\cos(x) = c\sin(x + \phi)$, där $c = \sqrt{a^2 + b^2}$ och $\phi = \arctan\left(\frac{b}{a}\right)$[3]. Variabeln Tid har negativ skattning och tyder på att sannolikheten att dö i en trafikolycka minskar med tiden. Utfallet stämmer överens med den historiska bakgrunden, då en av Trafikverkets främsta mål är 'Nollvisionen' - en vision om en vägtrafik där ingen människa ska dödas eller skadas allvarligt, som beslutades av riksdagen år 1997[18].

'P:Korsning' och 'P:Väg' har 'P:Annat' som referensgrupp vid regressionen. Den negativa skattningen som vi har fått för nivå 'Korsning' kan förklaras med att farten i korsningar är lägre än på raka sträckor när bilar kör ut från en trafikplats samt att det oftast finns trafikljus och skyltar i korsningar. Nivån i 'P:Väg' gav däremot en positiv skattning, det vill säga en ökad sannolikhet att dö i en trafikolycka. Mest troligt beror det på att det är högre fart på lands- och motorväg och risken att dö vid en sån kollision ökar.

'Ljus:Mörker', som har referensgrupp 'Ljus:Dagsljus', visar också på ökad sannolikhet att dö i en trafikolycka, det beror nog främst på att sikten är mycket sämre på natten och reaktionstiden minskar hos bilförare eftersom man inte ser eventuell fara i god tid.

Den geografiska inverkan, det vill säga effekten av Longitud och Latitud, riktning i öst-väst respektive norr-söder, var signifikant för responsutfallet. Vi konstaterar utifrån skattningar att effekten på responsen är relativt liten, men att det sker en trafikolycka med dödligt utfall söderut har högre sannolikhet än att det sker norrut. Däremot är sannolikheten för dödlighet i trafikolyckor i västra Sverige mindre än östra. Vägarna i södra Sverige är generellt sett hårdare trafikerade på grund av att dessa regioner har högre befolkningstäthet, men det behöver nödvändigtvis inte förklara varför dödligheten är högre, utan förklarar istället varför antalet olyckor är fler. Det kan dock

eventuellt finnas en korrelation mellan dessa. Resultatet från regressionen stärks om vi tittar på densiteten i Figur 1 och 2 på sidan 7.

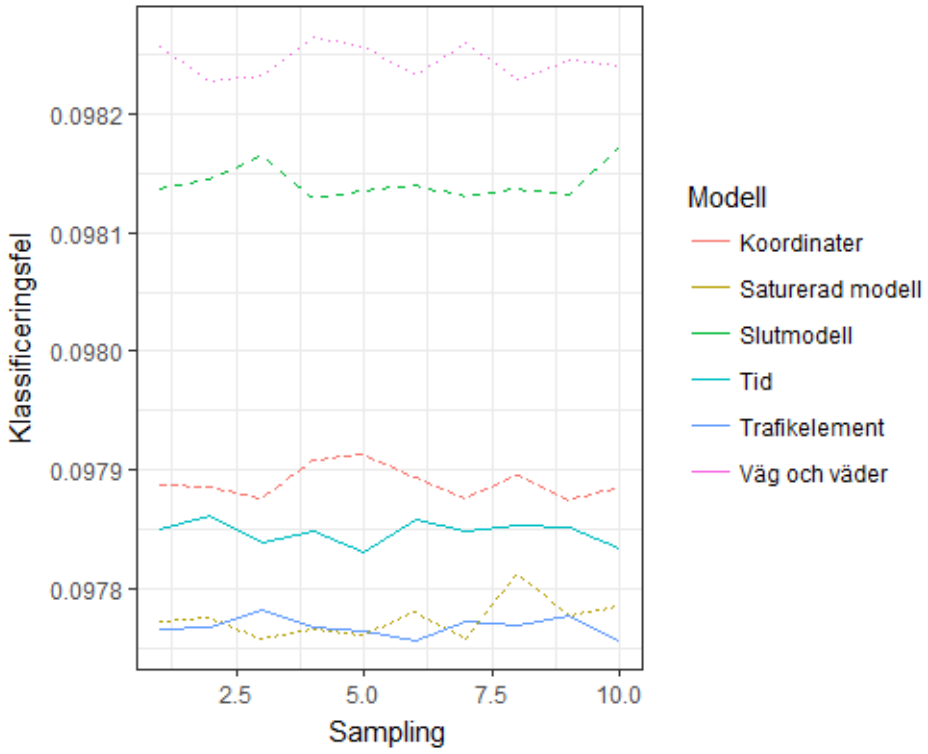
Slutligen tittar vi på de trafikelement som visade sig signifikanta för dödligt utfall. Dessa var Tungfordon, Person och Övrigt (se Figur 7 för kategorisering av respektive variabel). 'Tung' hade näst högsta skattningen av alla variabler i regressionen vilket tyder på att om det är en eller flera trafikelement från kategori Tungfordon som blir inblandade i en trafikolycka så ger det högre sannolikhet för dödligt utfall. Variabeln 'Ö' som inkluderar trafikelement så som snöskoter, släp och motorredskap, gav även den ökad chans att dö i en trafikolycka. Informationen kan dock vara missvisande eftersom antalet trafikolyckor med dessa trafikelement är betydligt mindre jämfört med 'Tung'. Skattningen för variabeln 'Per' indikerar på positiv dödligt utfall i en olycka, vilket tycks vara rimligt då även förklaringsvariabeln 'O:Singelolycka', det vill säga olycka med endast ett motorfordon inblandat, finns med i modellen.

