Populärvetenskaplig presentation

När man genomför en statistisk studie är ett vanligt tillvägagångssätt att man börjar med att välja ut ett antal studieobjekt. Studieobjekten hämtas ur den grupp av individer, populationen, som man vill kunna dra slutsatser om. Därefter följer man dessa studieobjekt under en längre tidsperiod för att studera hur de faktorer man är intresserad av utvecklas över tid. Denna typ av studiemetod kallas *prospektiv*, vilket betyder framåtblickande.

Om den typ av händelse som man vill studera är sällsynt, som exempelvis en ovanlig sjukdom, så är denna studiemetod ofta i praktiken omöjlig att genomföra. Man skulle behöva följa ett mycket stort antal individer. I annat fall kan man inte vara säker på att ett tillräckligt stort antal fall av händelsen hinner uppstå under tiden för studiens genomförande. Det stora antalet studieobjekt skulle innebära stora ekonomiska resurser, och ett mycket omfattande arbete med att samla in och analysera data. En mer effektiv metod kan i en sådan situation vara att närma sig problemet från andra hållet:

Det vill säga man väljer ut fall av händelsen ifråga, exempelvis patienter drabbade av en sällsynt sjukdom. Dessa fall studeras sedan, efter att sjukdomen redan har inträffat. Studien syftar till att försöka analysera vilka faktorer som har betydelse för risken att drabbas av sjukdomen i fråga. Denna typ av studiemetod kallas *retrospektiv*, vilket betyder tillbakablickande. Retrospektiva studier kräver inte alls lika stora ekonomiska resurser eller omfattande arbete, eftersom de inte innebär att man behöver följa ett lika stort antal individer, och inte över tid. Det finns dock flera svårigheter med denna typ av studie. Ett problem är att man inte följer studieobjekten aktivt innan sjukdomen i fråga har inträffat. På grund av det problemet blir det svårare att kontrollera och ta hänsyn till alla faktorer som kan ha påverkat uppkomsten av sjukdomen.

En vanligt förekommande typ av retrospektiv studie är så kallade *fall-kontrollstudier*. Det är analysen av denna typ av studier som vår egen studie berör. Namnet fall-kontrollstudie kommer av att man studerar *fall* av exempelvis patienter drabbade av en viss sjukdom, mot *kontroller*. Dessa kontroller är individer som inte är drabbade av sjukdomen. Uppdelningen i fall och kontroller gör man för att på ett bra sätt kunna analysera orsaksfaktorerna. Man kan jämföra detta med en vanlig prospektiv studie, där kontrollerna istället består av individer som inte är påverkade av de orsaksfaktorer vars betydelse man vill undersöka.

Ett exempel på en situation där en fall-kontrollstudie kan vara lämplig, är om man vill undersöka sambandet mellan rökning och lungcancer. Detta är också något som också har gjorts i flera välkända studier. Man går tillväga så att man väljer ett visst antal fall av patienter som är drabbade av lungcancer. Tillsammans med varje fall grupperas en eller flera kontroller av individer som inte är drabbade. Därefter analyseras data från dessa fall och kontroller i syfte att utforska om, och i så fall i vilken grad, rökning påverkar risken för att drabbas av lungcancer.

Det finns två olika metoder att välja dessa kontroller. Ett sätt är att man väljer ut dem slumpmässigt. Ett annat sätt att välja ut kontrollerna är att man matchar vissa av dess egenskaper mot egenskaper hos fallen de ska grupperas ihop med. Exempel på sådan matchning kan vara att man utser kontroller av samma kön eller ålder som sitt motsvarande fall.

Dessa två sätt att välja kontroller på har sina olika fördelar och nackdelar. Antag att man är intresserad av hur rökning påverkar risken att drabbas av lungcancer, som i exemplet ovan. Då är det en fördel om det går att, så långt som möjligt, förvissa sig om att det inte är något annat som är den egentliga orsaken bakom. Personens ålder skulle kunna vara en faktor som påverkar risken att drabbas. Om man då väljer kontroller av samma ålder som fallen,

så blir åldersfaktorn i sig inte något som påverkar fallet och kontrollen på olika sätt. Detta kan samtidigt vara en nackdel med metoden. Skulle man även vara intresserad av hur åldern eventuellt påverkar, så går det inte att utläsa något om detta när man har matchat på just ålder. Vid matchning av många egenskaper samtidigt, till exempel kön, ålder, utbildningsnivå och boendeort, kan det också bli svårare att hitta lämpliga kontroller.

Om kontroller väljs slumpmässigt har man istället möjlighet att ta med alla dessa faktorer i sin analys. Det gör det också lättare att undersöka hur en kombination av faktorer, som exempelvis kön och rökning, eventuellt påverkar risken för att drabbas av lungcancer. Det skulle kunna vara så att en samverkan mellan faktorerna rökning och kön påverkar risken på ett annat sätt, än bara summan av effekterna av de olika faktorerna var för sig. En större mängd faktorer att ta hänsyn till kan dock leda till att det blir svårare att urskilja vilka av dem som är av betydelse för sjukdomsrisken.

Man vill gärna kunna minimera nackdelarna med de olika sätten att välja kontroller, men samtidigt dra nytta av deras fördelar. Detta kan gå att uppnå genom att använda sig av en kombination av de båda metoderna. Vår studie går i huvudsak ut på att finna en lämplig metod för hur en sådan sammanvägning ska gå till, och undersöka hur mycket säkrare resultat man kan uppnå på detta sätt.

Den typ av händelse vi utgår ifrån i vår studie är trafiksituationer. Vi använder oss inte av verklig data, utan skapar data genom datorsimuleringar. Den tänkta situationen är att vi har en bil som färdas på en landsväg, med en annan bil framför sig. Plötsligt bromsar bilen framför kraftigt, och händelsen vi är intresserade av är om en kollision inträffar. Denna händelse är ovanlig i förhållande till händelsen att kollision inte uppstår. Därför är en fall-kontrollstudie en lämplig modell för analys av denna typ av händelser. Utfallet som vi simulerar är om en kollision uppstår eller inte. Den ena faktorn som påverkar risken för kollision i vår modell är färdhastigheten för bilen bakom. Den andra faktorn är om föraren i den bakre bilen har blicken riktad framåt mot trafiken eller inte. Datasimuleringarna av dessa faktorer och utfallet gör vi med hjälp av programspråket `R`.

Utifrån vår data genomför vi sedan en delstudie med slumpmässigt valda kontroller och en delstudie med matchade kontroller. Fallen är gemensamma för de två delstudierna. Faktorn som vi matchar är hastigheten. Vi väljer alltså kontroller där bilen har samma hastighet som i motsvarande fall där kollision har uppstått. Kontrollerna består då i det här fallet av situationer där bilarna inte kolliderar. Därefter väger vi samman resultaten med hjälp av en matematisk metod som vi har tagit fram, en tillämpning av något som kallas för *minsta kvadratmetoden*.

Vår studie visar att man ofta kan uppnå en påtaglig förbättring av resultatet, genom att väga samman resultaten från studier med slumpmässigt valda och med matchade kontroller. Förbättringen är i jämförelse med de resultat man kan uppnå genom de båda metoderna var för sig.

CHALMERS | GÖTEBORGS UNIVERSITET

Sammanvägning av parameterskattningar i fall-kontrollstudier med matchade och slumpmässigt valda kontroller

Examensarbete för kandidatexamen i matematik vid Göteborgs universitet

Mikael Boman
Ove Holm
Leo Jansson
Daniel Odin
Erik H. Rezazadeh

Institutionen för matematiska vetenskaper
Chalmers tekniska högskola
Göteborgs universitet
Göteborg 2017

Sammanvägning av parameterskattningar i fall-kontrollstudier med matchade och slumpmässigt valda kontroller

Examensarbete för kandidatexamen i matematisk statistik inom matematikprogrammet vid Göteborgs universitet

Mikael Boman Ove Holm Leo Jansson
Daniel Odin Erik H. Rezazadeh

Handledare: Prof. Olle Nerman, Henrik Imberg
Examinator: Marina Axelson-Fisk, Maria Roginskaya

Institutionen för matematiska vetenskaper
Chalmers tekniska högskola
Göteborgs universitet
Göteborg 2017

Sammanfattning

I statistiska studier är det vanligt att man väljer ut ett antal studieobjekt ur en population, för att sedan följa dessa och studera hur en faktor av intresse utvecklas över tid. I synnerhet inom epidemiologiska studier är detta ett typiskt tillvägagångssätt. Om händelsen av intresse i studien är sällsynt är metoden dock ofta ogenomförbar, och det är lämpligare att genomföra en *retrospektiv* studie, en studie av fall som redan har inträffat.

En sådan typ av studie är *fall-kontrollstudier*. Där väljer man ut *fall* av händelsen man är intresserad av ur en population, och *kontroller*, "icke-fall", ur samma population, för att kunna göra en statistisk analys av orsakande faktorer. Kontrollerna kan man välja slumpmässigt, eller så kan man välja kontroller där man har matchat vissa egenskaper så att de liknar fallen. Exempelvis kan man välja kontroller av samma kön eller ålder som fallen.

Dessa två olika sätt att välja kontroller har olika för- och nackdelar, och det kan vara av intresse att försöka kombinera fördelarna med bägge metoderna. Detta är huvudsyftet med denna studie, att undersöka hur mycket det går att förbättra precisionen i resultaten i en fall-kontrollstudie, genom att på ett optimalt sätt väga samman resultat från två delstudier med matchade respektive slumpmässigt valda kontroller, men med gemensamma fall.

I vår studie använder vi oss av simulerade data. För simuleringar och beräkningar använder vi oss av olika paket inom programspråket `R`. De statistiska metoder vi huvudsakligen använder oss av är *logistisk regression*, *betingad* logistisk regression, *bootstrap* och *minsta kvadratmetoden*.

Vår analys visar att det går att åstadkomma en väsentlig förbättring av precisionen i resultaten genom att sammanväga resultaten från studier med matchade respektive slumpmässigt valda kontroller.

Nyckelord: Retrospektiv studie, Fall-kontrollstudie, Matchade kontroller, Slumpmässigt valda kontroller, Logistisk regression, Betingad logistisk regression, Maximum likelihoodskattning, Bootstrap, Generaliserade linjära modeller

Abstract

In statistical studies, it is a common practice to have a number of study objects that are sampled from a population. Then they are followed for the purpose of studying how a factor of interest develops over time. For epidemiological studies in particular, this is a typical approach. However, if the event of interest in the study occurs rarely, this approach is often impractical. In these cases, it might be more appropriate to carry out a *retrospective* study, a study of events that have already occurred.

One type of retrospective study that is frequently used, is the *case-control study*. The basic concept of a case-control study is that a number of *cases* of the event of interest is selected, from a population. Then *controls*, "non-cases", are sampled from the same population, and a statistical analysis is performed on this group of cases and controls. The controls can be sampled randomly, or they could be selected by matching certain factors against those of the cases, for example *gender* or *age*.

These two different methods for selecting controls come with various advantages and disadvantages. Therefore, it would be beneficial to be able to combine results produced using each of the methods, in a way that preserves their advantages, but minimizes their drawbacks. This is the main purpose of this study. We want to investigate to what degree it is possible to improve the accuracy of the results from a case-control study, by using a combination of the results from two substudies, using matched and randomly sampled controls, but with the same cases.

