



SINTEF



# Rapport

## Utvikling av ulykkesmodeller for riks- og fylkesvegnettet i Norge (2016-2021)

### Forfatter:

Gunhild Elisabeth Berget

### Rapportnummer:

2024:00497 - Åpen

### Oppdragsgiver:

Statens Vegvesen



# Rapport

## Utvikling av ulykkesmodeller for riks- og fylkesvegnettet i Norge (2016-2021)

[Klikk eller trykk her for å skrive inn tekst.]

EMNEORD  
Trafikksikkerhet

VERSJON  
1

DATO  
2024-04-22

FORFATTER  
Gunhild Elisabeth Berget

OPPDRAGSGIVER  
Statens Vegvesen

OPPDRAGSGIVERS REFERANSE  
Arild Engebretsen

PROSJEKTNUMMER  
102027589

ANTALL SIDER  
21

### SAMMENDRAG

Rapporten beskriver ulykkesmodeller basert på data for riks- og fylkesvegnettet fra Nasjonal Vegdatabank (NVDB) for perioden 2016-2021. Modellene beregner det normale antallet personskadeulykker, lettere skadde, hardt skadde, drepte, samt drepte eller hardt skadde langs norske veger. Modellene er videreutviklet fra tidligere ulykkesmodeller basert på samme vegnett, og kan benyttes for å identifisere risikofaktorer og utvikle sikkerhetstiltak. Beregningene er basert på negative binomiale modeller, som tar hensyn til overspredning i ulykkesdata. Rapporten oppdaterer tidligere modeller ved å inkludere nye ulykkesdata og gjøre noen justeringer i prediktorvariablene.

UTARBEIDET AV  
Gunhild Elisabeth Berget

SIGNATUR  
*Gunhild E Berget*

KONTROLLERT AV  
Anders Kroksæter

SIGNATUR  
*Anders Kroksæter*  
Anders Kroksæter (24. apr. 2024 13:59 GMT+2)

GODKJENT AV  
Unn Karin Thorenfeldt

SIGNATUR  
*Unn Karin Thorenfeldt*

RAPPORT NR.  
2024:00497

ISBN  
978-82-14-07207-5

GRADERING  
Åpen

GRADERING DENNE SIDE  
Åpen

# Historikk

VERSJON	DATO	VERSJONSBESKRIVELSE
0.7	2023-06-01	Rapportutkast til gjennomlesing
0.9	2023-07-07	Gjennomlest og oppdatert rapport
1.0	2024-04-22	Ferdig rapport

# Innholdsfortegnelse

<b>1</b>	<b>Introduksjon .....</b>	<b>5</b>
<b>2</b>	<b>Ulykkesdata .....</b>	<b>6</b>
2.1	Antall ulykker, drepte og skadde .....	6
2.2	Segmentlengde og antall år .....	7
2.3	Trafikkmengde .....	7
2.3.1	Kommentar til bruk av ÅDT i modellene .....	8
2.4	Fartsgrense.....	9
2.4.1	Kommentarer til fartsgrenser .....	9
2.5	Antall kjørefelt .....	9
2.5.1	Kommentar til antall kjørefelt .....	9
2.6	Kryss, rundkjøringer og ramper .....	10
2.7	Type veg .....	11
2.8	Midtdeler og midtrekkverk .....	11
2.9	Forsterket midtoppmerking.....	12
2.10	Automatisk trafikkontroll (ATK) .....	12
2.11	Vegbelysning .....	13
2.12	Fylke .....	13
2.13	Oppsummering av prediktorvariablene brukt i regresjonsmodellen .....	13
2.13.1	Dummyvariabler .....	13
<b>3</b>	<b>Regresjonsmodell for normale ulykkestall .....</b>	<b>14</b>
3.1	Modellform .....	14
3.1.1	Negativ binomial fordeling .....	15
3.1.2	Bruk av offsetvariabel .....	15
3.1.3	Variabel overspredning .....	15
3.2	Estimerte koeffisienter .....	16
3.2.1	Kommentar til modellene.....	18
3.3	Estimert variabel overspredning.....	19
3.4	Hvor gode er modellene?.....	19
<b>4</b>	<b>Referanser .....</b>	<b>21</b>



## 1 Introduksjon

Rapporten gir en detaljert beskrivelse av ulykkesmodeller som er utviklet basert på data for riks- og fylkesvegnettet fra Nasjonal Vegdatabank (NVDB) for perioden 2016-2021. Modellene brukes til å beregne det normale antall personskadeulykker (PSU), lettere skadde (LS), hardt skadde (HS), drepte (DR), samt drepte eller hardt skadde (DR/HS). Rapporten bygger på det grunnleggende arbeidet som tidligere er gjort med utvikling av ulykkesmodeller for det norske vegnettet (Ragnøy, Christensen & Elvik, 2002; Ragnøy & Elvik, 2003; Høye, 2014; Høye, 2016). Disse modellene er særlig nyttige for vegmyndigheter og forskere som har som mål å forbedre trafikksikkerheten. Gjennom modellering av antall ulykker langs spesifikke vegstrekninger, kan risikofaktorer identifiseres og tiltak kan utvikles for å redusere ulykkesfrekvensen, med spesielt fokus på ulykker med høy alvorlighetsgrad. Ulykkesmodellene er et sentralt verktøy for å identifisere hvilke deler av vegnettet som vil ha mest nytte av sikkerhetstiltak. Gjennom å identifisere mønstre og sammenhenger mellom vegelementer, trafikkmengde og ulykkesforekomster, kan modellene bidra til å prioritere innsatsen for trafikksikkerhet der den er mest nødvendig.

Modellene beregner det normale antall ulykker som en funksjon av trafikkmengde og ulike egenskaper i vegnettet. Med "normale ulykkestall" refereres det til gjennomsnittlig antall ulykker på et vegsegment av en bestemt lengde, med bestemte egenskaper, gitt trafikkmengde, over en bestemt tidsperiode. Det er benyttet negative binomiale modeller med variabel overspredningsparameter for å beregne disse tallene. Negative binomiale modeller med variabel overspredningsparameter er særlig egnet for å analysere ulykkesdata, fordi de tar høyde for overspredning. Dette er viktig fordi trafikkulykker ikke nødvendigvis distribueres jevnt over tid og sted.

Datasettet er oppdelt i homogene strekninger, altså har segmentene samme egenskaper når det gjelder fartsgrense, antall felt, type veg osv. Dette konseptet er viktig i utforming og analyse av ulykkesmodeller fordi det gir en konsistent måte å vurdere og forutsi risiko og ulykker. Når man deler opp vegnettet i homogene strekninger, sikrer man at hver enhet (strekning) som analyseres, har samme egenskaper.

Rapporten bygger på tidligere modeller utviklet i Høye (2016), men inneholder oppdateringer basert på nye data og justeringer i modellene. De viktigste endringene i de oppdaterte modellene er som følger:

- Bruk av ulykkesdata fra 2016-2021.
- I modellene for PSU og LS er det flere kategorier for fart, inkludert egne kategorier for fartsgrense 100 og 110.
- I modellene for HS, DR og HS/DR er det flere kategorier for fart. 90 er nå lagt til som egen kategori.
- I modellen for DR er alle flere enn 4 kjørefelt slått sammen til kategorien Kjørefelt 4+.
- SATK (en retning) og SATK (begge retninger) er nå kombinert til en enkelt kategori, SATK.
- Nord-Trøndelag og Sør-Trøndelag er slått sammen til et enkelt fylke, Trøndelag.
- Kryss, rundkjøringer og ramper inngår i modellen som dummyvariabler.