Konfidensintervallet för samtliga variabler i vår modell illustreras i Figur 10. Vi konstaterar ur tabellen att samtliga variabler har en signifikant effekt på responsvariabeln, antingen negativ eller positiv, eftersom inget intervall inkluderar värdet $\Theta = 1$ utom 'Ljus:Gryning/skymning'. Då denna nivå inte var signifikant i regressionen behöver vi dock inte ta hänsyn till den. Förklaringsvariabeln 'P:Korsning' har oddskvoten på 0.681, vilket betyder att oddset att dö i en trafikolycka om den sker i en korsning minskar med 0.681, oberoende av koefficienterna av de andra förklaringsvariablerna i vår modell. Likaså för Tungfordon, som har oddskvoten på 1.7, ger nästan dubbelt så höga odds att dö om ett tungt fordon är inblandad jämfört med om det inte är det.

3.2 Prediktion

För att kontrollera modellens prediktionsförmåga använder vi k-faldig korsvalidering. Vi skattar CV värdet för sex olika modeller och jämför dessa med varandra.

K-faldig korsvaliderings metod var applicerad på våd data för att estimera klassifikationsfelet. Valideringsmetoden repeterades tio gånger där varje gång vi använder slumpmässigt uppdelning av observationerna i träningsset och valideringsset. Figur 3 illustrerar variationen hos felet som förekommer i



Figur 3: Plot över CV-felet för de 6 olika modeller

denna metod.

Bilden visar hur de olika modellerna presterar för prediktion av dödlighet på oberoende datamängd. De plottade linjerna i Figur 3 representerar felfrekvenser för sex olika modeller. Vi utgår från modellen med samtliga variabler som representeras av den bruna linjen. De andra linjerna symboliserar modellerna **utan** tid, dagliga väg- och väderförhållandena, geografiska koordinater och trafikelement från vänster till höger. Den sista gröna linjen är den slutliga modellen som vi har tagit fram i avsnitt 3.1 ovan. Trots att skillnaden bland modellerna är liten är det modellen med alla prediktorer som gav lägsta felfrekvensen. Att CV-felet för vår framtagna modell är högre beror mest troligt

på överanpassning, det vill säga modellen är kompatibel med vår datamängd men presterar sämre på oberoende data.

Det vi ser i Figur 3 är ett fall av ett så kallat *bias-variance tradeoff*, som är ett av de stora problemen inom inlärningsmetoder. Det man eftersträvar är att välja en modell som på ett bra sätt kan fånga upp variationen i träningsdatan och även kan generaliseras till oberoende datamängder. Oftast är det inte möjligt att åstadkomma båda samtidigt. Metoder som använder sig av hög varians kan producera bra resultat på träningsset men kan överanpassa på datamängder med mycket vitt brus. Däremot kan algoritmer med hög bias producera enklare modeller som kan underanpassa träningsdata och därmed misslyckas med att fånga viktiga regulariteter[17].

Modeller med låg bias är oftast de som är mer komplexa. Det leder till att dessa modeller kan representera träningssetet bättre. Nackdelen är dock att de även representerar en stor mängd av vitt brus från träningssetet och prediktionerna blir därmed sämre. I kontrast tenderar modeller med högre bias att vara relativt enkla men till följd producerar prediktioner med lägre varians när de tillämpas på andra datamängder.

Figuren ovan är ett exempel på ett sådant dilemma. Vår modell är den som innehåller flest prediktorer och därmed är mer komplex, vilket resulterar i överanpassningen. Vi kan dock konstatera att även om det föreligger variation i modellernas prestationsförmåga vad gäller prediktion, så är inte skillnaden mellan värdena markant.

4 Slutsatser

Vi ska nu sammanfatta de viktigaste resultaten vi har fått fram från sektion 3.1-3.2. Ett antal kovariater försvann under stegvis eliminering som man från början intuitvt trodde skulle vara signifikanta, såsom Väderlek och Väglag. Men eftersom modellen beskriver sannolikheten för att dö i trafikolyckor och inte sannolikheten för en trafikolycka att inträffa är det fullt rimligt att de ovannämnda variablerna uteslöts. Vid närmare granskning av dessa parametrar visade det sig att de var högt korrelerade, troligen eftersom väglaget beror mycket på vädret. Endast Väglag följde med i modellen, men den blev dock insignifikant på 5-procent signifikansnivån. Detta beror möjligtvis på

att vid sämre väglag minskas hastigheterna och med det även sannolikheten för att dö.

Skattningarna för respektive förklaringsvariabel visar på hur mycket inverkan de har på responsutfallet. Att den förklarande variabeln 'Tid' samt de periodiska variabler som vi lade till i modellen visade sig vara signifikanta tyder på att vår modell har lyckats fånga upp variationen i datamängden. Resultatet stärks ytterligare då antalet trafikolyckor har minskat markant under åren enligt en rapport från Trafikverket[11].