The data used in our study are produced from computer simulations. For simulations and calculations we use different packages within the `R` programming language. The statistical methods we mainly use are *logistic regression*, *conditional* logistic regression, *bootstrapping* and the *least squares method*.

We conclude from our analysis that it is possible to achieve an essential improvement in accuracy of the results, by comparing results from studies with matched and randomly sampled controls.

Keywords: Retrospective study, Case-control study, Matched controls, Randomly selected controls, Logistic regression, Conditional logistic regression, Maximum-likelihood estimation, Bootstrap, Generalized linear models

Innehåll

Förord

Denna rapport återger ett examensarbete i matematisk statistik. Alla fem av arbetets delaktiga går matematikprogrammet på Göteborgs universitet. Individuella loggböcker har förts över gruppmedlemmarnas prestationer. Det har även skrivits en gemensam dagbok där gruppens arbete och de individuella insatserna sammanfattats vecka för vecka.

Projektet delades från början upp i två inläsningsdelar, där Ove och Daniel fokuserade på teori bakom multivariata normalfördelningar och Leo, Mikael och Erik fokuserade på logistisk regression och hur man bootstrappar kovarianser. Detta utvecklades senare till att alla läste på om fall-kontrollstudier och simuleringar användes för att få en förståelse om detta.

Alla gruppmedlemmar har varit med och sammanställt denna rapport. Ove har haft huvudansvaret för *Inledning* och tillsammans med Erik även *Logistisk regression*. Leo har haft huvudansvaret för *Fall-kontrollstudier* och tillsammans med Daniel även *Utvärdering genom simuleringsexempel* och *Resultat*. Mikael har haft huvudansvaret för *Betingad logistisk regression*. Daniel har haft huvudansvaret för *Översikt om linjära regressionsmodeller* och *Metod för sammanvägning av skattade parametrar* och tillsammans med Erik även *Bootstrap*. Diskussionen har alla varit delaktiga i.

Utöver detta har Ove och Erik haft huvudansvaret för populärvetenskapliga presentationen och sammanfattningen. Daniel har haft huvudansvaret för programmering och sammanställande av resultat med hjälp från Leo och Mikael. Mikael har haft huvudansvar för stil och struktur i rapporten, framförallt med tekniska detaljer i LaTeX-implementeringen. Ove har haft huvudansvar för korrekturläsning och språket i hela rapporten.

Vi vill tacka vår handledare Olle Nerman som tagit fram detta projekt och hjälpt oss med teorin. Vi vill också tacka vår handledare Henrik Imberg som bland annat hjälpt oss med implementering i `R` och återgett teori på en grundläggande nivå. Tack till Olle och Henrik som tagit sig tid att träffa oss varje vecka. Vi vill även tacka avdelningen för fackspråk som hjälpt oss med projektplan och rapportskrivande.

1 Inledning

I en välkänd studie av Richard Doll och Austin B. Hill från 1952 undersöktes sambandet mellan rökning och lungcancer [1]. Man sökte upp 1465 *fall* av patienter som var drabbade av lungcancer, på olika sjukhus i London-området. Till varje enskilt fall parades en *kontroll*, en patient som inte var drabbad av lungcancer men av någon annan cancersjukdom. Därefter studerade man de två grupperna för att se om man kunde finna ett samband mellan patienternas tidigare tobaksvanor och risken att drabbas av lungcancer. Studien visade på ett statistiskt signifikant samband mellan tobaksrökning och lungcancer, och dess genomslag blev mycket stort. Detta var första gången man på ett mycket tydligt och säkert sätt kunde visa på ett sådant samband.

Denna typ av studier, där man från början väljer ut fall och kontroller, kallas följaktligen för *fall-kontrollstudier*. Ofta är det just studier av epidemiologisk karaktär som lämpar sig för denna metod, där man vill undersöka olika faktorers samband med en sällsynt händelse eller sjukdom. En fall-kontrollstudie är en *retrospektiv* studie, vilket innebär att man exempelvis studerar patienter som redan har utvecklat en sjukdom, till skillnad från en *prospektiv* studie, där man följer ett antal studieobjekt och väntar på att fall skall utvecklas. Att den aktuella händelsen är ovanlig är just det som gör att detta retrospektiva tillvägagångssätt ter sig mer lämpligt. Detta då man annars hade behövt följa en orimligt stor grupp och dessutom ändå möjligen fått ett för litet antal fall att studera.

I fall-kontrollstudier samlar man alltså flera fall av den eftersökta händelsen från en population, och ur samma population samlar man en kontrollgrupp av "icke-fall". Antalet kontroller väljs vanligen antingen till lika många som antalet fall, eller till en multipel av detta antal. Hur man väljer en passande kontrollgrupp är av stor betydelse och i huvudsak finns det två sätt att gå tillväga. Ett sätt är att välja kontroller helt slumpmässigt, det andra är att matcha vissa variabler mellan fall och kontroll, som till exempel kön och ålder. Med hjälp av *logistisk regression*, eller i det andra fallet *betingad* logistisk regression, kan man sedan skatta parametervärden för effekterna av de förklarande variablerna.

En fördel med att använda sig av matchade kontroller är att man kan få en högre precision i denna parameterskattning, för de variabler som man inte har matchat på. En annan fördel är att man kan uppnå en balans i urvalet, med avseende på viktiga riskfaktorer som ej är av primärt intresse, men viktiga att ta hänsyn till för att kunna göra en korrekt jämförelse av fall och kontroller.

Nackdelen är att man tappar information om de matchade variablerna, gentemot om man använder sig av slumpmässigt valda kontroller. Orsaken till detta är att man för varje par av fall och kontroll eliminerar de matchade variablernas betydelse för utfallet.

För att kunna dra nytta av fördelarna med matchning, men samtidigt minimera nackdelarna, kan det vara av intresse att använda sig av både matchade och slumpmässigt valda kontroller. Genom en kombinerad analys, där man på ett optimalt sätt väger samman parameterskattningarna från en delstudie med matchade och en med slumpmässigt valda kontroller, med gemensamma fall, kan man uppnå ett resultat med högre precision än genom de två metoderna var för sig. Det är just en sådan analys vi ska titta närmare på i denna studie. Ett översiktligt och förenklat schema över designen för en sådan studie kan ses i Figur 1.3.1.

Vi har under arbetets gång funnit andra studier där man väger samman resultat från separata studier med matchade respektive slumpmässigt valda kontroller [2], och där man väger samman resultat från två delstudier på ett liknande sätt som vårt [3]. Dessa var inte kända för oss från början av arbetet med att utveckla vår metod, och vi undersöker också i en vidare utsträckning hur faktorer som stickprovsstorlek och korrelation mellan variabler påverkar nyttan av sammanvägningen av parameterskattningarna.

1.1 Syfte

Syftet med denna uppsats är att undersöka i vilken grad parameterskattningarna i multivariata logistiska regressionsmodellerade fall-kontrollstudier kan förbättras, genom sammanvägning av skattningar från två delstudier baserade på samma fall, men med matchade respektive slumpmässigt utvalda kontroller.

1.2 Metod

Som grund för vår analys har vi ett antal statistiska metoder. För vår parameterskattning i fallet med slumpmässigt valda kontroller använder vi oss av *logistisk regression*, och i fallet med matchade kontroller använder vi oss av *betingad logistisk regression*. För att skatta kovariansstrukturen mellan parameterskattningarna använder vi oss av *bootstrap*. Sammanvägningen av parameterskattningarna från de två metoderna sker med hjälp av en variant av *minsta kvadrat-metoden*. Dessa metoder förklaras närmare i kapitel 2.

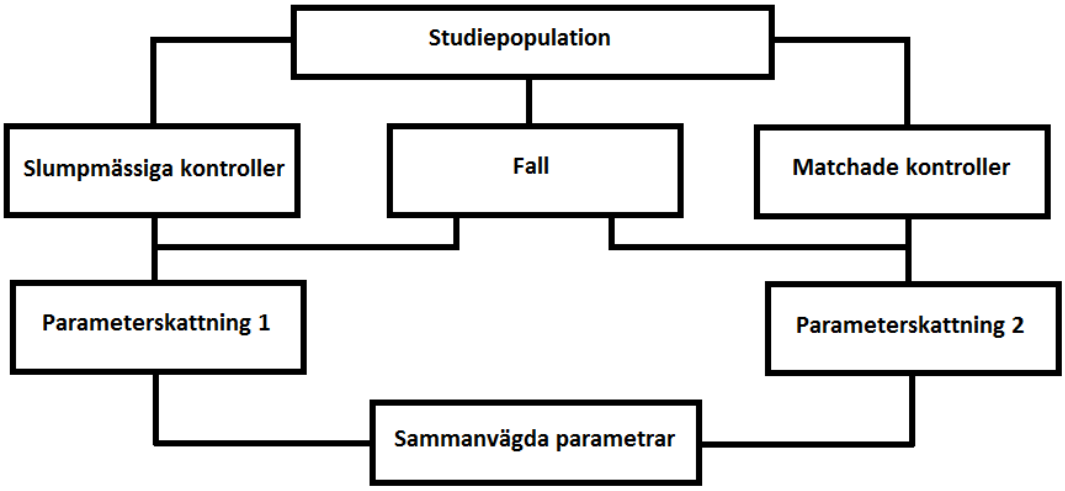
Vårt arbete bygger dels på teori och dels på datorsimuleringar. Samtliga simuleringar och beräkningar görs i programspråket `R` [4]. Vi analyserar och gör beräkningar utifrån vår data med hjälp av en kombination av inbyggda funktioner, och metoder och algoritmer som vi utvecklar.

Våra simuleringar utgår från en tänkt trafiksituation, där händelsen av intresse, *responsvariabeln*, är om en olycka inträffar eller inte. I vår modell låter vi sannolikheten för att en olycka ska inträffa avgöras av två faktorer eller så kallade *förklarande* variabler. Den ena faktorn är om föraren har blicken på vägen eller inte, och den andra är det egna fordonets hastighet då risksituationen uppstår.

1.3 Avgränsningar

Utöver att använda oss av simulerad data hade vi ursprungligen en ambition att även analysera data från verkliga trafiksituationer. Datan består av observationer av färdhastighet och eventuellt borttittande, från situationer då olyckor har skett, och motsvarande situationer då olyckor ej har inträffat. På grund av tidsbrist, och vissa praktiska svårigheter med att få tillgång till och analysera materialet, fick vi dock välja att avstå från detta.

Vi har av praktiska skäl valt att begränsa vår analys till att omfatta två orsakande faktorer och hur de påverkar utfallsvariabeln, olycka eller ej. Detta är dock enbart en begränsning som vi har valt för våra simuleringar, teorin och metoderna vi använder oss av är allmänt formulerade och går att applicera på studier med fler orsaksfaktorer, och matchning på flera av dessa.



Figur 1.3.1: Översiktligt schema över designen av en fall-kontrollstudie, med sammanvägning av parameterskattningar från delstudier med matchade respektive slumpmässigt valda kontroller.