Utover disse endringene har de samme prediktorvariablene, samme modellform og de samme vegkategoriene blitt benyttet som i de tidligere modellene.

## 2 Ulykkesdata

Ulykkesmodellene er beregnet basert på data fra NVDB, for riks- og fylkesvegnettet og årene 2016-2021. I dette avsnittet blir datasettet gjennomgått, og deskriptiv statistikk blir presentert for de variablene som inngår i regresjonsmodellen.

Datasettet er oppdelt i homogene strekninger, altså i vegsegmenter som har samme egenskaper når det gjelder viktige parametere som fartsgrense, antall kjørefelt, type veg osv.

Det totale datasettet består av 97100 observasjoner av 47 variabler. Til å begynne med er det ryddet i datasettet. Følgende rader er slettet (tallene i parentes beskriver antall segmenter dette gjelder for):

- Segmenter med manglende ÅDT (**2291**)
- Segmenter med manglende lengde (**0**)
- Segmenter med manglende fartsgrense (**6**)
- Segmenter med manglende år (**0**)
- Segmenter der ÅDT er lavere enn 10 (**314**)
- Segmenter med kun ett kjørefelt (**5510**)
- Segmenter som er arm, kryss eller rampe (**9028**)
- Segmenter som er tunneler (**2716**)
- Segmenter som er bruer (**976**)
- Segmenter med fartsgrense satt til 20 km/t (**2**)

Etter sletting av segmenter står vi igjen med et datasett med totalt 84264 observasjoner av 47 variabler. Totalt 12836 segmenter ble fjernet i ryddingen av datasettet. Basert på tallene for de fjernede segmentene, er det tydelig at flere av segmentene har manglende verdier for flere av variablene. Det totale antallet fjernede segmenter er dermed ikke lik summen av de individuelle fjernede segmentene for hver variabel.

Videre i kapittelet blir det presentert deskriptiv statistikk for prediktorvariablene som inngår i regresjonsmodellen.

### 2.1 Antall ulykker, drepte og skadde

I datasettet er ulykker, skadde og drepte delt inn i åtte kategorier:

- Antall drepte (DR)
- Antall ulykker med drepte (UDR)
- Antall meget alvorlig skadde (MAS)
- Antall ulykker med meget alvorlig skadde (UMAS)
- Antall alvorlig skadde (AS)
- Antall ulykker med alvorlig skadde (UAS)
- Antall lett skadde (LS)
- Antall ulykker med lett skadde (ULS)

Dersom en ulykke inneholder skadde/drepte innen flere kategorier telles den i gruppen med den alvorligste skadegraden.

I denne rapporten utvikles det fem ulike modeller for å beregne det normale antallet ulykker på de inkluderte vegsegmentene. Disse relateres til tallene over på følgende måte:



- **Personskadeulykker (PSU):** Her brukes det summen UDR+UMAS+UAS+ULS. Altså det totale antall ulykker langs vegen.
- **Lett skadde (LS):** Her benyttes tallene for LS
- **Hardt skadde (HS):** Her benyttes summen MAS+AS
- **Drepte (DR):** Her benyttes tallene for DR
- **Drepte eller hardt skadde (DR/HS):** Her benyttes summen DR+AS+MAS

Deskriptiv statistikk for variablene over kan sees i Tabell 1.

**Tabell 1: Deskriptiv statistikk for antall personskadeulykker (PSU), antall lett skadde (LS), antall hardt skadde (HS), antall drepte (DR) og antall drepte/hardt skadde (DR/HS).**

	PSU	LS	HS	DR	DR/HS
<b>Totalt antall</b>	12464	13733	2118	417	2535
<b>Min</b>	0	0	0	0	0
<b>Maks</b>	11	16	8	5	8
<b>Antall segmenter med null</b>	74731	75860	82451	83876	82151
<b>Andel segmenter med null</b>	88,7%	90,0%	97,8%	99,5%	97,5%

Fra tabellen kan det sees at de fleste segmenter inneholder ingen drepte eller skadde, som betyr at ulykker langs vegen er en relativ sjelden hendelse og burde hensyntas i utvikling av modellene.

## 2.2 Segmentlengde og antall år

Tabell 2 viser deskriptiv statistikk for segmentlengde og periodelengde for de i alt 84264 segmentene. Segmentlengde beskriver lengden av vegsegmentet i meter. Periodelengden beskriver antall år segmentet gjelder for. Vi ser at det benyttes data for en total strekning på 51813 km veg. Den opprinnelige filen med 97001 segmenter inneholdt data for 55392 km veg. Altså er 3579 km fjernet fra den opprinnelige filen (pga. broer, ramper, manglende data osv.).

**Tabell 2: Deskriptiv statistikk for segmentlengde og periodelengde (uvektet).**

	Segmentlengde (m)	Periodelengde (antall år)
<b>Sum</b>	51 812 826	
<b>Gjennomsnitt</b>	615	5,50
<b>Standardavvik</b>	389	1,31
<b>Min.</b>	1	0,02
<b>Maks</b>	1200	6,00

Den gjennomsnittlige periodelengden vektet med lengden på segmentet er også beregnet. Vi får da en snittlengde på 5,62 år. Vi ser altså at majoriteten av segmentene gjelder for størstedelen av den totale perioden på 6 år. På grunn av kutting i datasettet for å oppnå homogene segmenter (med ensartede egenskaper i perioden) vil det forekomme segmenter som gjelder for en kortere tidsperiode.

## 2.3 Trafikkmengde

Trafikkmengde blir målt i ÅDT (ÅrsDøgnTrafikk) og er i prinsippet summen av antall kjøretøy som passerer et punkt på en vegstrekning (for begge retninger sammenlagt) gjennom året, dividert på årets dager, altså et gjennomsnittstall for daglig trafikkmengde. Deskriptiv statistikk for ÅDT er vist i Tabell 3.

Tabell 3: Deskriptiv statistikk for trafikkmengde (ÅDT).

	ÅDT	ÅDT (Høye, 2016)
Gjennomsnitt (uvektet)	2678	2393
Standardavvik	6034	5555
Min.	10	5
Maks	92584	98043
Gjennomsnitt (vektet med segmentlengde)	1727	1733

I tabellen ser vi at gjennomsnittlig uvektet ÅDT er 2678 mens gjennomsnittet vektet med segmentlengde er 1727. Dette tyder på at segmentene med størst lengde oftere har lavere ÅDT. Fra standardavviket i tabellen kan det sees at det er stor variasjon på ÅDT. Dette sees også fra min og maks-verdiene som spriker fra 5 til 92584. Tabell 4 viser hyppighetsfordelingen for ÅDT.

Tabell 4: Hyppighetsfordeling av ÅDT.