Med hjälp av skattningar från modellen samt av oss framtagna kartor ser vi att densiteten för olyckor i allmänhet samt de med dödligt utfall, är hög i södra Sverige. Troligtvis beror det på att det är fler som bor i dessa regioner men vi noterar även att de skattade värdena både för 'Lon' och 'Lat' var markant låga. Variabeln för Longitud pekar på minskad sannolikhet för dödligt utfall västerut. Det beror nog främst på att de sträckorna som körs och de mest trafikerade vägarna som till exempel E4:an och E18:n ligger i norr-syd riktning, därav den negativa skattningen för 'Lon'.

Enligt en rapport från STRADA har antalet dödade fotgängare och cyklister i trafikolyckor ökat. I vår modell kan man observera liknande effekt från skattningen av förklaringsvariabeln 'Per' som inkluderar både fotgängare och cyklister. Detta fenomen kan förklaras med att de satsningar inom infrastrukturen som görs av Trafikverket har varit inriktade mer för att öka säkerheten på motorväg och inte lika mycket för de områden där fotgängare och cyklister befinner sig. Det kan vara en av möjliga orsaker till att minskningen av trafikolyckorna har stagnerat under de senaste åren[14].

Ur modellen tolkas variabeln 'Tung' ha betydande effekt på dödligt utfall i en trafikolycka, vilket inte var så oväntat, då denna kategori inehåller de största och tyngsta fordonen. Att kollisioner med tung trafik har hög effekt på dödligt utfall beror inte endast på fordonens vikt men även att de har oerhört mycket längre bromssträcka, något som personbilsförare och fotgängare inte alltid tar hänsyn till. En lastbilsförare har även sämre sikt i döda vinkeln än personbilsförare som gör det svårt att märka cyklister och gångtrafikanter. Koefficientskattningen för variabeln 'Tung' är näst störst i vår modell. Vi konstaterar att 'Tung' har stor effekt på utfall av en trafikolycka och är därför mycket signifikant.

Något mer oväntat är att variabeln 'Ö' från kategori Övrigt följde med i den

slutliga regressionen. En möjlig tolkning kan vara att eftersom fordon från denna kategori har betydligt mindre storlek än en personbil så löper de även högre risk att få svårare skador vid en krock. Förklaringen kan även vara att man är mindre skyddad som förare när man manövrerar dessa fordon, därav ökning av dödligheten. Vi noterar dock att det finns väldigt få observationer med dessa trafikelement och vi kan därför inte förlita oss på detta resultat helt och hållet. För att dra mer konkreta slutsatser skulle vi behöva fler fall där just trafikelement från kategori 'Ö' var inblandade.

5 Diskussion

Resultatet hade definitivt varit annorlunda om vi skulle använda samma datamängd för att förklara och prediktera förekomsten av trafikolyckor. Variablerna 'Väg' och 'Väd' borde ha spelat signifikant roll för en trafikolycka att utfalla. Eftersom informationen som används i vår studie kommer från Transportstyrelsen som använder sig av en speciell mall för att registrera alla trafikolyckor som förekommer, är informationen bättre anpassad för att prediktera och förklara trafiolyckor i sig och inte dödligheten.

Förklaringsvariablerna för tid kunde ha haft ännu starkare effekt på responsutfallet om vi hade minskat perioden och på så sätt eventuellt fångat in mer fluktuationer i vår data. Att minska perioden ytterliggare kan dock leda till överanpassning av modellen som betyder att modellen blir för komplex för vår datamängd. Modellen skulle då endast beskriva variationen för vår specifika datamängd och fungera bristfälligt för ett annat stickprov, något som vi redan har observerat vid skattning av prediktionsförmågan i avsnitt 3.2

En av slutsatserna som vi kan dra efter genomfört arbetet är att för att kunna ge bättre svar på våra frågeställningar behövs mer data för varje olycka. I rapporter från olika instanser förekommer ofta informationen om kön samt ålder för de som har varit inblandade i olyckorna. På grund av PUG (personuppgiftslagen) får vi inte använda denna information i vår modell så vi kan endast spekulera om det hade hjälpt oss att generera en bättre modell.

I rapporten från STRADA från Värmland är det framförallt kvinnorna vars ålder överstiger 55 år som skadas i trafikolyckor med personbil. I det fallet hade tillgång till informationen om kön kunnat hjälpa vår modell att fånga

upp mer av variationen i datan. Även variabeln 'O:Singel olycka' hade kunnat ge oss högre skattningar, då majoriteten av fotgängare som omkommer i trafikolyckor har varit i kontakt med en personbil[14].

Andra exempel på information som oftast inte är tillgänglig när olyckan sker är hastighet. Det är en variabel som bör ha stor påverkan på dödlighet i trafikolyckor. Enligt NTF (Nationalförening för Trafiksäkerhetens Främjande) rapport om myter och sanningar om hastighet i trafiken bidrar ökning/sänkning av 5% av hastigheten till 25% ökning/minskning risk för dödsolyckor[6]. Ett sätt att inkludera denna effekt är att använda sig av hastighetsbegränsningar på vägen där olyckan har inträffat. På det viset får man ett approximativt värde på hastigheten när olyckan sker. Nackdelen med denna metod är dock att den riktiga hastigheten kan skilja sig markant och skattningar blir inte väntevärdesriktiga.