2 Teoretisk bakgrund

I detta kapitel beskrivs kortfattat de bakomliggande teoriena till de olika metoderna som används i detta projekt. Avsnitt 2.1 ger en grundläggande insikt i teori och utförande av en fall-kontrollstudie. En beskrivning av matchade och slumpmässigt valda kontroller ges och deras för- och nackdelar diskuteras. I de fall där responsvariabeln är binär så är *logistisk regression* en lämplig matematisk modell för att skatta parametervärden för prediktorvariablerna, vilket introduceras i avsnitt 2.2. Om man dessutom vill använda sig av matchning på vissa variabler inom stickprovet så kan man använda sig utav *betingad logistisk regression*, avsnitt 2.3. För sammanvägning av parametrar från två studier använder vi oss av en *linjär regressionsmodell*. I avsnitt 2.4 ges en översikt över linjära regressionsmodeller, samt hur man skattar dess koefficienter med *minsta kvadrat-metoden*. För skattning av fördelning och varians av statistiska parametrar är *bootstrap* en användbar metod, se avsnitt 2.5. Bootstrap används i detta projekt för att skatta en kovariansmatris, som är nödvändig när man sammanväger parameterskattningar.

För att undvika alltför omfattande och tekniskt avancerat innehåll ger vi ibland enbart litteraturhänvisningar för mer djuplodande studier inom området.

2.1 Fall-kontrollstudier

En *fall-kontrollstudie* är en typ av retrospektiv studie, där bakomliggande faktorer kopplade till en förhållandevis sällsynt händelse undersöks, till exempel en ovanlig sjukdom. Att den är *retrospektiv* innebär att data samlas in från redan dokumenterade fall, för att man sedan ska kunna studera en eller flera potentiellt orsakande faktorer. Motsatsen till en retrospektiv studie är en *prospektiv* studie, där man följer ett antal objekt för att under studiens gång analysera uppkomst av fall. När det är praktiskt genomförbart så är prospektiva studier att föredra, då sådana i lägre utsträckning genererar resultat med systematiska fel, och ger säkrare skattningar av vad som verkligen orsakar uppkomst av fall. I situationer där fallen är sällsynta är det dock ofta praktiskt och ekonomiskt ohållbart att följa tillräckligt många individer för att möjliggöra en prospektiv studiedesign. Huvudsaklig referens till detta kapitel är [5].

2.1.1 Introduktion till fall-kontrollstudier

I en fall-kontrollstudie modelleras det aktuella fallet som en slumpvariabel Y vilken antar värdena 1, för fall och 0, för kontroll. De orsakande faktorerna modelleras som slumpvariabler X, och var och en av dessa kan vara antingen kategoriska eller kontinuerliga. Vi söker alltså relationen mellan Y och X. Ett sätt att se på denna relation är risken att vara ett fall, givet någon uppsättning förklarande variabler $X = \mathbf{x}_i$, det vill säga $P(Y = 1|X = \mathbf{x}_i)$ och jämföra den mot $P(Y = 1|X = \mathbf{x}_j)$ som är risken att vara ett fall givet en annan uppsättning förklarande variabler $X = \mathbf{x}_j$. En kvot mellan dessa bildar vad som kallas en relativ risk. Den relativa risken går dock inte omedelbart att utläsa ur en fall-kontrollstudie, då man redan i utformningen av studien har bestämt förhållandet mellan antalet fall och kontroller. Vi behöver då jämföra de två ovan nämnda sannolikheterna på ett annat sätt och vi kommer istället att studera sambandet genom ett *odds*. Oddset anger den relativa sannolikheten mellan att en händelse inträffar och att den inte inträffar. Låt sannolikheten för en godtycklig händelse A betecknas $P(A)$, då kan vi definiera ett odds som

$$\frac{P(A)}{P(A^{\mathsf{c}})}$$

där A^{c} betecknar komplementet till A, det vill säga att A inte inträffar. Med denna definition kan vi också uttrycka ett odds för en händelse A givet en händelse B genom

$$\frac{P(A|B)}{P(A^{\mathsf{c}}|B)}.$$

Kvoten mellan oddset för A givet en händelse B och oddset för A givet B^{c} kallas för en *oddskvot* och kan då skrivas som

$$\frac{P(A|B)/P(A^{\mathsf{c}}|B)}{P(A|B^{\mathsf{c}})/P(A^{\mathsf{c}}|B^{\mathsf{c}})}$$

som är ett mått på association mellan de två händelserna A och B. Vi ska längre fram i detta avsnitt se varför denna oddskvot är möjlig att skatta även i en retrospektiv studie.

Låt oss betrakta det mest grundläggande scenariot, då vi har en studie med slumpmässiga kontroller, och endast en förklarande variabel X som antar värdena 0 eller 1. Då kan data summeras som i Tabell 2.1.1, där n_0 och n_1 är totala antalet kontroller respektive fall, och r_0 och r_1 är antalet kontroller respektive fall med $X = 1$.

Tabell 2.1.1: Korstabell för grundläggande fall-kontrollstudie med slumpmässiga kontroller.

	Kontroller	Fall	Total
$X = 0$	$n_0 - r_0$	$n_1 - r_1$	$n - r.$
$X = 1$	r_0	r_1	$r.$
Total	n_0	n_1	n

Det empiriska oddset för att $X = 1$ bland fallen är $r_1/(n_1 - r_1)$ och oddset för att $X = 1$ bland kontrollerna är $r_0/(n_0 - r_0)$. Vi får då oddskvoten

$$\frac{r_1/(n_1 - r_1)}{r_0/(n_0 - r_0)} \, . \tag{2.1.1}$$

Här innebär kvoten 1 ingen relation mellan X och Y. Om vi istället antar att vår data hade kommit från en prospektiv studie, så hade oddsen mellan fall och kontroll när $X = 1$ kunnat beräknas som r_1/r_0 och när $X = 0$ beräknas som $(n_1 - r_1)/(n_0 - r_0)$. Kvoten mellan dessa två odds blir då samma som kvoten i (2.1.1). Det är detta som möjliggör denna grundläggande fall-kontroll-analys, och även mer avancerade studier.

Vi kan också se på proportionerna i de fyra cellerna i Tabell 2.1.1 som sannolikheter,

$$\pi_{00}, \, \pi_{01}, \, \pi_{10}, \, \pi_{11} \, ,$$

där

$$\pi_{xy} = P(X = x, Y = y)$$

och

$$\pi_{00} + \pi_{01} + \pi_{10} + \pi_{11} = 1 \, .$$

Då kan vi uttrycka den relativa risken att vara ett fall givet $X = 1$ eller $X = 0$ genom

$$\frac{P(Y = 1|X = 1)}{P(Y = 1|X = 0)} = \frac{\pi_{11}(\pi_{10} + \pi_{11})}{\pi_{01}(\pi_{00} + \pi_{01})}. \tag{2.1.2}$$

Men denna risk kan endast skattas i en prospektiv studie. Om vi istället tittar på oddskvoten (2.1.1) så kan det visas att följande gäller,

$$\frac{P(Y = 1|X = 1)/P(Y = 0|X = 1)}{P(Y = 1|X = 0)/P(Y = 0|X = 0)} = \frac{P(X = 1|Y = 1)/P(X = 0|Y = 1)}{P(X = 1|Y = 0)/P(X = 0|Y = 0)} \, ,$$

det vill säga att oddskvoten är densamma för retrospektiv data som för prospektiv. Eftersom fall är sällsynt förekommande i situationer där fall-kontrollstudier används, så kommer sannolikheten $P(Y = 0|X = x)$ att vara nära ett, och vi får då en approximation till den relativa risken (2.1.2). Att beskriva oddskvoten på en logaritmisk skala är ofta praktiskt, och oddskvoten kan skrivas som

$$e^{\psi} = \frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}} \, ,$$

där ψ är oddskvoten på log-skala. Vi har nu sett hur oddskvoten, och log-oddskvoten, ger ett mått på risken för att vara ett fall givet inverkan av att vara utsatt för en faktor som antar

värdena 0 och 1. Detta går att generalisera till en förklarande variabel med flera nivåer och log-oddskvoten skattas då med en linjär funktion på följande sätt

$$\hat{\psi_x} = \log\left(\frac{r_{x1}r_{00}}{r_{x0}r_{01}}\right) ,$$

där r_{x1} och r_{x0} är antalet fall respektive kontroller under påverkan av $X = x$ där $x \neq 0$. Här är $X = 0$ valt som en referensnivå, men det skulle kunna vara vilken nivå som helst. Vi kan också generalisera detta vidare till en studie med flera förklarande kategoriska variabler, och log-oddskvoten skattas då enligt

$$\hat{\psi_s} = \log\left(\frac{r_{1s}(n_{0s} - r_{0s})}{r_{0s}(n_{1s} - r_{1s})}\right) ,$$

där n_{1s}, n_{0s} anger antalet fall respektive kontroller med samma värde för en uppsättning förklarande variabler, vilka definierar ett strata s av individer. Antalet fall och kontroller utsatta för en viss variabel anges av r_{1s} respektive r_{0s} i samma uppsättning förklarande variabler s. Varje uppsättning s ger en korstabell liknande Tabell 2.1.1.

I tidiga fall-kontrollstudier studerades huvudsakligen binära eller kategoriska förklarande variabler och då fungerade ovan nämnda metoder bra. Senare ville man även kunna studera mer komplexa situationer, även inkluderande kontinuerliga variabler. En ny modell behövdes för dessa situationer, och man introducerade den *logistiska* regressionsmodellen för att skatta oddskvoten. Log-oddset givet en vektor av förklarande variabler $\mathbf{x}$ modelleras då på följande sätt

$$\log\left(\frac{P(Y = 1|X = \mathbf{x})}{P(Y = 0|X = \mathbf{x})}\right) = \alpha + \boldsymbol{\beta}\mathbf{x}$$

där α betecknar interceptet och $\boldsymbol{\beta}$ är en vektor av log-oddskvoter svarande mot respektive förklarande variabel. Den logistiska regressionsmodellen introduceras mer ingående i avsnitt 2.2.

Fall-kontroll-studier kan alltså innefatta en kombination av flera kategoriska och kontinuerliga förklarande variabler. Vi ser ovan att metoderna skiljer sig till viss del åt, men i grunden ligger skattningen av en oddskvot. Att denna oddskvot blir densamma i en prospektiv studie som i en retrospektiv studie är vad som möjliggör en fall-kontrollstudie och är en viktig del av förståelsen. Logistisk regression är en vanligt förekommande metod för att skatta oddskvoten och det är den metoden vi kommer att använda oss av.

Det finns huvudsakligen två olika sätt att gå till väga för att välja ut kontroller till studien. Dessa beskrivs i de nästföljande avsnitten.

2.1.2 Studie med slumpmässiga kontroller

I en fall-kontrollstudie används normalt alla fall som finns att tillgå, sedan skall kontroller väljas ur en population. Det enklaste sättet att välja dessa kontroller är att göra valet helt slumpmässigt. Förutom den praktiska och ekonomiska fördelen med denna metod, så kan vi på detta sätt få information om ett större antal variabler än de vi primärt är intresserade av. Ett problem med detta är dock att dessa variabler kan göra det svårare att utläsa sambandet mellan utfallet och variabeln vi i huvudsak är intresserade av.

Antag att vi till exempel har en studie där lungcancer är det betraktade utfallet, och rökning är den förklarande variabeln av huvudsakligt intresse. Om kontroller väljs slumpmässigt tenderar fall och kontroller att vara obalanserade med avseende på viktiga riskfaktorer, så som ålder eller kön, vilka inte är av primärt intresse. Eftersom dessa riskfaktorer kan påverka responsen måste vi ta hänsyn till dem i analysen, exempelvis genom att inkludera dessa variabler i regressionsmodellen. Hur kan vi annars veta att det faktiskt är just rökningen som orsakar lungcancer? Exempelvis kanske män är mer benägna att röka, och i sin tur har en större risk att utveckla lungcancer. Dessa variabler, som kan ha en direkt påverkan på utfallet, och som ofta är korrelerade med övriga variabler, kallas *confoundingvariabler*.