ÅDT	Antall segmenter	Andel segmenter	Kumulativ andel
<100	5700	6,8%	6,8%
100-199	9847	11,7%	18,4%
200-399	14020	16,6%	35,0%
400-599	7924	9,4%	44,5%
600-799	5018	6,0%	50,4%
800-999	4726	5,6%	56,1%
1000-1999	12554	14,9%	71,0%
2000-2999	6851	8,1%	79,1%
3000-3999	3912	4,6%	83,7%
4000-4999	2626	3,1%	86,8%
5000-7499	3702	4,4%	91,2%
7500-9999	2294	2,7%	94,0%
10000-24999	4019	4,8%	98,7%
25000-49999	768	0,9%	99,6%
50000-74999	233	0,3%	99,9%
75000-	70	0,1%	100%
SUM	84264	100%	

### 2.3.1 Kommentar til bruk av ÅDT i modellene

I (Høye, 2016) blir det gjort testberegninger for å se hvordan ÅDT bør inngå i modellen. Flere forskjellige varianter har blitt utprøvd, blant annet å inkludere  $\ln(\text{ÅDT})$  og  $\ln(\text{ÅDT})^2$ . Resultatene viste kun en liten forbedring i modellen ved å inkludere  $\ln(\text{ÅDT})^2$ . I tillegg medfører det en ulempe i den praktiske bruken av modellresultatene ved at den prosentvise økningen av predikerte ulykker ved økning av trafikkmengden vil endre seg med trafikkmengden. I de aktuelle modellene er det derfor  $\ln(\text{ÅDT})$  som inngår som prediktorvariabel slik konklusjonen også ble i (Høye, 2016).

## 2.4 Fartsgrense

Fartsgrensen inngår i modellen som en dummyvariabel. Altså blir variabelen satt til 1 for den respektive fartsgrensen og 0 ellers. 80 km/t blir satt som referansekategori. Tabell 5 viser deskriptiv statistikk for fartsgrenser.

Tabell 5: Deskriptiv statistikk for fartsgrenser.

Fartsgrense	Antall segmenter	Andel av segmenter	Samlet lengde (km)	Andel lengde	Snitt ÅDT (vektet)	Millioner kjøt.km per år	Andel av all kjøt.km
30	1180	1,40%	326	0,63%	1828	217	0,67%
40	3753	4,45%	1115	2,15%	3037	1236	3,78%
50	12856	15,26%	4856	9,37%	2225	3944	12,1%
60	16777	19,91%	8450	16,31 %	1852	5711	17,5%
70	3938	4,67%	1839	3,55%	5243	3519	10,8%
80	43006	51,04%	33800	65,24%	1057	13044	39,9%
90	1557	1,84%	995	1,92%	4710	1711	5,2%
100	489	0,58%	171	0,33%	24355	1518	4,6%
110	708	0,84%	261	0,50%	18505	1766	5,4%
SUM	84246	100%	51813	100%	1727	32666	100%

### 2.4.1 Kommentarer til fartsgrenser

I (Høye, 2016) ble fartsgrensene 100 og 110 slått sammen til en kategori i modellene for PSU og LS. Med de oppdaterte dataene kan vi se at det har blitt flere segmenter innenfor disse fartsgrensene, og vi velger å bruke de som individuelle kategorier i de estimerte modellene i denne rapporten.

For modellene for HS, Drepte og DR/HS velger vi å slå sammen fartsgrensene 100 og 110 til en kategori. Dette gjøres fordi antallet drepte og hardt skadde er svært få, og det finnes ikke signifikante sammenhenger mellom antall hardt skadde eller drepte for disse små kategoriene. I (Høye, 2016) ble også 90 inkludert i denne kategorien, men vi velger i denne versjonen å la 90 være en individuell kategori.

## 2.5 Antall kjørefelt

Også for antall kjørefelt er det definert en dummyvariabel slik som for fartsgrensen. Antall kjørefelt varierer fra 2 til 8 og deskriptiv statistikk er vist i Tabell 6. Segmenter med kun ett kjørefelt er fjernet fra datasettet fordi dette er en liten og svært heterogen gruppe av helt ulike typer vegger.

### 2.5.1 Kommentar til antall kjørefelt

I (Høye, 2016) ble segmenter med seks eller flere kjørefelt slått sammen til en dummyvariabel for modellene for PSU og LS og i modellene for HS, Drepte og DR/HS er alle segmenter med fem eller flere kjørefelt slått sammen. Det samme blir gjort for modellene i denne rapporten.

Tabell 6: Deskriptiv statistikk for antall kjørefelt.

Kjørefelt	Antall segmenter	Andel av segmenter	Samlet lengde (km)	Andel lengde	Snitt ÅDT (vektet)	Millioner kjt.km per år	Andel av alle kjt.km
2	80899	96,00%	50897	98,23%	1384	25710	78,7%
3	897	1,06%	212	0,41%	8936	692	2,1%
4	2172	2,58%	644	1,24%	21596	5080	15,6%
5	178	0,21%	32	0,06%	43270	505	1,5%
6	80	0,09%	17	0,03%	67776	432	1,3%
7	21	0,03%	2	0,004%	63745	53	0,2%
8	17	0,02%	7	0,01%	73784	193	0,6%
<b>SUM</b>	84264	100%	51813	100%	1727	32666	100%

## 2.6 Kryss, rundkjøringer og ramper

Prediktorvariabler for X-kryss, T-kryss, rundkjøringer, av-ramper, på-ramper og uspesifiserte ramper inngår i modellen som dummyvariabler. Dersom det finnes ett eller flere kryss langs segmentet blir variabelen satt lik 1, ellers er den lik 0.

**Plankryss (X- og T-kryss):** X-kryss er firearmede plankryss og T-kryss er trearmede plankryss. En del av T-kryssene i datamaterialet kan være avkjørsler. Det skilles ikke mellom ulike typer kryssregulering. Kryss er i datamaterialet alltid lagt til hovedvegen i krysset. Det betyr at sidevegene i datamaterialet ikke har noen kryss (ulykkene er som regel uansett stedfestet på hovedvegen).

**Rundkjøringer:** Rundkjøringer omfatter alle typer rundkjøringer. Det er gjort en opptelling av antall rundkjøringer som ligger på hvert segment. På samme måte som for plankryss er rundkjøringene alltid lagt til hovedvegen. Ulykkene som skjer i rundkjøringer, er da også lagt til nærmeste segment på hovedvegen.

**Ramper:** Av- og på-ramper er ramper til og fra (hoved-)veger med planskilte kryss. Ramper finnes både på motorveger/tofeltsveger med planskilte kryss og på andre typer veger. Som for plankryss og rundkjøringer beskriver rampevariablene antall ramper som ligger i tilknytning til hvert segment, selve rampen inngår ikke i datamaterialet.

Tabell 7: Deskriptiv statistikk for kryss, rundkjøringer og ramper. Tabellen viser antall kryss/rundkjøringer, andel av totale kryss/rundkjøringer og kryss/rundkjøringer per km etter vegtype. Vegtyper er definert i neste avsnitt.

	X-kryss			T-kryss			Rundkjøringer		
	Antall	Andel	Per km	Antall	Andel	Per km	Antall	Andel	Per km
Motorveg	0	0%	0	0	0%	0	1	0,0%	0,002
Tofelts-planskilt	0	0%	0	0	0%	0	7	0,3%	0,021
Øvrig TEN-T	38	2,3%	0,011	1786	4,7%	0,495	144	6,3%	0,040
Øvrig Ev/Rv	115	6,9%	0,026	2628	6,9%	0,597	263	11,5%	0,060
Fylkesveg	1505	90,8%	0,035	33868	88,5%	0,788	1881	82,0%	0,044
<b>Alle veger</b>	1658	100%	0,032	38282	100%	0,739	2296	100%	0,044

Sammenhengen mellom plankryss, rundkjøringer og ramper, og antall ulykker kan være forskjellig avhengig av bl.a. fartsgrense, vegtype, om vegen har skille mellom kjøretningene eller ikke, og kurver og stigninger. Slike interaksjonseffekter er imidlertid ikke tatt hensyn til i modellene.