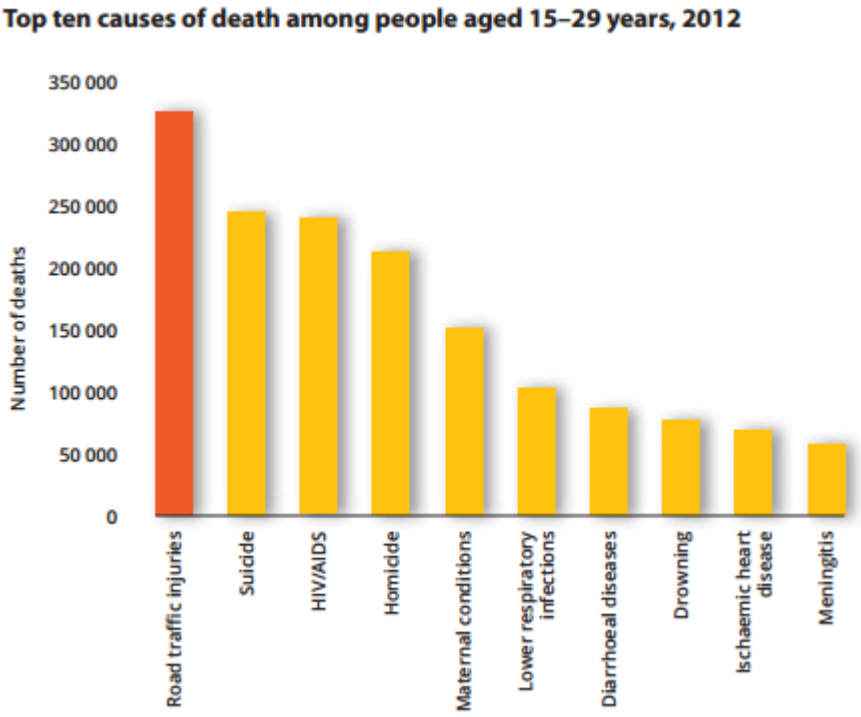
Världshälsoorganisationen WHO (World Health Orgasnization) presenterar en rapport där det framgår att trafikolyckor är det främsta dödsorsaken runt om i världen[10]. I Figur 4 från rapporten kan vi se topp 10 dödliga orsaker i världen i år 2012.

Under det senaste årtiondet har man kunnat observera en avtagande ökning vad gäller dödsfall som vi kan se i Figur 5 nedan. Mest troligt beror det på utvecklingen av infrastrukturen runtom i världen.

Trots dessa satsningar ser vi dock fortsatt ökade dödsfall i trafiken. För att bryta trenden krävs mer resurser och även inblandning av andra instanser som kan influera och påverka utvecklingen av säkrare vägar.

I Figur 6 kan vi se att frekvensena för dödsfall i trafikolyckor på global skala singifikant varierar mellan olika regioner. Likaså som i Sverige har det varit lite ändring på dödsfrekvenserna sen 2010[10]. Vi kan dock konstatera att Sverige som är med i stapeln 'European' ligger markant under det globala medelvärdet med dödlighet på 9.3 per 100 000 invånare gentemot 'World' på 17.4 per 100 000 invånare.

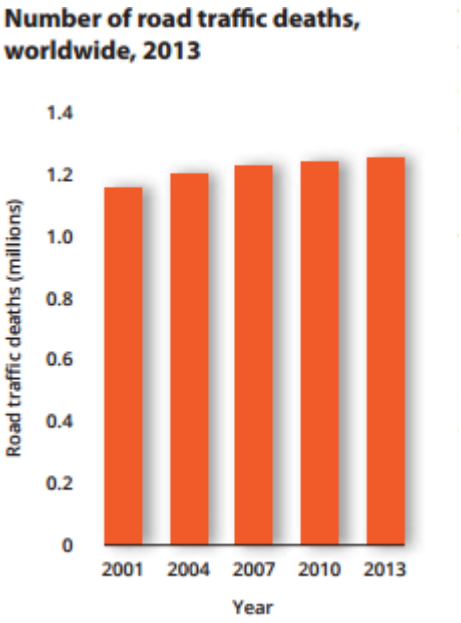
Värt att notera är att utav de 41015 trafikolyckor i vår dataset är det endast 4 588 som var dödliga. I framtida arbetet hade man eventuellt kunnat använda sig av olika instanser för att på så sätt få fler dödliga olyckor i datamängden. Man får ha i åtanke att eftersom infrastrukturen skiljer sig markant mellan länder kan resultatet bli felaktig vid tolkning och prediktion av dödliga utfall.



Figur 4: Topp 10 dödsorsaker år 2012[10]

Eftersom att vår modell visade sig vara sämre på att prediktera än modeller med färre förklaringsvariabler och högre AIC värde tyder det på att modellen mest troligt är överanpassad till specifika observationer som finns i stickprovet och kan inte generaliseras till andra datamängder.