Ett problem med att försöka modellera alla dessa variabler är att det kan leda till en sämre precision i skattningen av betydelsen av vår huvudsakliga förklarande variabel, rökning i exemplet ovan. Om ett stort antal confoundingvariabler används finns också en risk att

någon av dem är sällsynt eller rentav aldrig återfinns i fall- eller kontrollgruppen. Detta gör att analysens stabilitet och tillförlitlighet blir negativt påverkad.

Vid användande av slumpmässiga kontroller är logistisk regression (se avsnitt 2.2) det primära verktyget.

2.1.3 Studie med matchade kontroller

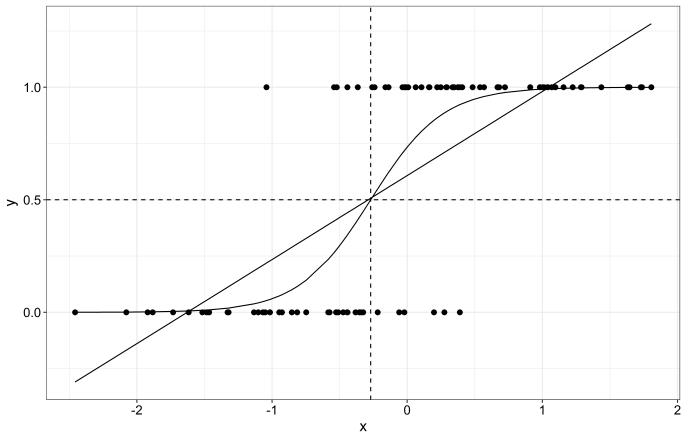
I en *matchad* studie väljs fallen ut på samma sätt som i en studie med slumpmässiga kontroller. Kontrollerna väljs dock ut på ett lite annorlunda sätt. Antalet kontroller som skall matchas ihop med ett fall väljs ofta till en liten heltalsmultipel, exempelvis 1, 2, 4 eller 8. Sedan väljs givet antal kontroller ut slumpmässigt, men som samtidigt matchar på önskad förklarande variabel.

Eftersom effekten av matchningsvariablerna elimineras kan vi få en bättre precision i skattningen av effekterna för de andra variablerna. En annan fördel är att vi förhindrar förväxling av egentliga bakomliggande faktorer, vilket kan förekomma vid slumpmässiga kontroller. Nackdelen vid matchade kontroller är att man förlorar information om de variabler man matchat på. Det finns också problem med att oförsiktiga matchningar kan leda till systematiska fel eller så kallad övermatchning. Om antalet variabler som skall matchas är stort kan det också vara svårt att hitta lämpliga kontroller.

Med matchade kontroller använder man sig av betingad logistisk regression (se avsnitt 2.3), som tar till vara på matchningen, för att skatta effekterna av de givna variablerna.

2.2 Logistisk regression för retrospektiva studier med slumpmässiga kontroller

När responsvariabeln är kontinuerlig är *linjär regression* (se avsnitt 2.4) ofta den naturliga prediktionsmetoden, men denna metod stämmer sämre överens med verkligheten i andra situationer. I synnerhet är linjär regression otillräcklig när responsvariabeln är binär, det vill säga att den endast kan anta två värden, som exempelvis 0 eller 1. I sådana fall är *logistisk regression* en lämpligare metod. En illustration av kurvanpassning genom linjär regression jämfört med logistisk regression visas i Figur 2.2.1. Den logistiska regressionsmodellen predikterar värden mellan 0 och 1 för sannolikheten att reponsvariabeln antar värdet 1, medan den linjära regressionsmodellen förutsäger det faktiska värdet på responsvariabeln, och kan anta värden både under 0 och över 1.



Figur 2.2.1: Den krökta linjen visar sannolikheten för utfallet $y = 1$ utifrån värdet på x-variabeln i en logistisk regressionsmodell. Den räta linjen representerar en ordinarie linjär kurvanpassning. Kurvorna är anpassade efter punkterna i bilden, vilket är observationer på x.

Den logistiska regressionsmodellen utvecklades av David Cox under slutet av 50-talet. I artikeln *The Regression Analysis of Binary Sequences* förklaras teorin bakom denna [6]. Grundtanken med logistisk regression är att man modellerar *sannolikheten* för utfallet $y = 1$ utifrån det specifika värdet på prediktorvariablerna, $\mathbf{x} = (x_1, \ldots, x_p)^T$, genom tillhörande

koefficienter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ och intercept α. Här betecknas *transponat* som T, det vill säga $\mathbf{x}^T$ motsvarar $\mathbf{x}$ transponerat. Låt oss kalla denna sannolikhet, för den i:te observationen, för $\pi_i = P(y_i = 1 \mid \mathbf{x}_i)$. Sannolikheten för utfallet $y = 0$ blir då $P(y_i = 0 \mid \mathbf{x}_i) = 1 - \pi_i$.

Vidare antar vi att detta samband kan modelleras genom *logit*-funktionen, logaritmen av *oddset* (se avsnitt 2.1.1) för utfallet $y = 1$ givet $\mathbf{x}_i$, enligt följande:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} = \alpha + \boldsymbol{\beta}\,\mathbf{x}_i \ .$$

Då följer att oddset

$$\frac{\pi_i}{1-\pi_i} = e^{\alpha + \boldsymbol{\beta}\,\mathbf{x}_i} \ , \tag{2.2.1}$$

och sannolikheten för utfallet $y = 1$ blir

$$\pi_i = \frac{e^{\alpha + \boldsymbol{\beta}\,\mathbf{x}_i}}{1 + e^{\alpha + \boldsymbol{\beta}\,\mathbf{x}_i}} \ . \tag{2.2.2}$$

Likelihood-funktionen, med avseende på α och $\boldsymbol{\beta}$, ser ut som följer, där n betecknar antalet observationer i vårt stickprov:

$$\mathcal{L}(\alpha, \boldsymbol{\beta}) = \prod_{i=1}^{n} {\pi_i}^{y_i} (1-\pi_i)^{1-y_i} \ . \tag{2.2.3}$$

Regressionskoefficienterna, α och $\boldsymbol{\beta}$, estimeras genom maximering av *maximum likelihood*-funktionen i (2.2.3) med hjälp av en iterativ metod, kallad *Fisher's scoring estimation method* [7, s. 88].
Här är y_i utfallet av den i:te observationen, och π_i är sannolikhetsfunktionen i (2.2.2). Kovariansmatrisen för koefficentskattningarna $\hat{\boldsymbol{\beta}}$ estimeras därefter enligt följande:

$$\text{Cov}[\hat{\boldsymbol{\beta}}] \approx \left(\sum_{i=1}^{n} \hat{\pi}(1-\hat{\pi})\mathbf{x}_i\mathbf{x}_i^T\right)^{-1} \ . \tag{2.2.4}$$

Här är $\hat{\pi}_i$ den estimerade sannolikheten från den logistiska modellen. Denna kovariansmatrisskattning baseras på inversen av *informationsmatrisen* i *Fisher scoring estimation method* [7, s. 110].

I vanliga fall krävs det att ett antal grundläggande antaganden om datamängden i fråga är uppfyllda, för att den logistiska regressionsmodellen ska vara tillämpbar. Ett av dessa är att alla y_i ska vara oberoende observationer. Ett annat är att vi inte har några *outliers*, det vill säga att inga av våra observationer avviker på något onaturligt sätt ifrån de övriga.

Den logistiska regressionsmodellen är i grunden anpassad för den typ av data som genereras i en prospektiv studie. Då oddskvoter skattas på ett liknande sätt i retrospektiva studier (se avsnitt 2.1.1) kan det finnas anledning att anta att logistisk regression kan vara applicerbar även på data från denna typ av studie. Det finns också mycket riktigt ett antal intressanta samband mellan retrospektiva och prospektiva studier, rörande parameterskattningarna. Det viktigaste av dessa är följande:

Antag att man analyserar en fall-kontrollstudie, med ett stort antal fall och ett stort antal slumpmässigt valda kontroller, som om det vore en oberoende prospektiv studie. I sådant fall bryter man mot grundantagandet för logistisk regression, att alla observationer är oberoende. De är inte oberoende eftersom man har valt ut observationer utifrån värdet av responsvariabeln y. Det visar sig dock att maximum likelihoodskattningarna av regressionskoefficienterna i detta fall ändå kommer att konvergera mot de från en motsvarande prospektiv studie [8]. Detta med undantag för α-parametern, interceptet, som tappar sin betydelse i en fall-kontrollstudie. Orsaken till det är att man redan vid designen av studien har definierat förhållandet mellan antalet fall och kontroller, och därmed har bestämt sannolikheten för att en observation är ett fall. Denna sannolikhet avspeglas i interceptet, som därmed inte innehåller någon information av större intresse. I retrospektiva studier, som vår, kommer därför parametervektorn $\boldsymbol{\beta}$, och skattningar av denna, att tolkas utan tillhörande interceptkomponent.

2.3 Betingad logistisk regression för retrospektiva studier med matchade kontroller

Vid genomförande av en fall-kontrollstudie med matchade kontroller är vanlig logistisk regression otillräcklig som metod, eftersom matchningen påverkar kovariatfördelningen och den underliggande risken bland kontrollerna. Då krävs istället att den vanliga logistiska regressionsmodellen betingas med matchningen som gjorts vid valet av kontroller. Denna modifierade metod benämns därför *betingad* logistisk regression. För att få en överskådlig introduktion till metoden inleder vi med att betrakta en situation där vi har ett fall och en kontroll, utan matchning. Vi antar också att variablerna i kovariatvektorerna endast kan anta ett begränsat antal diskreta värden. Detta scenario följs sedan av ett likartat scenario, med skillnaden att vi använder oss av matchade kontroller. Resultaten kan även generaliseras till den allmänna situationen.

Betrakta $y_1 = 1$ som fall och $y_2 = 0$ som kontroll, med tillhörande kovariatvektorer $\mathbf{x}_1$ och $\mathbf{x}_2$. Kovariatvektorerna kan innehålla *samspelsvariabler* som är variabler bestående av produkten av två, eller flera, av de enskilda variablerna. Vi tänker oss att det är okänt vilken av kovariatvektorerna som hör ihop med fallet och vilken som hör ihop med kontrollen. Vi är intresserade av den betingade sannolikheten för att $\mathbf{x}_1$ hör ihop med fallet, givet kovariatvektorparet $(\mathbf{x}_1, \mathbf{x}_2)$, och givet att paret består av exakt ett fall och en kontroll.

Den betingade sannolikheten för att $\mathbf{x}_1$ hör ihop med fallet i det givna fall-kontrollparet ser ut på följande vis:

$$\frac{P(\mathbf{x}_1 \mid y = 1)P(\mathbf{x}_2 \mid y = 0)}{P(\mathbf{x}_1 \mid y = 1)P(\mathbf{x}_2 \mid y = 0) + P(\mathbf{x}_1 \mid y = 0)P(\mathbf{x}_2 \mid y = 1)} \,. \tag{2.3.1}$$

Här är $P(\mathbf{x} \mid y)$ den betingade sannolikhetsfunktionen för kovariatfördelningen givet y.