## 2.7 Type veg

Det er definert fem dummyvariabler for ulike typer veg:

1. **Motorveg**
2. **Tofelts-veg med planskilte kryss** (motortrafikkveg i tidligere terminologi)
3. **TEN-T-veg for øvrig** (ikke motorveg eller tofelts-veg med planskilte kryss)
4. **Europa-/riksveg (Ev/Rv) for øvrig** (ikke motorveg eller tofelts-veg med planskilte kryss eller TEN-T veg)
5. **Fylkesveg** (referansekategori)

Vegtypene er valgt fordi det finnes generelle forskjeller i vegstandarden, blant annet med tanke på kjørefelt- og skulderbredde. F.eks. er det slik at alle motorveger skal ha midtrekkverk. Det forventes også at TEN-T-veger har høyere standard enn øvrige Europa- og riksveger. Deskriptiv statistikk for type veg kan sees i Tabell 8.

**Tabell 8: Deskriptiv statistikk for type veg.**

Type veg	Antall segmenter	Andel av segmenter	Sum lengde (km)	Andel av veglengde	Snitt ÅDT (vektet)	Millioner kjt.km per år	Andel av alle kjt.km
Motorveg	1578	1,8%	499	1,0%	24873	4528	13,9%
Tofelts-veg Planskilt	901	1,1%	326	0,6%	10678	1269	3,9%
TEN-T	6243	7,4%	3602	7,0%	3818	5020	15,4%
Øvrig Ev/Rv	7592	9,0%	4403	8,5%	3295	5295	16,2%
Fylkesveg	67950	80,6%	42983	83,0%	1055	16553	50,7%
<b>SUM</b>	<b>84264</b>	<b>100%</b>	<b>51813</b>	<b>100%</b>	<b>1727</b>	<b>32666</b>	<b>100%</b>

## 2.8 Midtdeler og midtrekkverk

For denne kategorien er det definert fire dummyvariabler som følger:

- Verken midtdeler eller midtrekkverk
- Midtdeler, ikke midtrekkverk
- Midtrekkverk, ikke midtdeler
- Midtdeler og midtrekkverk

Dette er gjort for å ta hensyn til eventuelle interaksjonseffekter (altså at virkningen av midtrekkverk kan påvirkes av om det er midtdeler på vegen eller ikke). Deskriptiv statistikk kan sees i Tabell 9.

Tabell 9: Deskriptiv statistikk for midtdeler og midtrekkverk.

	Antall segmenter	Andel av segmenter	Sum lengde (km)	Andel av lengde	Snitt ÅDT (vektet)	Millioner kjt.km per år	Andel av alle kjt.km
Uten midtdeler/rekkverk	81217	96,4%	50857	98,2%	2012	27309	83,6%
Kun midtdeler	827	1,0%	157	0,3%	14491	621	1,9%
Kun midtrekkverk	901	1,1%	266	0,5%	25622	1829	5,6%
Midtdeler m/ rekkverk	1319	1,6%	532	1,0%	20621	2940	9,0%
SUM	84246	100%	51813	100%	1727	32666	100%

## 2.9 Forsterket midtoppmerking

Forsterket midtoppmerking inngår i datasettet med en verdi for bredden av midtoppmerkingen. Det er i tidligere modeller (Høye, 2014) blitt benyttet flere kategorier for smal og bred midtoppmerking med en grenseverdi satt til 1 meter. I (Høye, 2016) gikk man bort fra dette og slo kategoriene sammen. Altså så man kun på om det fantes forsterket midtoppmerking eller ikke. Denne rapporten følger metoden brukt i (Høye, 2016) og benytter en samlet kategori.

Tabell 10: Deskriptiv statistikk for forsterket midtoppmerking.

Forst. midtoppmerking	Antall segmenter	Andel	Sum (km)	Andel av veglengden	Snitt ÅDT (vektet)	Millioner kjt.km per år	Andel av alle kjt.km
Med	3028	3,6%	1585	3,1%	5248	3036	9,3%
Uten	81236	96,4%	50228	96,9%	1616	29630	90,7%
SUM	84264	100%	51813	100%	1727	32666	100%

## 2.10 Automatisk trafikk kontroll (ATK)

For automatisk trafikk kontroll ser vi på to typer; punkt-ATK (PATK) og streknings-ATK(SATK). Det er definert influensstreknings for begge typene i NVDB (streknings der man antar at trafikantene påvirkes av ATK.)

Tabell 11: Deskriptiv statistikk for automatisk trafikk kontroll (ATK).

	Antall segmenter	Andel	Sum (km)	Andel av veglengden	Snitt ÅDT (vektet)	Millioner kjt.km per år	Andel av alle kjt.km
PATK	1224	1,5%	489	0,9%	9326	1666	5,1%
SATK	283	0,3%	150	0,3%	5633	308	0,9%
Ingen ATK	82757	98,2%	51173	98,8%	1643	30692	94,0%
SUM	84264	100%	51813	100%	1727	32666	100%

**Kommentar:** I forrige modell ble SATK delt opp to; en for segmenter med SATK i kun en retning og en der SATK var i begge retninger. Denne informasjonen inngår ikke i det nåværende datagrunnlaget.

## 2.11 Vegbelysning

Vegbelysning er definert i en dummyvariabel som indikerer om vegen er belyst eller ikke. Det er ikke skilt mellom ulike typer belysning, eller om det er satt opp belysning på en eller begge sider av vegen. Deskriptiv statistikk for vegbelysning kan sees i Tabell 12.

Tabell 12: Deskriptiv statistikk for vegbelysning.

	Antall segmenter	Andel	Sum (km)	Andel av veglengden	Snitt ÅDT (vektet)	Millioner kjt.km per år	Andel av alle kjt.km
Med lys	30250	96,4%	40449	78,1%	1052	15533	47,6%
Uten lys	54014	1,0%	11364	21,9%	4131	17133	52,4%
SUM	84246	100%	51813	100%	1727	32666	100%

## 2.12 Fylke

Fylke inngår også i modellen som en dummyvariabel. Fylke vil kunne være med på å forklare generelle forskjeller mellom vegene som ikke blir beskrevet av andre variabler (som f.eks. forskjeller i vær, befolkningstetthet og topografi). Rogaland er satt som referansekategori, da dette er fylket med laveste avvik for antall PSU og antall DR/HS per million kjøretøykilometer fra de gjennomsnittlige tallene.

## 2.13 Oppsummering av prediktorvariablene brukt i regresjonsmodellen

Nedenfor finnes en oppsummering av de ulike prediktorvariablene og måten de inngår i regresjonsmodellen.

- **ÅDT:** Inngår som  $\ln(\text{ÅDT})$
- **Fartsgrense:** Inngår som dummyvariabel. Referansekategori: 80 km/t.
- **Kjørefelt:** Inngår som dummyvariabel. Referansekategori: 2 felt.
  - For PSU og LS 6 felt og over slått sammen.
  - For HS, Drepte og DR/HS blir alt over 5 felt slått sammen
- **X-kryss:** Inngår som dummyvariabel.
- **T-kryss:** Inngår som dummyvariabel.
- **Rundkjøringer:** Inngår som dummyvariabel.
- **Ramper:** Inngår som dummyvariabel.
- **Type veg:** Inngår som dummyvariabel. Referansekategori: Fylkesveg.
- **Midtdeler/midtrekkverk:** Inngår som dummyvariabel.
- **Forsterket midtoppmerking:** Inngår som dummyvariabel.
- **ATK:** Inngår som dummyvariabel.
- **Vegbelysning:** Inngår som dummyvariabel.
- **Fylke:** Inngår som dummyvariabel. Referansekategori: Rogaland.