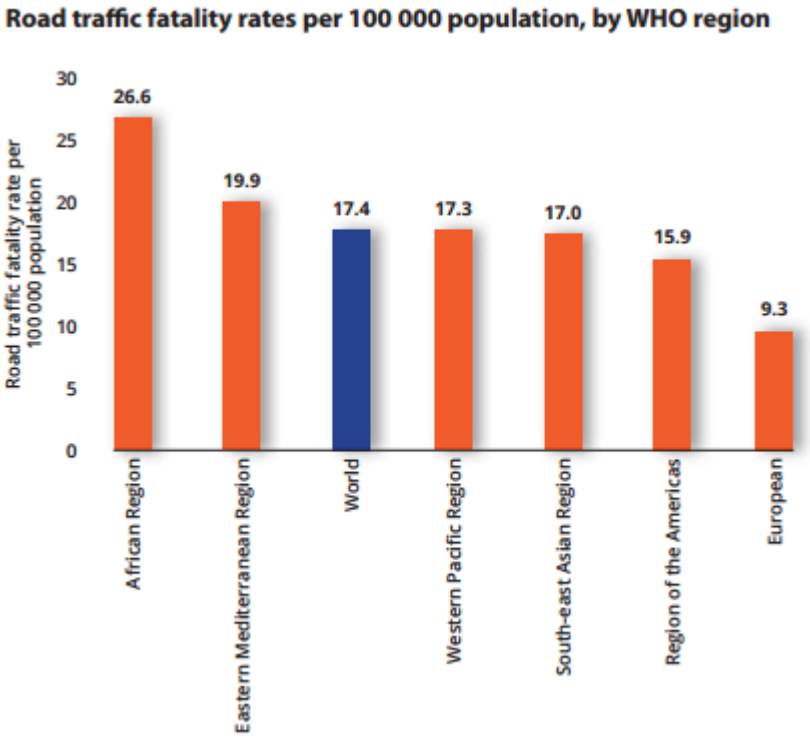
För att undvika överanpassning skulle vi behöva en större datamängd för att se till att samtliga stickprov vid korsvalidering är tillräckligt stora. En metod som skulle hjälpa oss för att minska överanpassning är regularisering. Regulariseringsalgoritmer bestraffar vissa egenskaper hos de skattade parametrarna och kan vara en metod att föredra när man har tillgång till en mindre mängd förklaringsvariabler som är signifikanta för modellen[17]. En kort beskrivning av Ridge regularisering ges i bilaga 6.2.



Figur 5: Utveckling av dödsfall i trafikolyckor 2001-2013[10]

En slutsats som man kan dra efter genomfört arbete är att det vore lämpligare att skapa en modell för prediktion och en annan modell som är mer förklarande. Den andra delen har vi lyckats att åstadkomma då modellen som vi har fått fram är lättolkad och man kan se en tydlig effekt av varje prediktor på responsvariabeln. För att förbättra prediktionsförmågan skulle vi behöva ta fram en modell som är mer komplex, då det vid prediktion inte spelar så stor roll om modellen är svårtolkad och på vilket sätt dödligheten beror på en variabel. Det viktiga är att den kan prediktera rätt.

Då uppsatsens syfte i första hand var att ta fram en förklarande modell och i andra hand titta på dess prediktionsförmåga kan vi konstatera att det är sällsynt att samma modell är lämpligast för båda ändamål. Därför valde vi den modell som bäst förklarar vår huvudsakliga syftet.



Figur 6: Frekvens av dödliga trafikolyckor per 100 000 invånare[10].

6 Bilagor

6.1 Härledning av Wald-konfidensintervall

Nedan ska vi härleda Wald-konfidensintervall för en binomialfördelad parameter. Konfidensintervallet för en parameter ges av

$$\hat{\theta}_{ML} \pm \lambda_{0.025} \sqrt{I(\hat{\theta}_{ML})^{-1}}$$

Eftersom kvantilen från normalfördelningen är känd återstår att hitta $\hat{\theta}_{ML}$ och $I(\hat{\theta}_{ML})$.

Likelihoodfunktionen och log-likelihoodfunktionen för binomialfördelningen ges av

$$L(\theta|y) = \theta^y(1-\theta)^{n-y}$$
$$l(\theta|y) = y\log(\theta) + (n-y)\log(1-\theta)$$

För att få Score-funktionen $u(\theta)$ deriverar vi log-likelihoodfunktionen

$$u(\theta) = \tfrac{y}{\theta} - \tfrac{n-1}{1-\theta}$$

ML-skattningen för θ ges genom att ansätta Score-funktionen till noll och lösa ut θ. Vi får följande uttryck för $\hat{\theta}_{ML}$

$$\begin{aligned} 0 = u(\theta) &= \frac{y}{\theta} - \frac{n-y}{1-\theta} \\ \frac{y}{\theta} &= \frac{n-y}{1-\theta} \\ \hat{\theta}_{ML} &= \frac{y}{n} \end{aligned}$$