Med hjälp av *Bayes sats* för betingade sannolikheter, vilken ger att

$$P(\mathbf{x} \mid y) = \frac{P(\mathbf{x})P(y \mid \mathbf{x})}{P(y)} \,,$$

kan vi skriva om (2.3.1) som

$$\frac{P(y = 1 \mid \mathbf{x}_1)P(y = 0 \mid \mathbf{x}_2)}{P(y = 1 \mid \mathbf{x}_1)P(y = 0 \mid \mathbf{x}_2) + P(y = 0 \mid \mathbf{x}_1)P(y = 1 \mid \mathbf{x}_2)} \,. \tag{2.3.2}$$

Från den vanliga logistiska regressionsmodellen har vi att

$$P(y = 1 \mid \mathbf{x}_i) = \frac{e^{\alpha + \boldsymbol{\beta}\,\mathbf{x}_i}}{1 + e^{\alpha + \boldsymbol{\beta}\,\mathbf{x}_i}} \,,$$

vilket gör att (2.3.2) kan förenklas till

$$\frac{e^{\boldsymbol{\beta}\,\mathbf{x}_2}}{e^{\boldsymbol{\beta}\,\mathbf{x}_2} + e^{\boldsymbol{\beta}\,\mathbf{x}_1}} \,. \tag{2.3.3}$$

Den sökta sannolikheten för att $\mathbf{x}_1$ hör ihop med fallet, är alltså relaterad till regressionskoefficienterna $\boldsymbol{\beta}$. Notera att α, interceptet, har eliminerats.

I scenariot med matchade kontroller ändras fördelningen för kovariatvektorn för kontrollerna på grund av matchningsurvalet. En del av dessa kovariater kommer överensstämma med kovariaterna hörande till fallen, eftersom de är matchade på en eller flera variabler. Vi behöver nu betinga $P(\mathbf{x}_1 \mid y = 0)$ och $P(\mathbf{x}_2 \mid y = 0)$ med hänsyn till att de är matchade. Sannolikheterna betingas genom att de normaliseras med hjälp av en konstant, κ. Denna normaliseringskonstant bestäms så att summan av sannolikheterna för varje tänkbar kovariatvektor blir lika med 1. Denna konstant blir samma för båda kovariatvektorerna $\mathbf{x}_1$ och $\mathbf{x}_2$, då värdet på κ enbart beror på matchningsvariablerna. De betingade sannolikheterna ges alltså av $\kappa P(\mathbf{x}_1 \mid y = 0)$ och $\kappa P(\mathbf{x}_2 \mid y = 0)$. Den sökta sannolikheten för att $\mathbf{x}_1$ hör ihop med fallet i fall-kontrollparet ges fortfarande av (2.3.1). Det beror på att normaliseringskonstanten är samma för båda sannolikheterna eftersom κ endast beror på matchningsvariablerna. Således kan konstanten strykas i alla termer. Den sökta sannolikheten ser därmed precis likadan ut som tidigare, uttrycket i (2.3.3).

För ett matchat fall-kontrollstickprov fås skattningen av regressionskoefficienterna $\boldsymbol{\beta}$ genom maximering av den betingade likelihooden:

$$\prod_{i=1}^{I} \frac{e^{\boldsymbol{\beta}\mathbf{x}_{2_i}}}{e^{\boldsymbol{\beta}\mathbf{x}_{2_i}} + e^{\boldsymbol{\beta}\mathbf{x}_{1_i}}} = \prod_{i=1}^{I} \frac{1}{1 + e^{\boldsymbol{\beta}(\mathbf{x}_{1_i} - \mathbf{x}_{2_i})}}. \tag{2.3.4}$$

Här är I antal matchade par, och $\mathbf{x}_{1_i}$ och $\mathbf{x}_{2_i}$ svarar mot fallet respektive kontrollen i det i:te matchade paret.

Notera att värdet av matchningsvariablerna är konstant inom de matchade paren. Som resultat av detta förhållande är vissa koordinater i vektorn $\mathbf{x}_{1_i} - \mathbf{x}_{2_i}$ alltid lika med 0. Därmed försvinner matchingsvariablerna från den betingade logistiska regressionsmodellen, och till följd av detta kan inte deras effekt på utfallsvariablen skattas.

Den betingade sannolikheten vi har studerat hittills är $P(\mathbf{x} \mid y)$, givet en uppsättning kovariatvektorer och ett antal fall-kontrollpar. Detta är en naturlig konsekvens av designen av en fall-kontrollstudie. I och med att man väljer observationer utifrån vilka som är fall och icke-fall, kan man se det som att det är $\mathbf{x}$ som är responsvariabeln i modellen. Dock är man oftast intresserad av den omvända sannolikheten, det vill säga $P(y \mid \mathbf{x})$. Precis som i det allmänna fallet gäller här att regressionsparametrarna för motsvarande prospektiva modell, $P(y \mid \mathbf{x})$, kan skattas genom maximering av den betingade likelihooden (2.3.4).

Det går att visa att motsvarande resonemang håller även för fallet där variabler i kovariatvektorer är kontinuerliga, och därmed kan anta ett oändligt antal olika värden, men denna härledning blir ganska omfattande [9]. Vidare går resultatet även att generalisera till en situation där man matchar flera kontroller till varje fall. Skattningen av regressionskoefficienterna $\boldsymbol{\beta}$ för m_i matchade kontroller per fall fås genom maximering av den betingade likelihooden:

$$\prod_{i=1}^{I} \frac{1}{1 + \sum_{j=1}^{m_i} exp[\boldsymbol{\beta}(\mathbf{x}_{ij} - \mathbf{x}_{i0})]}\ .$$

Här är $\mathbf{x}_{ij}$ kovariatvektorn för den j:te matchade kontrollen tillhörande det i:te fallet, och $\mathbf{x}_{i0}$ är kovariatvektorn för det i:te fallet, och I är det totala antalet fall i stickprovet.

2.4 Översikt om linjära regressionsmodeller

Linjär regression är en metod som används för att sammanfatta ett linjärt samband mellan ett antal förklarande variabler och en responsvariabel. Innehållet om linjär regression i detta kapitel refereras till [10, s. 1-92] och innehållet om multivariata normalfördelningar refereras till [11].

Låt oss anta att en linjär modell är tillräcklig för att beskriva sambandet mellan variabler i en datamängd. Låt n vara stickprovsstorleken. En linjär regressionsmodell med en förklarande variabel x och en responsvariabel y kan beskrivas med formeln

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

där $i = 1, \ldots, n$. Feltermen, ε_i, beskriver det som inte kan förklaras av modellen. Vi antar att $\varepsilon_i \sim N(0, \sigma^2)$ för alla i och att feltermerna är okorrelerade. Parametrarna α och β skattas med minsta kvadrat-metoden, genom att minimera

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

Skattningarna erhålls genom att derivera Q med avseende på α och β och sätta derivatorna lika med 0. Skattningarna betecknas $\hat{\alpha}$ och $\hat{\beta}$ och är väntevärdesriktiga, det vill säga $E[\hat{\alpha}] = \alpha$ och $E[\hat{\beta}] = \beta$. Skattningarna för β och α ges av

$$\begin{aligned} \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}$$

där $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ och $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Eftersom $\text{Var}[\varepsilon_i] = \sigma^2$ följer det att $\text{Var}[y_i] = \sigma^2$. Eftersom feltermerna är okorrelerade får vi

$$\text{Var}[\hat{\beta}] = \sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \right)^2 \text{Var}[y_i] \Rightarrow$$
$$\text{Var}[\hat{\beta}] = \frac{\sigma^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \sigma_\beta^2$$

Det gäller också att $\hat{\beta} \sim N(\beta, \sigma_\beta^2)$.

Låt oss istället anta att vi har p antal variabler betecknade $y, x_1, \ldots, x_{p-1}$. En multivariat linjär regressionsmodell kan beskrivas med formeln

$$y_i = \alpha + \sum_{j=1}^{p-1} \beta_j x_{j_i} + \varepsilon_i$$

som är ekvivalent med

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \,, \tag{2.4.1}$$

där $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\beta} = (\alpha, \beta_1, \ldots, \beta_{p-1})^T$ och $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ är vektorer och

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \ldots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{n,p-1} \end{pmatrix}$$

är en $(n \times p)$-matris, där antalet rader är storleken på stickprovet och antalet kolonner är antalet förklarande variablerna i modellen. En rad motsvarar en observation och en kolumn motsvarar värden för en variabel. Även här antar vi att $\varepsilon_i \sim N(0, \sigma^2)$ för alla i. För att skatta $\boldsymbol{\beta}$ används minsta kvadrat-metoden igen, nu genom att minimera

$$\mathbf{Q}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Skattningen erhålls genom att derivera $\mathbf{Q}$ med avseende på $\boldsymbol{\beta}$ och sätta derivatan lika med $\mathbf{0}$:

$$\frac{d\mathbf{Q}}{d\boldsymbol{\beta}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \Rightarrow$$

$$(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \,.$$

$\boldsymbol{\beta}$ skattas nu genom att multiplicera med inversen av $\mathbf{X}^T\mathbf{X}$ på båda sidor:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \,. \tag{2.4.2}$$

Precis som i det endimensionella fallet är $\hat{\boldsymbol{\beta}}$ väntevärdesriktig, det vill säga $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ och variansen av $\hat{\boldsymbol{\beta}}$ ges av

$$\text{Var}[\hat{\boldsymbol{\beta}}] = \text{Var}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{Var}[\mathbf{y}]((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \Rightarrow$$
$$\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \,,$$

ty $\text{Var}[\mathbf{y}] = \sigma^2 I$.