Det blir benyttet to **offsetvariabler**; Lengde og År. Se 3.1.2 for mer detaljer om offsetvariabler og grunnen til at disse benyttes.

### 2.13.1 Dummyvariabler

Flere av kategoriene inngår som dummyvariabler. En dummyvariabel representerer ulike kategorier som ikke kan rangeres eller måles direkte på en kontinuerlig skala. Dummyvariabler tillater oss å inkludere kategorisk informasjon i statistiske modeller ved å tildele tallverdier til de forskjellige kategoriene. En dummyvariabel kan ta verdien 0 eller 1, der 1 representerer tilstedeværelsen av den spesifikke kategorien.

### 3 Regresjonsmodell for normale ulykkestall

Negativ binomial regresjon er en statistisk teknikk som kan brukes til å modellere antall sjeldne hendelser som kan skje over tid. Negativ binomial regresjon er en spesielt egnet teknikk for å modellere ulykker langs veger, fordi ulykker ofte skjer sjeldent og variansen er stor. En negativ binomial modell tar hensyn til disse egenskapene og gir mer nøyaktige estimater og prediksjoner enn f.eks. en Poisson-modell som ofte benyttes når hendelser skal «telles». I dette kapittelet beskrives det hvordan negativ binomial regresjon kan brukes til å modellere antall ulykker langs veger. Vi vil også diskutere hvordan man kan inkludere en variabel overspredningsparameter i modellen for å ta hensyn til variasjonen i spredningen av dataene.

Negativ binomial regresjon brukes til å modellere telledata, spesielt når disse dataene viser tegn på overspredning. Overspredning oppstår når den observerte variansen i dataene er større enn den som forventes under en bestemt statistisk modell. I en Poisson-regresjonsmodell, forventes variansen av den avhengige variabelen å være lik gjennomsnittet. Det vil si at hvis vi har en responsvariabel  $Y$ , da ville vi forvente at  $\text{Var}(Y) = E(Y)$ , hvor  $\text{Var}()$  og  $E()$  henholdsvis er variansen og forventningen. Imidlertid er dette ofte ikke tilfellet i virkelige scenarier, spesielt når vi har å gjøre med sjeldne hendelser. Antall ulykker langs en veg vil typisk vise en varians som er større enn gjennomsnittet, noe som indikerer overspredning.

#### 3.1 Modellform

I dette avsnittet beskrives matematikken bak negativ binomial regresjon med variabel overspredning. La  $Y$  være den avhengige variabelen (i dette tilfellet antall ulykker). Modellen kan da uttrykkes som:

$$Y \sim NB(\mu, k),$$

der NB representerer den negative binomialfordelingen,  $\mu$  er den forventede verdien av  $Y$  (det vil si det normale antall ulykker), og  $k$  er overspredningsparameteren.

For å binde sammen prediktorvariablene, her definert som  $X$ , med den avhengige variabelen  $Y$ , kan vi bruke en log-lineær link-funksjon. Denne link-funksjonen kan uttrykkes som:

$$\ln(\mu) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p,$$

hvor  $\beta_0$  er en konstanterterm, og  $\beta_1$  til  $\beta_p$  er regresjonskoeffisientene (som beregnes i regresjonsanalysen) for prediktorvariablene  $X_1$  til  $X_p$ .

En fullstendig negativ binomial modell kan da skrives som:

$$Y \sim NB(e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}, k).$$

Med denne modellen kan vi estimere effekten av prediktorvariablene på antall ulykker, samtidig som vi tar hensyn til overspredning i dataene ved hjelp av overspredningsparameteren  $k$ .

En utfordring med å modellere antall trafikkulykker er at det ofte er høy varians sammenlignet med gjennomsnittet av antall ulykker. Dette kalles overspredning, og kan skape problemer for modellene fordi mange modellformer (f.eks. Poisson-modeller) antar at variasjonen er lik forventningen. Negativ binomial modell tar hensyn til overspredning i dataene. Modellen tillater at variasjonen i antall trafikkulykker er større enn forventningen, og dermed kan den gi mer nøyaktige estimater av risikoen for trafikkulykker.



### 3.1.1 Negativ binomial fordeling

#### Forventning:

Forventningen (eller gjennomsnittet) av en negativ binomialfordelt variabel er  $\mu$ . I regresjonskontekst er dette vanligvis en funksjon av de uavhengige variablene, som vi uttrykte med link-funksjonen tidligere. Forventningen til  $Y$ , gitt de uavhengige variablene  $X$ , kan uttrykkes som:

$$E(Y | X) = \mu = e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}$$

#### Varians:

Variansen til en negativ binomialfordelt variabel er gitt ved

$$\mu + \frac{\mu^2}{k}$$

Dette er en funksjon av både gjennomsnittet ( $\mu$ ) og overspredningsparameteren ( $k$ ). Hvis overspredningsparameteren  $k$  er null, reduseres variansen til  $\mu$ , som er det vi forventer under en Poisson-prosess. Men når  $k$  er forskjellig fra null, har vi ekstra varians utover Poisson-prosessen, noe som gjør at modellen kan håndtere overspredning bedre.

Det bemerkes at forventningen og variansen er avhengig av prediktorvariablene gjennom link-funksjonen, og derfor vil endre seg med forskjellige verdier av  $X$ . Dette er kjernen i regresjonsanalysen; vi modellerer hvordan forventningen og variansen av responsvariabelen endrer seg som en funksjon av de uavhengige variablene.

### 3.1.2 Bruk av offsetvariabel

I regresjonsanalyse, og spesielt i Poisson- og negativ binomialregresjon, er en offsetvariabel en type variabel som brukes for å ta hensyn til eksponering. Offsetvariabler er viktig når antall observasjoner ikke er konstant over alle observasjonene i datasettet. F.eks. når det gjelder modellering av trafikkulykker, vil antall ulykker langs en bestemt vegstrekning kunne variere avhengig av lengden på en strekning. En lang vegstrekning vil sannsynligvis ha flere ulykker, alt annet likt. Derfor, når vi modellerer antall ulykker, vil vi ta hensyn til denne eksponeringen.

En typisk offsetvariabel i dette tilfellet kan altså være lengden på en vegstrekning. For å ta hensyn til dette, kan vi inkludere segmentlengde som en offsetvariabel i regresjonsmodellen. Dette gjør at vi kan modellere ulykkesraten, i stedet for det totale antall ulykker.

Hvis vi lar  $Z$  være vår offset variabel, kan den inkluderes i vår negativ binomial regresjonsmodell som følger:

$$\ln(\mu) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \ln(Z)$$

Her er  $\ln(Z)$  lagt til på høyre side av ligningen. Offset variabelen  $Z$  inngår med en koeffisient på 1, og er derfor ikke en estimert parameter i modellen. Det er bare en fast skaleringsfaktor som gjør at vi kan modellere ulykkesraten per km i stedet for det totale antall ulykker. Dette gir oss en mer rettferdig sammenligning mellom vegstrekninger med forskjellige lengder.

I estimeringen av modellene er det benyttet to offsetvariabler; Lengde og År.

### 3.1.3 Variabel overspredning

Å modellere variabel overspredning i en negativ binomial regresjonsmodell kan gjøres ved å tillate at overspredningsparameteren  $k$  varierer i stedet for å være konstant.