Det återstår att hitta Fisher informationsmatrisen, det vill säga $I(\theta)$, som ges av den negativa derivatan av Score-funktionen

$$I(\theta) = \frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}$$

Vi byter ut θ mot $\hat{\theta}_{ML} = \frac{y}{n}$ och får följande uttryck

$$\begin{aligned} I(\hat{\theta}_{ML}) &= \frac{y}{\left(\frac{y}{n}\right)^2} - \frac{n-y}{\left(1-\frac{y}{n}\right)^2} = \frac{n^2}{y} + \frac{n^2}{n-y} \\ &= \frac{n^3}{y(n-y)} = n\left(\frac{n}{y}\right)\left(\frac{n}{n-y}\right), \end{aligned}$$

där vi i sista steget kan skriva om ekvationen som

$$I(\hat{\theta}_{ML}) = n(\hat{\theta}_{ML})^{-1}(1-\hat{\theta}_{ML})^{-1}$$

Vi kan nu ersätta $\hat{\theta}_{ML}$ och $I(\hat{\theta}_{ML})$ i Wald-konfidensintervallet och får

$$\hat{\theta}_{ML} \pm \lambda_{0.025}\sqrt{\frac{(\hat{\theta}_{ML})(1-\hat{\theta}_{ML})}{n}},$$

som är konfidensintervallet för en binomialfördelad stokastisk variabel[9].

6.2 Regularisering för logistisk regression

Ett alternativ för att förbättra modellens prediktionsförmåga är att använda en metod som begränsar skattningarna. Det vill säga metoden krymper skattningarna mot noll som i sin tur leder till minskad varians. Nedan ges en kort beskrivning av idén bakom *Ridge logistisk regression*.

Ridge log likelihoodfunktion definieras som

$$l_{ridge}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\lambda\sum_{i=1}^{N}\beta_i^2, \tag{7}$$

där $l(\boldsymbol{\beta})$ är den 'ostraffade' log likelihooden och λ är en regulariseringskonstant som bestämmer hur mycket variablerna ska modifieras. Ridge skattningarna $\hat{\boldsymbol{\beta}}_{ridge}$ följer från (7) enligt

$$\hat{\boldsymbol{\beta}}_{ridge} = argmax(l_{ridge}(\boldsymbol{\beta})).$$

λ, som är vikten av den regulariserade termen, styr storleken på koefficienterna. Att summan i ekvationen (7) startar från ett och inte noll beror på att vi inte vill minska interceptet. Eftersom interceptet inte påverkar kurvans lutning utan endast läget, finns det ingen anledning att bestraffa den. Vidare används korsvalidering för att välja den mest passande λ för att sedan estimera klassifikationsfelet. För att de kontinuerliga variablerna ska straffas likvärdigt används standardisering. På det viset får samtliga variabler standardavvikelse ett och som resultat kommer den slutliga modellen bli oberoende av skala. För det ändamål används formeln

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_{ij} - \bar{x})^2}},$$

där nämnaren motsvarar en skattad standardavvikelse av j:te prediktor i observation i[16].

6.3 Tabeller med kategorisering av förklaringsvariabler samt koefficientskattningar

Personbil	Personbil
Person	Fotgängare, Cyklist
Tungfordon	Lastbil (lätt), Lastbil (tung), Lastbil (okänd), Traktor, Tåg, Buss, Spårvagn
Djur	Rådjur/Hjort, Älg, Häst, Övriga tamdjur, Övrigt vilt, Ren, Okänt djur, Nötkreatur, Vildsvin, Dovhjort, Kronhjort
Motorcykel	Motorcykel (lätt), Moped (klass 2), Moped (okänd), Motorcykel (tung), Moped (klass 1), Motorcykel (okänd)
Övrigt	Terrängskoter, Terränghjuling, Övrigt fordon, Snöskoter, Motorredskap, Släp, Terrängvagn, Okänt fordon

Raksträcka	Gatu-/Vägsträcka, Gångbana/trottoar, Gång- och cykel-bana/väg
Korsning	Cirkulationsplats, Gatu-/Vägkorsning
Annat	Trafikplats, Annan

Singelolycka	G0 (fotgängare singel), S (singel-motorfordon), G2 (moped singel), G1 (cykel singel), F (fotgängare-motorfordon), G2 (moped singel), G6 (moped-fotgängare), G3 (fotgängare-cyklist), G5 (cykel-moped), G8 (fotgängare-fotgängare), G4 (cykel-cykel)
Annan olycka	U (upphinnande-motorfordon), A (avsvängande motorfordon), C (cykel/moped-motorfordon), O (omkörning-motorfordon), K (korsande-motorfordon), M (möte-motorfordon), V5 (parkerat fordon), V3 (traktor/snöskoter/terränghjuling/motorredskap), V6 (backning/vändning/u-sväng), V0 (övrigt), G7 (moped-moped), W2 (älg), W1 (rådjur/hjort), J (spårvagn), W5 (vildsvin), W4 (annat vilt), J (tåg), W3 (ren), V1 (djur, häst/annat tamdjur), J (tåg/spårvagn övrigt)

Väglag	Okänt, Vägbanan torr, Vägbanan våt/fuktig, Lös snö/snömodd, Tunn is (vägbanan synlig), Tjock is/packad snö
Väderlek	Okänt, Uppehållsväder, Regn, Snöfall, Snöblandat regn, Dis/dimma
Ljusförhållande	Okänt, Dagsljus, Mörker, Gryning/Skymning

Figur 7: Kategorisering av förklaringsvariabler.