Ovan visas en linjär regressionsmodell där variansen av feltermen antas vara konstant. Annorlunda blir det när variansen varierar bland feltermerna. Då kallas det för *generaliserad linjär regression* [10, s. 417-418]. Antag att $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$, det vill säga att $\boldsymbol{\varepsilon}$-vektorn är multivariat normalfördelad med väntevärde $\mathbf{0}$ och varians $\boldsymbol{\Sigma}$. Konstant varians krävs för att kunna tillämpa teorin för linjära modeller och skatta $\boldsymbol{\beta}$ med (2.4.2). Det är möjligt att transformera $\boldsymbol{\varepsilon}$ så att dess kovariansmatris blir en identitetsmatris genom att multiplicera båda sidorna i (2.4.1) med en lämplig matris $\mathbf{C}$.

$$\mathbf{Cy} = \mathbf{CX}\boldsymbol{\beta} + \boldsymbol{\varepsilon}',$$
$$\text{där} \quad \boldsymbol{\varepsilon}' = \mathbf{C}\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon}' \sim N(0, I) \ .$$

Vi vill hitta matrisen $\mathbf{C}$ och visa dess existens. Vi antar att $\boldsymbol{\Sigma}$ har full rang och därmed är inverterbar. Vi har att

$$\begin{aligned} \text{Cov}[\boldsymbol{\varepsilon}'] = \text{Cov}[\mathbf{C}\boldsymbol{\varepsilon}] = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T = I \Rightarrow \\ \boldsymbol{\Sigma} = \mathbf{C}^{-1}(\mathbf{C}^T)^{-1} = (\mathbf{C}^T\mathbf{C})^{-1} \ . \end{aligned}$$

Då en kovariansmatris alltid är positivt semidefinit kan den delas upp enligt $\boldsymbol{\Sigma} = \mathbf{PDP}^T$, där $\mathbf{P}$ är en ortogonalmatris som består av $\boldsymbol{\Sigma}$:s normerade egenvektorer, och $\mathbf{D}$ är en diagonalmatris, med $\boldsymbol{\Sigma}$:s egenvärden i diagonalen. För den sökta matrisen $\mathbf{C}$ har vi alltså sambandet

$$\begin{aligned} \boldsymbol{\Sigma} = (\mathbf{C}^T\mathbf{C})^{-1} = \mathbf{PDP}^T \Rightarrow \\ \mathbf{C}^T\mathbf{C} = (\mathbf{PDP}^T)^{-1} = (\mathbf{P}^T)^{-1}\mathbf{D}^{-1}\mathbf{P}^{-1} \Rightarrow \\ \{\mathbf{P}^T = \mathbf{P}^{-1} \text{ för } \mathbf{P} \text{ är ortogonalmatris}\} \Rightarrow \\ \mathbf{C}^T\mathbf{C} = \mathbf{PD}^{-1}\mathbf{P}^T \Rightarrow \\ \mathbf{C}^T\mathbf{C} = \mathbf{PD}^{-1/2}\mathbf{D}^{-1/2}\mathbf{P}^T \Rightarrow \\ \mathbf{C}^T\mathbf{C} = (\mathbf{D}^{-1/2}\mathbf{P}^T)^T\mathbf{D}^{-1/2}\mathbf{P}^T \Rightarrow \\ \mathbf{C} = \mathbf{D}^{-1/2}\mathbf{P}^T = \mathbf{P}^T\mathbf{D}^{-1/2} \ , \end{aligned}$$

vilket visar att $\mathbf{C}$ existerar, ty både $\mathbf{P}$ och $\mathbf{D}$ är inverterbara, och därmed är produkter av de båda också inverterbara [12].

Eftersom $\mathbf{C}$ existerar, finns det en transformation som gör att $\boldsymbol{\varepsilon}$ har konstant varians. $\boldsymbol{\beta}$ kan därmed skattas enligt (2.4.2). Vi får

$$\begin{aligned} \hat{\boldsymbol{\beta}} =& ((\mathbf{CX})^T\mathbf{CX})^{-1}(\mathbf{CX})^T(\mathbf{Cy}) \Rightarrow \\ \hat{\boldsymbol{\beta}} =& (\mathbf{X}^T\mathbf{C}^T\mathbf{CX})^{-1}\mathbf{X}^T\mathbf{C}^T\mathbf{Cy} \Rightarrow \\ \hat{\boldsymbol{\beta}} =& (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y} \ , \end{aligned} \tag{2.4.3}$$

ty $\boldsymbol{\Sigma}^{-1} = \mathbf{C}^T\mathbf{C}$.

Variansen för $\hat{\boldsymbol{\beta}}$ ges av $\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T$ där $\mathbf{D} = (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}$. Det är av intresse att förenkla $\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T$ för att få en tydligare bild hur kovariansmatrisen är uppbyggd och få ett mindre och lätthanterligt uttryck. Förenklingen ses nedan:

$$\begin{aligned} \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T = \\ (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}((\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1})^T = \\ (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T((\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1})^T = \\ (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}^T\boldsymbol{\Sigma}^{-1})^T((\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1})^T = \\ (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})((\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1})^T = \\ ((\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1})^T = (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}. \end{aligned} \tag{2.4.4}$$

Sista likheten följer av att $\boldsymbol{\Sigma}$ är symmetrisk vilket gör att $\boldsymbol{\Sigma}^{-1}$ också är symmetrisk, ty inversen av en symmetrisk matris är också symmetrisk. Således följer det att $\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}$ också är symmetrisk vilket medför att $(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}$ är symmetrisk. Sammanfattningsvis så gäller det att $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}$.

2.5 Bootstrap

Bootstrap är en metod som används då skattning av till exempel varians eller skattning av konfidensintervall för en viss parameter är av intresse. Idén med bootstrap är att skatta fördelningen av en parameterskattning genom att återsampla stickprov.

Låt $x_1, \ldots, x_n$ vara ett oberoende stickprov från någon fördelning. Antag att vi är intresserade av en parameter förknippad med fördelningen som stickprovet kommer ifrån. Låt θ vara denna parameter och låt $\hat{\theta}$ vara skattningen från stickprovet. Eftersom stickprovet är slumpmässigt, kommer också $\hat{\theta}$ vara det. Det kan då vara intressant att veta fördelningen av $\hat{\theta}$, för att sedan kunna skatta exempelvis variansen eller konfidensintervall för denna parameter. Om detta inte är möjligt att ta reda på analytiskt, kan bootstrap vara till hjälp. Det finns två typer av bootstrap, *icke-parametrisk* och *parametrisk*. Skillnaden mellan de båda är antagandet om vilken fördelning stickprovet är taget ifrån.

Icke-parametrisk bootstrap används när fördelningen är okänd och syftar till att sampla från stickprovet n gånger med återläggning och på så sätt skapa ett nytt stickprov, med samma antal n element som ursprungsstickprovet. Detta görs B gånger och vi har nu B nya bootstrapstickprov. Från varje bootstrapstickprov får vi en skattning av θ. Vi kallar dessa skattningar $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. Detta ger en *empirisk fördelning* av $\hat{\theta}$, det vill säga en approximativ fördelning av den riktiga fördelningen, från vilken varians och konfidensintervall kan beräknas. Sammanfattningsvis utförs den icke-parametriska bootstrapen på följande sätt:

1. Börjar med att man har ett stickprov, S, med okänd fördelning och av storlek n.
2. Sampla ifrån stickprovet n gånger med återläggning och spara detta som ett nytt stickprov S^*.
3. Utifrån stickprovet S^* skattas θ, sparar det som $\hat{\theta}^*$.
4. Upprepa steg 2 & 3 B gånger för att få B stycken skattningar av θ.

Om vi antar att det ursprungliga stickprovet kommer från en bestämd typ av fördelning, kan vi använda oss av parametrisk bootstrap. Istället för att återsampla nya stickprov från ursprungliga stickprovet, kan vi generera B nya stickprov genom att simulera sådana utifrån den antagna fördelningen med skattad parameter $\hat{\theta}$. Vi skattar $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ från dessa stickprov [13, s. 45-47].

2.5.1 Icke-parametrisk bootstrap för att skatta kovariansen mellan parameterskattningar

Låt $\mathbf{z}_1, \ldots, \mathbf{z}_n$ vara ett stickprov av oberoende observationer från någon multivariat fördelning F. Låt $\boldsymbol{\theta}$ vara en vektor av q intressanta parametrar, där $\hat{\boldsymbol{\theta}}$ är skattningen. Vi samplar stickprovet med återläggning B gånger och får en $\boldsymbol{\theta}$-skattning för varje bootstrapstickprov. Vi kallar dessa skattningar $\hat{\boldsymbol{\theta}}_1^*, \ldots, \hat{\boldsymbol{\theta}}_B^*$. Detta är en empirisk fördelning av $\hat{\boldsymbol{\theta}}$. Denna empiriska fördelning kan nu användas för att skatta $\mathrm{Cov}[\hat{\boldsymbol{\theta}}]$, kovariansmatrisen av $\hat{\boldsymbol{\theta}}$, med följande formel

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_b (\hat{\boldsymbol{\theta}}_b^* - \bar{\hat{\boldsymbol{\theta}}}^*)(\hat{\boldsymbol{\theta}}_b^* - \bar{\hat{\boldsymbol{\theta}}}^*)^T}{B-1} \tag{2.5.1}$$

där $\bar{\hat{\boldsymbol{\theta}}}^* = \sum_b \hat{\boldsymbol{\theta}}_b^*/B$, det vill säga bootstrapmedelvärdet av $\boldsymbol{\theta}$ [13, s. 61-64].

2.5.2 Icke-parametrisk bootstrap med två stickprov

Bootstrap kan tillämpas även när man har två eller flera stickprov. Låt $\mathbf{x} = (x_1, \ldots, x_n)$ och $\mathbf{z} = (z_1, \ldots, z_m)$ vara två oberoende stickprov. Låt $\mathbf{y} = (\mathbf{x}, \mathbf{z})$, vilket motsvarar en vektor av storlek $n+m$. Eftersom $\mathbf{x}$ och $\mathbf{z}$ är tagna från två olika fördelningar, är det viktigt att skilja på dem när bootstrapskattningarna görs. Låt $\mathbf{x}^*$ vara bootstrapstickprovet återsamplat från $\mathbf{x}$ och låt $\mathbf{z}^*$ vara bootstrapstickprovet återsamplat från $\mathbf{z}$. Bootstrapstickprovet för $\mathbf{y}$ blir $\mathbf{y}^* = (\mathbf{x}^*, \mathbf{z}^*)$. Således kan parameterskattningar göras för varje bootstrapstickprov [13, s. 88-89]. Detta kan även generaliseras till multivariata stickprov som i avsnitt 2.5.1. Principen ligger i att bootstrapen användas enskilt på varje stickprov.

I nästa kapitel presenteras icke-parametrisk bootstrap på stickprov innehållande fall och matchade kontroller och stickprov innehållande fall och helt slumpmässiga kontroller. Det är viktigt att fall och slumpmässiga kontroller återsamplas separat. Parameterskattningar från den första delstudien kombineras med det andra och således kan en kovariansmatris mellan parametrarna skattas fram med hjälp av (2.5.1).

3 Metod för sammanvägning av skattade parametrar

I detta kapitel beskrivs metoden som vi tagit fram för att sammanväga skattade parametrar från två delstudier med matchade respektive slumpmässiga kontroller. För att underlätta läsningen kommer metoden hädanefter beskrivas som *vår metod*. Metoden bygger på att en kovariansmatris skattas med hjälp av bootstrap för att i sin tur användas i en linjär regressionmodell för att sammanväga parametrarna från de två delstudierna. Avsnitt 3.1 beskriver hur linjär regression tillämpas för att sammanväga parametrarna och avsnitt 3.2 beskriver hur skattningen av kovariansmatrisen går till. Avsnitt 3.3 introducerar implementationen för ett specialfall, som senare används i simuleringar.

3.1 Sammanvägning av skattade parametervektorer

Låt $\boldsymbol{\gamma}_1$ och $\boldsymbol{\gamma}_2$ vara vektorer med parameterskattningar från två delstudier. Vi antar att $\boldsymbol{\gamma}_1$ och $\boldsymbol{\gamma}_2$ är normalfördelade vektorer med väntevärden $\mathbf{A}_1\boldsymbol{\beta}$ och $\mathbf{A}_2\boldsymbol{\beta}$, där $\mathbf{A}_1$ och $\mathbf{A}_2$ är matriser där raderna svarar mot linjärkombinationer av parametrarna i $\boldsymbol{\beta}$. Antalet kolonner i matriserna är lika med antalet element i $\boldsymbol{\gamma}_1$. Anta att $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T)$ är multivariat normalfördelad och låt $\mathbf{A}^T = (\mathbf{A}_1^T, \mathbf{A}_2^T)$. Låt $\mathrm{Cov}(\boldsymbol{\gamma}) = \boldsymbol{\Sigma}$. Då gäller

$$\boldsymbol{\gamma} = \mathbf{A}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

där $\boldsymbol{\varepsilon} \approx N(0, \boldsymbol{\Sigma})$, det vill säga $\boldsymbol{\varepsilon}$ är approximativt normalfördelad. Vi ser nu att detta är en multivariat linjär regressionsmodell med undantag att $\boldsymbol{\varepsilon}$ inte har konstant varians. Skattningen $\hat{\boldsymbol{\beta}}$ blir då

$$\hat{\boldsymbol{\beta}} = (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma} \tag{3.1.1}$$

enligt (2.4.3). I detta projekt antar vi att $\boldsymbol{\gamma}$ är approximativt multivariat normalfördelad för stora stickprov.