En slik modell, kjent som en "generalisert negativ binomial" modell, har en form som ligner på den vanlige negative binomialmodellen, men med  $k$  gjort til en funksjon av de uavhengige variablene. Her er den generelle formen på den fulle modellen:

$$\ln(\mu) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \ln(Z)$$
$$\ln(k) = \gamma_0 + \gamma_1 W_1 + \dots + \gamma_q W_q,$$

hvor  $W_1$  til  $W_q$  er de uavhengige variablene som påvirker overspredning, og  $\gamma_1$  til  $\gamma_q$  er de tilhørende koeffisientene. Den første linjen kjenner vi igjen fra 3.1.2 som beskrivelse av den «ordinære» negative binomiale modellen.

Denne modellen tillater at overspredningen varierer avhengig av de uavhengige variablene  $W_1$  til  $W_q$ , noe som gir mer fleksibilitet og kan tilby bedre tilpasning til dataene. For eksempel, i sammenheng med trafikkulykker, kan overspredning variere basert på egenskaper som trafikkvolum, segmentlengde og tid. Estimering av variabel overspredning for modellene er beskrevet i 3.3.

## 3.2 Estimerte koeffisienter

Tabell 13 viser koeffisientene for alle prediktorvariablene i de fem utviklede modellene. Koeffisientene er estimert i R ved hjelp av pakken «glimmTMB».

**Tabell 13: Koeffisienter i de endelige modellene for personskadeulykker (PSU), lett skadde (LS), hardt skadde (HS), drepte (DR) og drepte eller hardt skadde (DR/HS).**

	PSU		LS		HS		DR		DR/HS	
	Koeff	p	Koeff	p	Koeff	p	Koeff	p	Koeff	P
Ln(ÅDT)	<b>0.874</b>	0.000	<b>0.895</b>	0.000	<b>0.817</b>	0,000	<b>0.808</b>	0.000	<b>0.814</b>	0.000
Fgr, 30 km/t	<b>0.213</b>	0.023	<b>0.056</b>	0.646	<b>-0.214</b>	0.430	<b>-0.638</b>	0.409	<b>-0.237</b>	0.365
Fgr, 40 km/t	<b>0.181</b>	0.000	<b>0.047</b>	0.466	<b>0.037</b>	0.781	<b>-0.637</b>	0.092	<b>-0.027</b>	0.831
Fgr, 50 km/t	<b>0.112</b>	0.002	<b>0.055</b>	0.200	<b>-0.062</b>	0.495	<b>-0.764</b>	0.002	<b>-0.134</b>	0.123
Fgr, 60 km/t	<b>-0.032</b>	0.290	<b>0.008</b>	0.824	<b>-0.296</b>	0.000	<b>-0.453</b>	0.008	<b>-0.314</b>	0.000
Fgr, 70 km/t	<b>-0.101</b>	0.012	<b>-0.040</b>	0.385	<b>-0.219</b>	0.032	<b>-0.263</b>	0.194	<b>-0.230</b>	0.014
Fgr 80 km/t				(ref.)		(ref.)		(ref.)		(ref.)
Fgr 90 km/t	<b>-0.211</b>	0.020	<b>-0.112</b>	0.242	<b>-0.055</b>	0.809	<b>-0.137</b>	0.705	<b>-0.066</b>	0.737
Fgr 100 km/t	<b>-0.685</b>	0.000	<b>-0.632</b>	0.000						
Fgr 110 km/t	<b>-0.829</b>	0.000	<b>-0.765</b>	0.000						
Fgr, 100-110					<b>0.189</b>	0,683	<b>-1.209</b>	0.088	<b>-0.278</b>	0.470
Kjørefelt: 2		(ref.)		(ref.)		(ref.)		(ref.)		(ref.)
Kjørefelt: 3	<b>0.073</b>	0.415	<b>0.014</b>	0.895	<b>-0.613</b>	0.046	<b>-0.215</b>	0.702	<b>-0.500</b>	0.067
Kjørefelt: 4	<b>0.298</b>	0.000	<b>0.272</b>	0.002	<b>0.182</b>	0.415			<b>0,231</b>	0.267
Kjørefelt: 5	<b>0.517</b>	0.001	<b>0.601</b>	0.001						
Kjørefelt: 6+	<b>0.675</b>	0.000	<b>0.612</b>	0.000						
Kjørefelt: 5+					<b>0.292</b>	0.446			<b>0.001</b>	0.998
Kjørefelt 4+							<b>0.106</b>	0.835		
X-kryss	<b>0.416</b>	0.000	<b>0.391</b>	0.000	<b>0.430</b>	0.001	<b>0.572</b>	0.089	<b>0.433</b>	0.001
T-kryss	<b>0.147</b>	0.000	<b>0.190</b>	0.000	<b>0.080</b>	0.168	<b>0.049</b>	0.702	<b>0.073</b>	0.182
Rundkjøringer	<b>0.035</b>	0.455	<b>0.020</b>	0.725	<b>-0.131</b>	0.320	<b>-0.245</b>	0.506	<b>-0.146</b>	0.248
Ramper	<b>0.300</b>	0.097	<b>0.158</b>	0.454	<b>-0.716</b>	0.354	<b>1.418</b>	0.054	<b>0.070</b>	0.895
Motorveg	<b>-0.288</b>	0.031	<b>-0.139</b>	0.330	<b>-0.744</b>	0.092	<b>0.505</b>	0.470	<b>-0.373</b>	0.313
Tofelts planskilt	<b>-0.405</b>	0.001	<b>-0.314</b>	0.012	<b>-0.743</b>	0.031	<b>0.294</b>	0.583	<b>-0.497</b>	0.089
Øvrig TEN-T	<b>-0.035</b>	0.351	<b>0.031</b>	0.467	<b>0.124</b>	0.174	<b>0.469</b>	0.008	<b>0.192</b>	0.021
Øvrig Ev/Rv	<b>-0.012</b>	0.698	<b>0.039</b>	0.294	<b>0.070</b>	0.379	<b>0.360</b>	0.025	<b>0.116</b>	0.114
Fylkesveg		(ref.)		(ref.)		(ref.)		(ref.)		(ref.)
Ikke midtd./midtrekkv.		(ref.)		(ref.)		(ref.)		(ref.)		(ref.)
Kun midtdeler	<b>-0.056</b>	0.581	<b>0.034</b>	0.760	<b>0.103</b>	0.689	<b>-0.383</b>	0.581	<b>0.031</b>	0.899
Kun midtrekkverk	<b>-0.520</b>	0.000	<b>-0.426</b>	0.000	<b>-1.175</b>	0.000	<b>-0.846</b>	0.157	<b>-1.093</b>	0.000
Midtdeler+midtrekkv.	<b>-0.634</b>	0.000	<b>-0.712</b>	0.000	<b>-1.219</b>	0.000	<b>-0.702</b>	0.202	<b>-1.077</b>	0.000
Forst. midtoppm.	<b>-0.422</b>	0.000	<b>-0.422</b>	0.000	<b>-0.433</b>	0.001	<b>-0.069</b>	0.768	<b>-0.355</b>	0.003
SATK	<b>-0.427</b>	0.005	<b>-0.456</b>	0.006	<b>-0.044</b>	0.886	<b>-0.397</b>	0.526	<b>-0.093</b>	0.745
PATK	<b>0.111</b>	0.028	<b>0.085</b>	0.158	<b>0.007</b>	0.961	<b>0.039</b>	0.897	<b>0.007</b>	0.955
Vegbelysning	<b>0.033</b>	0.222	<b>0.029</b>	0.361	<b>0.032</b>	0.636	<b>-0.274</b>	0.070	<b>-0.022</b>	0.730
1-Østfold	<b>0.337</b>	0.000	<b>0.411</b>	0.000	<b>0.147</b>	0.339	<b>-0.380</b>	0.257	<b>0.064</b>	0.652
2-Akershus	<b>0.071</b>	0.170	<b>0.133</b>	0.031	<b>0.218</b>	0.109	<b>0.008</b>	0.977	<b>0.198</b>	0.115
3-Oslo	<b>0.166</b>	0.179	<b>0.130</b>	0.338	<b>0.237</b>	0.531	<b>-0.978</b>	0.255	<b>0.078</b>	0.824
4-Hedmark	<b>-0.137</b>	0.018	<b>-0.126</b>	0.066	<b>0.103</b>	0.464	<b>-0.309</b>	0.301	<b>0.040</b>	0.759
5-Oppland	<b>0.155</b>	0.005	<b>0.189</b>	0.004	<b>0.486</b>	0.000	<b>-0.330</b>	0.262	<b>0.367</b>	0.003
6-Buskerud	<b>-0.108</b>	0.058	<b>-0.059</b>	0.376	<b>0.093</b>	0.514	<b>0.090</b>	0.753	<b>0.065</b>	0.622
7-Vestfold	<b>0.233</b>	0.000	<b>0.286</b>	0.000	<b>0.289</b>	0.051	<b>0.017</b>	0.958	<b>0.247</b>	0.075
8-Telemark	<b>0.324</b>	0.000	<b>0.391</b>	0.000	<b>0.239</b>	0.111	<b>-0.092</b>	0.766	<b>0.195</b>	0.161
9-Aust-Agder	<b>0.161</b>	0.011	<b>0.264</b>	0.000	<b>-0.153</b>	0.392	<b>-0.809</b>	0.058	<b>-0.252</b>	0.131
10-Vest-Agder	<b>0.012</b>	0.847	<b>0.018</b>	0.806	<b>0.131</b>	0.407	<b>-0.210</b>	0.540	<b>0.084</b>	0.569
11-Rogaland		(ref.)		(ref.)		(ref.)		(ref.)		(ref.)
12-Hordaland	<b>0.287</b>	0.000	<b>0.287</b>	0.000	<b>0.335</b>	0.011	<b>-0.431</b>	0.155	<b>0.231</b>	0.062
14-Sogn og Fjør.	<b>0.154</b>	0.019	<b>0.177</b>	0.025	<b>-0.029</b>	0.866	<b>-0.270</b>	0.419	<b>-0.064</b>	0.684
15-Møre og Rom.	<b>0.111</b>	0.042	<b>0.130</b>	0.047	<b>-0.062</b>	0.662	<b>-0.156</b>	0.592	<b>0.036</b>	0.782
50-Trøndelag	<b>-0.107</b>	0.039	<b>-0,103</b>	0,093	<b>-0.190</b>	0.160	<b>-0.196</b>	0.460	<b>-0.180</b>	0.145
18-Nordland	<b>-0,077</b>	0.178	<b>-0.048</b>	0.481	<b>-0.011</b>	0.938	<b>-0.190</b>	0.501	<b>-0.030</b>	0.823
19-Troms	<b>-0,319</b>	0.000	<b>-0.286</b>	0.000	<b>-0.333</b>	0.052	<b>-0.251</b>	0.432	<b>-0.305</b>	0.049
20-Finnmark	<b>-0,175</b>	0.035	<b>-0.152</b>	0.115	<b>-0.236</b>	0.246	<b>-0.101</b>	0.780	<b>-0.173</b>	0.341
Konstantterm	<b>-16,436</b>	0.000	<b>-16,563</b>	0,000	<b>-17.658</b>	0,000	<b>-18.821</b>	0.000	<b>-17.398</b>	0.000