	term	estimate	std.error	statistic	p.value
1	(Intercept)	35.5608602	8.048e+00	4.419	9.934e-06
2	Tid	-0.0187099	4.006e-03	-4.670	3.012e-06
3	I(sin(2 * pi * imputeraddata$M/12))	-0.0592357	2.318e-02	-2.556	1.060e-02
4	I(cos(2 * pi * imputeraddata$M/12))	-0.0658065	2.407e-02	-2.733	6.267e-03
5	PKorsning	-0.3838850	8.890e-02	-4.318	1.572e-05
6	PVäg	0.1688883	8.261e-02	2.044	4.092e-02
7	OSingel olycka	-0.1191833	3.394e-02	-3.512	4.452e-04
8	LjusGryning/skymning	0.0683463	6.003e-02	1.139	2.549e-01
9	LjusMörker	0.2541463	3.863e-02	6.579	4.734e-11
10	Lon	-0.0021285	1.485e-04	-14.338	1.275e-46
11	Lat	0.0007167	7.699e-05	9.309	1.287e-20
12	Per	0.2373189	6.525e-02	-3.637	2.758e-04
13	Tung	0.5283548	4.655e-02	11.349	7.493e-30
14	Ö	0.5441568	1.235e-01	4.405	1.058e-05

Figur 8: Slutlig modell.

	term	estimate	std.error	statistic	p.value
1	(Intercept)	3.753e+01	8.273e+00	4.53635	5.724e-06
2	Tid	-1.953e-02	4.116e-03	-4.74565	2.078e-06
3	I(sin(2 * pi * df$M/12))	-7.424e-02	2.522e-02	-2.94435	3.236e-03
4	I(cos(2 * pi * df$M/12))	-8.619e-02	3.007e-02	-2.86666	4.148e-03
5	P:Korsning	-1.063e-01	1.003e-01	-1.05886	2.897e-01
6	P:Väg	3.226e-01	9.316e-02	3.46313	5.339e-04
7	OC (cykel/moped-motorfordon)	-4.679e-01	1.007e-01	-4.64720	3.365e-06
8	OF (fotgängare-motorfordon)	3.371e-01	9.747e-02	3.45872	5.427e-04
9	OG0 (fotgängare singel)	3.885e-01	6.490e-01	0.59867	5.494e-01
10	OG1 (cykel singel)	3.819e-01	1.675e-01	2.28037	2.259e-02
11	OG2 (moped singel)	-1.202e+00	1.791e-01	-6.71158	1.925e-11
12	OG3 (fotgängare-cyklist)	-5.528e-01	3.114e-01	-1.77497	7.590e-02
13	OG4 (cykel-cykel)	-1.127e+00	3.450e-01	-3.26569	1.092e-03
14	OG5 (cykel-moped)	-1.818e+00	5.152e-01	-3.52816	4.185e-04
15	OG6 (moped-fotgängare)	-1.362e+01	1.168e+02	-0.11668	9.071e-01
16	OG7 (moped-moped)	-2.985e+00	1.010e+00	-2.95683	3.108e-03
17	OG8 (fotgängare-fotgängare)	-5.091e-01	1.051e+00	-0.48433	6.282e-01
18	OJ (spårvagn)	-1.382e+01	2.446e+02	-0.05650	9.549e-01
19	OJ (tåg)	2.219e+00	2.319e-01	9.57034	1.066e-21
20	OJ (tåg/spårvagn övrigt)	1.718e+01	1.455e+03	0.01180	9.906e-01
21	OK (korsande-motorfordon)	1.373e-02	1.006e-01	0.13654	8.914e-01
22	OM (möte-motorfordon)	1.067e+00	9.674e-02	11.03243	2.665e-28
23	OO (omkörning-motorfordon)	-8.254e-01	2.377e-01	-3.47332	5.141e-04
24	OS (singel-motorfordon)	-1.574e-01	9.209e-02	-1.70926	8.740e-02
25	OU (upphinnande-motorfordon)	-1.264e+00	1.262e-01	-10.01843	1.265e-23
26	OV0 (övrigt)	-5.279e-02	1.533e-01	-0.34441	7.305e-01
27	OV1 (djur, häst/annat tamdjur)	8.754e-02	4.230e-01	0.20697	8.360e-01
28	OW1 (rådjur/hjort)	-1.455e+00	5.189e-01	-2.80377	5.051e-03
29	OW2 (älg)	-2.482e-01	1.509e-01	-1.64484	1.000e-01
30	OW3 (ren)	-1.362e+01	4.177e+02	-0.03261	9.740e-01
31	OV3 (traktor/snöskoter/terränghjuling/motorredskap)	-8.084e-02	1.596e-01	-0.50638	6.126e-01
32	OW4 (annat vilt)	-5.639e-01	6.131e-01	-0.91985	3.576e-01
33	OV5 (parkerat fordon)	-2.799e-01	2.157e-01	-1.29784	1.943e-01
34	OW5 (vildsvin)	5.729e-01	1.124e+00	0.50978	6.102e-01
35	OV6 (backning/vändning/u-sväng)	-2.426e-01	1.722e-01	-1.40843	1.590e-01
36	VägOkänt	-2.531e-01	1.747e-01	-1.44902	1.473e-01
37	VägTjock is / packad snö	-1.350e-01	1.293e-01	-1.04406	2.965e-01
38	VägTunn is, vägbanan synlig	1.463e-01	1.189e-01	1.23031	2.186e-01
39	VägVägbanan torr	3.848e-01	1.143e-01	3.36619	7.621e-04
40	VägVägbanan våt/fuktig	3.804e-01	1.132e-01	3.35956	7.807e-04
41	VädOkänt	4.398e-02	1.663e-01	0.26451	7.914e-01
42	VädRegn	-1.897e-01	1.151e-01	-1.64776	9.940e-02
43	VädSnöblandat regn	-3.533e-01	1.875e-01	-1.88409	5.955e-02
44	VädSnöfall	-4.369e-02	1.516e-01	-0.28828	7.731e-01
45	VädUppehållsväder	-1.373e-01	1.038e-01	-1.32224	1.861e-01
46	LjusGryning/skymning	8.570e-02	6.233e-02	1.37497	1.691e-01
47	LjusMörker	3.011e-01	4.110e-02	7.32624	2.367e-13
48	LjusOkänt	3.817e-01	1.070e-01	3.56822	3.594e-04
49	Lon	-1.893e-03	1.526e-04	-12.40880	2.341e-35
50	Lat	6.587e-04	8.104e-05	8.12817	4.358e-16
51	Pbil	-7.676e-01	1.521e-01	-5.04521	4.530e-07
52	Tung	-2.094e-01	1.534e-01	-1.36524	1.722e-01
53	Per	-9.762e-01	1.731e-01	-5.64030	1.698e-08
54	Djur	-3.662e-01	5.161e-01	-0.70945	4.780e-01
55	Moto	-5.476e-01	1.587e-01	-3.45017	5.602e-04