3.2 Skattning av kovariansmatris

För att kunna använda (3.1.1) måste $\boldsymbol{\Sigma}$ vara känd. Med två olika delstudier är inte $\boldsymbol{\Sigma}$ känd och måste därför skattas. Skattningen görs med hjälp av bootstrap. Det är kovariansen mellan skattningarna av $\boldsymbol{\beta}$-parametrarna i en delstudie med slumpmässiga kontroller och $\boldsymbol{\beta}$-parametrarna i en delstudie med matchade kontroller som är intressant att skatta. Icke-parametrisk bootstrap med två stickprov tillämpas då kontroller och fall måste återsamplas separat (se avsnitt 2.5.2). Skattning av kovariansmatris beskrivs stegvis nedan.

1. Antag att vi har två stickprov, med slumpmässiga respektive matchade kontroller. Vi betecknar dessa stickprov S_S och S_M. Stickprovet S_S innehåller n fall och k slumpmässigt valda kontroller och stickprovet S_M innehåller precis samma fall men med m kontroller matchade med varje fall.

2. Sampla fallen n gånger med återläggning. Vi har nu ett bootstrapsstickprov av fallen.

3. Ta ut tillhörande matchade kontroller till de fallen som samplats. Vi har nu ett nytt bootstrapstickprov med matchade kontroller som vi betecknar S_M^*.

4. Sampla de slumpmässiga kontrollerna k gånger med återläggning. Sätt ihop med bootstrapstickprovet av fallen i steg 2. Vi har nu ett nytt bootstrapstickprov med slumpmässiga kontroller som vi betecknar S_S^*.

5. Parametrar skattas genom en logistisk regressionsmodell för studien med slumpmässiga kontroller och betingad logistisk regression för studien med matchade kontroller. Skattningarna görs på stickproven S_S^* och S_M^*.

6. Upprepa steg 2-5 tills vi har B uppsättningar av parameterskattningar.

7. Kovariansmatrisen skattas med hjälp av bootstrapmetoden i avsnitt 2.5.1, där $\hat{\boldsymbol{\theta}}$-vektorn i detta avsnitt består av parametrarna från båda delstudierna.

3.3 Implementation i ett specialfall

I detta projekt inriktar vi oss på ett specialfall, där vi har tre parameterskattningar från en delstudie med slumpmässiga kontroller och två parameterskattningar från en delstudie med matchade kontroller. Vi är intresserade av att skatta β_1, β_2 och β_3 med vår metod, där β_1 och β_2 är effekten av två förklarande variabler, x_1 och x_2, och β_3 är samspelseffekten mellan dem. Parameterskattningarna från delstudien med slumpmässiga kontroller betecknar vi med $\hat{\beta}_1$, $\hat{\beta}_2$ och $\hat{\beta}_3$. I delstudien med matchade kontroller matchar vi på variabeln x_2 och således ges två skattningar som vi betecknar $\tilde{\beta}_1$ och $\tilde{\beta}_3$. För att skatta parametrarna i delstudien med slumpmässiga kontroller använder vi logistisk regression (se avsnitt 2.2). Parametrarna skattas med *glm*-funktionen i `R`. Parametrarna för delstudien med matchade kontroller skattas med betingad logistisk regression (se avsnitt 2.3). Funktionen *clogit* i paketet *survival* används till detta [14]. Implementationen beskrivs mer detaljerat i A.8.

När parametrarna är skattade från de båda delstudierna är det möjligt att väga ihop skattningarna, givet att kovariansmatrisen för dessa är känd, för att få ut en skattning för varje variabel. Kovariansmatrisen $\mathbf{\Sigma}$ skattas med hjälp av bootstrap (se avsnitt 2.5.2 och 3.2) med 500 återsamplingar. Låt $\hat{\mathbf{\Sigma}}$ vara skattningen och antag att $\boldsymbol{\varepsilon} \sim N(0, \hat{\mathbf{\Sigma}})$. Låt nu $\boldsymbol{\gamma} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \tilde{\beta}_1, \tilde{\beta}_3)^T$ och $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$. Låt

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \mathbf{A}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \Rightarrow$$

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

och $\boldsymbol{\beta}$ skattas enligt (3.1.1) med hjälp av $\hat{\mathbf{\Sigma}}$. Även här utförs skattningarna i `R` med hjälp av matrismultiplikationer och inversberäkningar med hjälp av funktionen *ginv* i paketet *MASS* [15].

4 Utvärdering genom simuleringsexempel

I detta kapitel presenteras hur simuleringar används för att undersöka parameterskattningar för simulerad trafikdata. Eftersom teorin bakom vår metod bygger på asymptotiska antaganden, kan simuleringsundersökningar användas för att härma ideala verklighetsscenarion som följer modellen. Simuleringsdatan kan ligga nära verkligheten, med fördelen att vi känner till parametrarna vi har simulerat ifrån. Dessutom kan samma population simuleras flera gånger, för att få en bild av hur bra skattningarna blir i genomsnitt.

Vi tänker oss en trafiksituation där en bil kör på en 70-väg med en bil framför sig som hastigt bromsar in. Fallet som vi undersöker är om en krock uppstår eller inte. Vi har två förklarande variabler, borttittande och hastighet. Bortittande-variabeln antar värdena 0 eller 1 och syftar på en situation där föraren inte fokuserar på vägen, som till exempel kan kontrolleras av en kamera. Det är alltså en skattning av risken att krocka givet dessa två variabler som är av intresse, men även samspelet mellan de två. Samspelet är någonting som ofta glöms bort i fall-kontrollstudier, speciellt i studier med matchade kontroller där samspelsvariabler kan finnas kvar även om minst en av variablerna har eliminerats på grund av matchningen.

Alla simuleringar utförs i programspråket `R`. Simuleringar ger fördelen att vi kan kontrollera hur både vår fall- och kontrollpopulation ska se ut. Vi kommer att veta hur den verkliga risken ser ut och hur nära våra parameterskattningar ligger i jämförelse med verkligheten.

Som tidigare nämnts är vi intresserade av hur mycket bättre eller sämre skattningar blir med vår metod, gentemot skattningar från delstudier med enbart slumpmässiga respektive matchade kontroller. Vi kommer att simulera olika typer av scenarion, där hastighet och borttittande påverkar risken för krock på olika sätt. För att jämföra metoderna mot varandra kommer vi först att titta på parameterskattningarnas varianser. Sedan kommer vi att undersöka en riskminskning för krock i en tänkt population där borttittandet har eliminerats.

4.1 Simulering av data

Vi börjar med att simulera vår hastighetsvariabel, som vi kallar x_2. Detta görs genom funktionen *rsnorm* som ligger i R-paketet *fGarch* [16]. Funktionen generar slumptal från en högerskev normalfördelning. Vi sätter väntevärdet till 70 och standardavvikelsen till 15. Vi väljer denna högerskeva fördelning istället för en symmetrisk normalfördelning då vi anser att det är vanligare att man kör för fort än för långsamt. En korrelation mellan hastighet och bortittande känns naturlig. Till exempel är man mindre benägen att titta bort från vägen vid höga hastigheter och då behöver vi en negativ korrelation mellan de två. Därför väljer vi att simulera bortittande, som vi benämner x_1, utifrån hastigheten. Detta görs genom en logistisk modell på följande sätt

$$P(x_1 = 1) = \pi_{x_1} = \frac{e^{\theta_1 + \theta_2 x_2}}{1 + e^{\theta_1 + \theta_2 x_2}}$$

där θ_1 och θ_2 styr förekomsten av bortittande och korrelationen mellan hastighet och bortittande. Vi ser också att ekvationen ovan ger oss en sannolikhet istället för ett binärt värde. Detta löser vi med hjälp av en bernoulli-fördelning som simuleras i R med funktionen *rbinom*, vi har då

$$x_1 \sim Bin(1, \pi_{x_1}).$$

Vi har nu genererat de förklarande variablerna och har responsvariabeln kvar. Vi kommer igen att använda oss av en logistisk modell och responsvariabeln, som vi kallar y, skapas på följande sätt

$$P(y = 1) = \pi_y = \frac{e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}}{1 + e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}}$$

där parametern α i första hand styr proportionen av krockar i populationen. Parametrarna β_1,β_2 och β_3 justerar risken för krock utifrån borttittande, hastighet och samspelet mellan de två. Således simuleras y med hjälp av en binomialfördelning, det vill säga

$$y \sim Bin(1, \pi_y).$$

På detta sätt skapar vi en population på 200 000 simulerade observationer. Observationerna tänks alltså representera alla inbromsningssituationer på 70 väg i en trafikstudie registrerade av efterföljande bilen. Antalet fall i denna population är beroende av parametrarna och kommer i denna studie variera mellan 30 och 200. Fallen plockas ut ur populationen och kontrollerna väljs från de observationer som är kvar. Vi kallar denna population för kontrollpopulationen.

4.2 Val av kontroller

Vi ska nu skapa två stickprov, ett med slumpmässigt valda kontroller och ett med matchade kontroller. Kontrollerna väljs med en multipel 1, 2, 4 eller 8 gånger antalet fall. För de slumpmässiga stickprovet samplar vi fram önskat antal kontroller från kontrollpopulationen. Detta görs utan återläggning, vilket betyder att varje kontroll är unik.

Stickprovet med matchade kontroller skall nu skapas. Vi vill matcha på hastighet då vi i huvudsak är intresserade av risken för krock vid borttittande. Att matcha på bortittande skulle i praktiken innebära videoannotering och är kostsamt. Vi går igenom varje fall och söker upp det eftersökta antalet (1, 2, 4 eller 8) kontroller i kontrollpopulationen som ligger närmast i hastighet. Om det finns fler kontroller med samma hastighet än vad som önskas, så samplar vi helt slumpmässigt från dessa kontroller till önskat antal. Dessa kontroller matchas sedan ihop med fallen och bildar en grupp. Exempel på hur dessa stickprov kan se ut ser vi i Tabell 4.2.1.

4.3 Jämförelse av paramaterskattningar

Beskrivning av hur regressionparametrarna skattas för studier med slumpmässigt valda kontroller respektive matchade kontroller, samt för vår metod, beskrivs i avsnitt 3.3. Efter att

Tabell 4.2.1: Exempel på ett stickprov med matchade kontroller (vänster) och ett med slumpmässiga kontroller (höger), där det är två kontroller per fall. Fall markerat med fet stil.