### 3.2.1 Kommentar til modellene

Dersom antall ulykker langs vegen er svært få sammenlignet med antall kjørte kilometer, kan dette føre til at dataene har lav statistisk styrke og begrenset presisjon. Dette er tilfelle i våre modeller, spesielt for modellene med hardt skadde og drepte der det skjer relativt få ulykker. I slike tilfeller kan det være vanskelig å finne signifikante sammenhenger mellom antall ulykker og prediktive variabler, og det kan være vanskelig å lage en pålitelig modell. F.eks. er p-verdiene for modellen Drepte i Tabell 13 svært høye og svært mange av prediktorvariablene er ikke signifikante. For fremtidige modeller bør dette undersøkes videre.

P-verdien er et mål på sannsynligheten for å observere resultatene du har fått (eller mer ekstreme resultater) gitt at nullhypotesen er sann. Nullhypotesen i denne konteksten er som oftest at det ikke er noen effekt eller sammenheng mellom variablene, det vil si at regresjonskoeffisienten er null. Hvis du får en høy p-verdi, betyr det at du ikke har nok bevis for å avvise nullhypotesen. En vanlig terskel for signifikans er en p-verdi mindre enn 0,05.

En høy p-verdi for en koeffisient i en regresjonsmodell indikerer at det ikke er signifikante bevis for at denne koeffisienten er forskjellig fra null, gitt dataene man har. Dette kan bety flere ting:

1. **Variabelen har ingen effekt:** Dette er den mest direkte tolkningen. Variabelen har ingen effekt på den avhengige variabelen, og dens inkludering i modellen er unødvendig.
2. **Det er for mye støy i dataene:** Hvis det er stor varians i dataene dine, kan det være vanskelig å påvise en effekt selv om den eksisterer.
3. **Det er ikke nok data:** Hvis datasettet er lite, kan man mangle den «statistiske kraften» som kreves for å påvise en effekt.
4. **Effekten er svak:** En variabel kan ha en ekte, men svak effekt som ikke er sterk nok til å detekteres som signifikant.

I ulykkesmodellene har flere av koeffisientene høye p-verdier, og det kan være verdt å vurdere om alle de inkluderte variablene er nødvendige eller om modellen kan forenkles. Det kan også være verdt å se om det er mulighet for å samle mer data, eller å undersøke andre mulige kilder til varians i dataene. En løsning kan være å benytte mer data for å øke antall observasjoner av antall drepte langs vegen (f.eks. benytte data fra et lengre tidsrom). Dette kan være spesielt viktig hvis man ønsker å inkludere flere prediktive variabler i modellen, siden et større datasett vil gi mer pålitelige resultater.

En annen løsning kan være å justere modellen til å bruke en modell som tar hensyn til at ulykker er sjeldne hendelser. Såkalte «sjeldenhetsmodeller» (rare event models) er designet for å håndtere datasett der det er svært få tilfeller av hendelser av interesse, som i dette tilfellet antall drepte langs vegen. Disse modellene tar hensyn til at sannsynligheten for å observere hendelser er svært liten, og kan derfor gi mer nøyaktige estimater selv med begrensede data. Såkalte zero-inflated modeller er en slik type modell og kan vurderes for dette tilfellet. Alternative metoder for å estimere risikoen for ulykker, som for eksempel maskinlæringsteknikker kan også benyttes. Disse metodene kan være bedre egnet for å håndtere datasett med få observasjoner, og kan dermed gi mer nøyaktige resultater.

### 3.3 Estimert variabel overspredning

Vanligvis brukes en fast overspredningsparameter i negativ binomial regresjon for å beskrive spredningen av responsvariabelen, men det kan være tilfeller der spredningen av responsvariabelen varierer avhengig av verdien av en eller flere forklaringsvariabler. Dette kan føre til at en fast overspredningsparameter gir en dårlig tilpasning til dataene og reduserer presisjonen til estimatene.