Figur 9: Utgångsmodell.

	OR	2.5 %	97.5 %
(Intercept)	2.779e+15	3.995e+08	2.009e+22
Tid	9.815e-01	9.738e-01	9.892e-01
I(sin(2 * pi * imputeraddata$M/12))	9.425e-01	9.006e-01	9.862e-01
I(cos(2 * pi * imputeraddata$M/12))	9.363e-01	8.931e-01	9.815e-01
P:Korsning	6.812e-01	5.736e-01	8.129e-01
P:Väg	1.184e+00	1.010e+00	1.396e+00
O:Singel olycka	8.876e-01	8.305e-01	9.487e-01
Ljus:Gryning/skymning	1.071e+00	9.506e-01	1.203e+00
Ljus:Mörker	1.289e+00	1.195e+00	1.391e+00
Lon	9.979e-01	9.976e-01	9.982e-01
Lat	1.001e+00	1.001e+00	1.001e+00
Per	1.137e+00	1.027e+00	1.251e+00
Tung	1.696e+00	1.547e+00	1.857e+00
Ö	1.723e+00	1.345e+00	2.185e+00

Figur 10: Konfidensintervall för oddskvoten.

Referenser

[1] Alan Agresti. *Categorical Data Analysis, 2nd Edition*. A John Wiley and Sons, Inc, 2002.

[2] Matthias Templ Alexander Kowarik. Imputation with the r package vim. *Journal of Statistical Software*, 2016.

[3] Larsw-Christer Böiers Arne Persson. *Analys i en variabel*. Studentlitteratur AB, 2001.

[4] Ulf Brude. Sveriges trafiksakerhet i ett 100-årigt persepektiv. *Mitt i trafiken*, 2013.

[5] Michael J. Crawley. *The R Book, 2nd Edition*. Michael J. Crawley, 2012.

[6] NTF (Nationalförening för Trafiksäkerhetens Främjande). Myter och sanningar om hastighet. *NTF, säker trafik*, 2004.

[7] Lantmäteriet Informationsförsörjning Geodesi. Gps och geodetisk mätning.

[8] Allan Gut. *An Intermideate Course In Probability, 2nd Edition*. Springer, 2009.

[9] Daniel Sabanés Bové Leonhard Held. *Applied Statistical Inference. Likelihood and Bayes*. Springer, 2014.

[10] Violence Management of Noncommunicable Diseases, Disability and Injury Prevention (NVI). Global status report on road safety 2015. *World Health Organization*, 2015.

[11] Mattias Rabe. Lägsta antalet döda i trafiken sedan 1944. *Teknikens Värld. Allt om bilar*, 2014.

[12] J. Ranstam. Why the p-value culture is bad and confidence intervals a better alternative. *Osteoarthritis and Cartilage*, 2012.

[13] Rolf Sundberg. *Lineära Statistiska Modeller*. Stockholm Universitet, 2016.

[14] Kaj Sundström. Personskador i trafiken, strada, värmland 2007-2012. *Transportstyrelsen*, 2013.

[15] Tom Britton Sven Erick Alm. *Stokastik - Sannolikhetsteori och statistikteori med tillämpningar*. Liber, 2008.

[16] Gareth James Daniella Witen Trevor Hastie Robert Tibshirani. *An Introduction To Statistical Learning*. Springer, 2013.

[17] Jerome Friedman Trevor Hastie Robert Tibshirani. *The Elements Of Statistical Learning: Data Mining, Inference And Prediction, 2nd Edition*. Springer, 2008.

[18] Trafikverket. Trafiksäkerhetsmål.

[19] Gerard L'Estrange Turner. Ward [née king], mary [pseud. the hon. mrs ward] (1827–1869), microscopist and author. *Oxford Dictionary of National Biography*, 2004.