Ved å bruke en variabel overspredningsparameter kan man fange opp variasjonen i spredningen av responsvariabelen og dermed produsere mer nøyaktige estimater og prediksjoner. Dette kan være spesielt nyttig i situasjoner der responsvariabelen har store variasjoner eller uventede mønstre i spredningen. Det er vist av (Geedipally, Lord & Park, 2009) at ulykkestall langs veg kan modelleres bedre ved å ta hensyn til at overspredningen varierer som en funksjon av strekningslengde, antall år og trafikkmengde.

Den variable overspredningsparameteren er estimert i R ved bruk av pakken «glmmTMB». Vi lar her overspredningsparameteren variere med  $\ln(\text{Lengde})$ ,  $\ln(\text{År})$  og med  $\ln(\text{ÅDT})$ . Resultatene fra estimeringen er oppsummert i Tabell 14, og overspredningsparameteren  $k$ , kan nå beregnes ved formelen:

$$k = e^{\text{Konstantterm} + (\ln(\text{Lengde}) * \text{Koeff}_1) + (\ln(\text{År}) * \text{Koeff}_2) + (\ln(\text{ÅDT}) * \text{Koeff}_3)}$$

**Tabell 14: Koeffisienter for beregning av overspredningsparameteren i de endelige modellene for personskadeulykker (PSU), lett skadde (LS), hardt skadde (HS), drepte (DR) og drepte eller hardt skadde (DR/HS).**

	PSU		LS		HS		Drepte		DR/HS	
	Koeff.	p	Koeff.	p	Koeff.	p	Koeff.	p	Koeff.	P
$\ln(\text{Lengde})$	0.653	0.000	0.702	0.000	0.476	0.000	1.509	0.000	0.671	0.000
$\ln(\text{År})$	0.086	0.873	1.060	0.000	1.338	0.000	1.575	0.345	1.275	0.000
$\ln(\text{ÅDT})$	0.375	0.000	0.708	0.000	0.508	0.000	0.848	0.000	0.595	0.000
Konstantterm	-6.650	0.000	-12.863	0.000	-11.024	0.000	-21.672	0.000	-12.891	0.000

### 3.4 Hvor gode er modellene?

Tabell 15 viser ulike goodness-of-fit (GOF) statistikker som beskriver hvor gode modellene predikerer registrerte ulykkestall. Statistikkene er beskrevet under:

- Pseudo-R<sup>2</sup>:** McFadden's pseudo R-kvadrat er en populær statistikk som brukes til å vurdere hvor godt en modell, som en logistisk eller negativ binomial regresjonsmodell, passer til dataene. Den kan sees på som en analog til den klassiske R-kvadrat-verdien som brukes i lineær regresjon, men dens tolkning og beregningsmetode er forskjellig. McFadden's pseudo R-kvadrat beregnes ved å ta forholdet mellom log-likelihood av den estimerte modellen til log-likelihood av en null-modell som kun inkluderer en konstant. Dette forholdet trekkes deretter fra 1 for å gi et mål på modellens forbedring sammenlignet med null-modellen. Verdier for McFadden's pseudo R-kvadrat ligger vanligvis mellom 0 og 1, hvor høyere verdier indikerer at den valgte modellen gir en bedre tilpasning til dataene enn null-modellen.
- Mean Squared Prediction Error (MSPE):** MSPE er et mål på nøyaktigheten til en statistisk modell. Det er gjennomsnittet av kvadratene av forskjellene (feilene) mellom modellens predikerte verdier og de observerte verdiene i datasettet. Lavere verdier av MSPE indikerer at modellens prediksjoner passer bedre til de observerte verdiene. En MSPE-verdi på null indikerer en perfekt modell, men i praksis er dette sjelden eller aldri oppnåelig. MSPE er sensitiv for store feil, siden disse gir større bidrag til summen på grunn av kvadreringen. Derfor kan denne metrikken være spesielt nyttig når det er viktig å unngå store avvik mellom de predikerte og observerte verdiene.

MSPE har ingen klar skala eller et referansepunkt for hva som er en "god" eller "dårlig" verdi. MSPE-verdien avhenger av skalaen til den avhengige variabelen og vil kunne variere betydelig mellom de forskjellige modellene.

**Tabell 15: Goodness-of-fit for de endelige modellene og avvik mellom normale og registrerte ulykkestall.**

	PSU	LS	HS	DR	DR/HS
<b>De aktuelle modellene</b>					
Pseudo R <sup>2</sup>	0,185	0,159	0,120	0,114	0,121
MSPE	0,179	0,304	0,033	0,006	0,041
<b>De forrige modellene (Høye, 2016)</b>					
Pseudo R <sup>2</sup>	0,190	0,140	0,072	0,062	0,070
MSPE	0,338	0,692	0,046	0,011	0,063
<b>Totale ulykkestall</b>					
Registrerte (sum)	12464	13733	2118	417	2535
Registrerte (maks)	11	16	8	5	8
Registrerte (varians)	0,233	0,375	0,034	0,006	0,043
Normale (sum)	12509	13873	2128	417	2542
Normale (maks)	8,40	10,86	1,41	0,49	1,23
Normale (varians)	0,066	0,088	0,001	0,000	0,002
<b>Avvik (antall)</b>	45	140	10	0	7
<b>Avvik (prosent)</b>	0,4%	1,0%	0,5%	0,0%	0,3%

Sammenlignet med modellene i (Høye, 2016) viser tabellen at pseudo-R<sup>2</sup>-verdiene er litt høyere for alle modellene, bortsett fra PSU-modellen, hvor det er en liten nedgang sammenlignet med de tidligere modellene. En høyere pseudo-R<sup>2</sup>-verdi indikerer bedre tilpasning av modellen, noe som betyr at modellene våre er på nivå med de tidligere utviklede modellene. Vi kan også se en svært liten forbedring i MSPE-verdiene for alle modellene sammenlignet med 2016.

Vi har også gjort prediksjoner for normale antall ulykker ved hjelp av hele datasettet. Resultatene i tabellen viser at summen av ulykker for hele perioden stemmer godt overens mellom de registrerte ulykkene og de normale (predikerte) ulykkene. Dette betyr at modellene fungerer godt for treningsdataene.

Det er imidlertid viktig å være oppmerksom på muligheten for overtilpasning. Overtilpasning oppstår når modellen er for kompleks og tilpasser seg treningsdataene for mye, noe som kan føre til dårlige resultater når modellen brukes på nye, uprøvde data. For å undersøke om overtilpasning forekommer, kan modellene evalueres ved hjelp av separate testsett eller kryssvalideringsteknikker for å vurdere deres ytelse på ukjente data.

## 4 Referanser

Geedipally, S. R., Lord, D., & Park, B. J. (2009). Analyzing different parameterizations of the varying dispersion parameter as a function of segment length. *Transportation Research Record*, 2103, 108-118.

Høye, A. (2014). Utvikling av ulykkesmodeller for ulykker på riks- og fylkesvegnettet i Norge. TØI-rapport 1323/2014. Oslo: Transportøkonomisk institutt.

Høye, A. (2016). Utvikling av ulykkesmodeller for ulykker på riks- og fylkesvegnettet i Norge. TØI-rapport 1522/2016. Oslo: Transportøkonomisk institutt.

Ragnøy, A. & Elvik, R. (2003). Trafikksikkerhetsanalyse av stamvegnettet i Norge. TØI-rapport 649/2003. Oslo: Transportøkonomisk institutt.

Ragnøy, A. & Christensen, P. & Elvik, R. (2002). Skadegradstetthet - SGT. Et nytt mål på hvor farlig en vegstrekning er. TØI-rapport 618/2002. Oslo: Transportøkonomisk institutt.