Grupp	Krock (Ja/Nej)	Borttittande (Ja/Nej)	Hastighet (km/h)
1	**1**	**1**	**110**
1	0	1	110
1	0	0	110
2	**1**	**0**	**72**
2	0	1	72
2	0	1	72
3	**1**	**1**	**89**
3	0	0	89
3	0	0	89

Krock (Ja/Nej)	Borttittande (Ja/Nej)	Hastighet (km/h)
1	**1**	**110**
0	1	85
0	1	93
1	**0**	**72**
0	0	42
0	0	124
1	**1**	**89**
0	0	63
0	1	87

ha skattat parametrarna jämförs resultaten med vår metod mot de övriga genom att jämföra skattade varianser av parameterskattningar och sedan skattningen på riskminskning då ögonen alltid håller sig på vägen, det vill säga när $x_1 = 0$ för alla observationer.

Vi tänker oss att ett bilföretag utvecklar ett varningssystem som varnar när föraren tittar bort. De är därför intresserade av att veta hur risken för krock kan minskas om man alltid håller ögonen på vägen, det vill säga om $x_1 = 0$ för alla observationer. Med hjälp av simuleringar kan man räkna ut denna riskminskning för en given population. Således kan denna riskminskning skattas med hjälp av de skattade sammanvägda parametrarna och de skattade parametrarna från delstudien med slumpmässiga kontroller. Dessa två modeller har varsin representation av $\boldsymbol{\beta}$-parametrarna. Delstudien med matchade kontroller saknar skattning på β_2 vilket gör att en skattad riskminskning i detta fall inte kommer göras.

4.3.1 Jämförelse av parametrarnas skattade varians

Eftersom $\hat{\boldsymbol{\beta}}$ är en asymptotiskt, approximativt för stora stickprov, väntevärdesriktig skattning av $\boldsymbol{\beta}$, är det relevant att titta på kovariansen av $\hat{\boldsymbol{\beta}}$ för att se hur bra skattningen är. Den skattade kovariansmatrisen av $\hat{\boldsymbol{\beta}}$ i den sammanvägda modellen ges av $(\mathbf{A}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{A})^{-1}$ enligt (2.4.4). Kovariansmatrisen för $\hat{\boldsymbol{\beta}}$ i de två delstudierna med matchade respektive slumpmässiga kontroller ges av formeln i (2.2.4) och beräknas med funktionen *vcov*. Diagonalen i kovariansmatriserna består av variansen för skattningar av β_1,β_2 och β_3. Dessa jämförs som en kvot mot respektive diagonalelement i den skattade kovariansmatrisen för de sammanvägda parameterskattningarna.

4.3.2 Skattning av riskminskning om borttittande elimineras

Givet en simulerad population, räknas antal fall ut, det vill säga antal krockar. Låt detta antal betecknas q. För att kunna veta riskminskningen krävs det att vi vet hur många fall som uppstår då borttittande utesluts. Detta görs genom att simulera nya fall, givet att $x_1 = 0$. Låt y' vara den nya responsvariabeln. Vi har att

$$P(y' = 1) = \pi_{y'} = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 x_2}}{1 + e^{\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 x_2}} = \frac{e^{\alpha+\beta_2 x_2}}{1 + e^{\alpha+\beta_2 x_2}}$$

och y' simuleras med *rbinom*, det vill säga

$$y' \sim Bin(1, \pi_{y'}).$$

Relativa riskminskningen räknas ut genom att ta antal fall i populationen utan borttittande delat med q:

$$\text{Relativ riskminskning} = \frac{\text{antal fall i population utan borttittande}}{q}.$$

Detta ses som den verkliga riskminskningen. Den bästa skattningen av riskminskning är den som ligger närmast detta värde.

Givet parameterskattningarna skattas riskminskningarna. Detta görs genom att skatta hur många fall det finns i populationen, när $x_1 = 0$, och dela detta med q:

$$\text{Skattad relativ riskminskning} = \frac{\text{skattat antal fall i population utan borttittande}}{q}.$$

För att kunna skatta antal fall, behövs α för båda parameteruppsättningarna. Låt $\hat{\beta}_1$, $\hat{\beta}_2$ och $\hat{\beta}_3$ vara parameterskattningarna från stickprovet med slumpmässiga kontroller och låt $\dot{\beta}_1$, $\dot{\beta}_2$ och $\dot{\beta}_3$ vara de sammanvägda parameterskattningarna. Låt N vara populationsstorleken. Parametrarna α och α_s bestäms genom att lösa ekvationerna nedan med hjälp av `R`-funktionen *optimize*.

$$q = \sum_{i=1}^{N} \frac{e^{\alpha+\dot{\beta}_1 x_{1_i}+\dot{\beta}_2 x_{2_i}+\dot{\beta}_3 x_{1_i} x_{2_i}}}{1+e^{\alpha+\dot{\beta}_1 x_{1_i}+\dot{\beta}_2 x_{2_i}+\dot{\beta}_3 x_{1_i} x_{2_i}}}$$

$$q = \sum_{i=1}^{N} \frac{e^{\alpha_s+\hat{\beta}_1 x_{1_i}+\hat{\beta}_2 x_{2_i}+\hat{\beta}_3 x_{1_i} x_{2_i}}}{1+e^{\alpha_s+\hat{\beta}_1 x_{1_i}+\hat{\beta}_2 x_{2_i}+\hat{\beta}_3 x_{1_i} x_{2_i}}}.$$

Sedan sätts $x_1 = 0$ och antalet skattas med hjälp av uttrycken nedan

$$\text{Skattat antal fall, vår metod} = \sum_{i=1}^{N} \frac{e^{\alpha+\dot{\beta}_2 x_{2_i}}}{1+e^{\alpha+\dot{\beta}_2 x_{2_i}}}$$

$$\text{Skattat antal fall, slumpmässiga kontroller} = \sum_{i=1}^{N} \frac{e^{\alpha_s+\hat{\beta}_2 x_{2_i}}}{1+e^{\alpha_s+\hat{\beta}_2 x_{2_i}}}.$$

5 Resultat från simuleringar

I detta kapitel presenteras simuleringsresultaten från jämförelserna mellan vår metod och standardmetoderna för studier med slumpmässiga kontroller respektive matchade kontroller. För att undersöka metodens styrka och svagheter har fem olika simuleringsscenarion använts. Dessa scenarion skiljer sig åt i den simulerade datans struktur, såsom antal fall, korrelationsstruktur mellan bortittande och hastighet, antal kontroller per fall och simuleringsparametrarnas storlekar.

5.1 Simuleringsscenarion

Varje scenario har ett utgångsläge med 200 fall i totalpopulationen och en kontroll per fall. Antalet fall styrs av parametern α i datasimuleringen. Datastrukturen för varje scenario avgörs av olika inställningar på parametrarna $\theta_1, \theta_2, \beta_1, \beta_2$ och β_3, där β-parametrarna styr de kausala effekterna av borttittande, hastighet och av samspelet mellan dessa. Samtidigt styr θ-parametrarna i huvudsak korrelationen mellan bortittande och hastighet samt andelen borttittande i populationen. Därefter simuleras kombinationer av antal fall och antal kontroller per fall, vi kallar dessa för simuleringskombinationer. Varje scenario tillsammans med en sådan kombination simuleras 400 gånger. Kombinationen av antal fall och antal kontroller per fall kan variera mellan scenarion då till exempel en hög korrelation mellan hastighet och borttittande kan göra analysen icke genomförbar med ett litet antal fall. Lösningsalgoritmerna för *glm* och *clogit* kan få problem att konvergera om ett visst bootstrap-stickprov innehåller för få eller ingen kombination av en viss kovariatuppsättning. Om det inträffar dras enligt vår algoritm ett nytt bootstrap-stickprov. För simuleringar med ett litet antal fall kan detta behöva ske många gånger och resultatet kan bli opålitligt. Då kan det krävas ett högt antal kontroller, så som 6 eller 8 gånger fler än fallen, för att kunna uppnå ett pålitligt resultat. En översikt av de olika scenarierna ges i Tabell 5.1.1.

De kommande avsnitten presenterar resultaten för varje scenario med hjälp av figurer och tabeller samt en kort diskussion som lyfter det mest intressanta i varje enskilt scenario. Figurerna består av två grafer vardera. Grafen till vänster i varje figur visar en relativ

jämförelse av variansskattningarna för varje β-parameter mellan de från vår metod och från de två delstudierna var för sig. Värdena som visas i denna graf är ett genomsnitt från 400 simuleringar. Grafen till höger i varje figur visar andelen av de 400 simuleringarna där skattningen av variansen gav ett lägre värde, alltså ett bättre resultat, med vår metod än för de två delstudierna. Den skattade relativa riskminskningen när borttittande elimineras, som diskuterats i avsnitt 4.3.1, används för att i varje simulering med en specifik simuleringskombination skatta den verkliga relativa riskminskningen. Ett medelvärde av dessa skattningar ger en bra approximation av den verkliga riskminskningen och fungerar som ett referensvärde. Referensvärdet används sedan för att se vilken av skattningarna, genom vår metod eller genom delstudien med slumpmässiga kontroller, som ligger närmast referensvärdet. Det bästa respektive sämsta resultatet för alla simuleringskombinationer i ett scenario presenteras i en tabell, där andelen simuleringar där vår metod med en sammanvägd parameterskattning är bäst anges.

Tabell 5.1.1: De olika simuleringsscenariona med tillhörande parametrar och resulterade datastruktur.

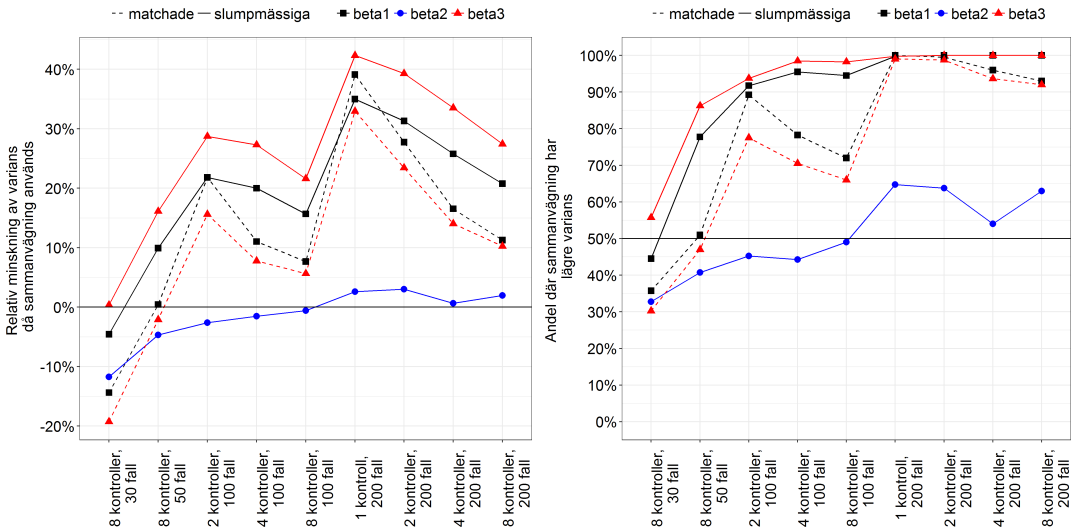
Scenario	θ_1	θ_2	β_1	β_2	β_3	**Beskrivning**
I	-0.02	-0.02	1.5	0.05	0.001	40-50% borttittande bland fall, 20% bortittande bland kontroller, $\text{Korr}(x_1, x_2) = -0.1$
II	-1.4	0	1.5	0.05	0.001	40-50% borttittande bland fall, 20% borttittande bland kontroller, $\text{Korr}(x_1, x_2) = 0$
III	-0.02	-0.02	1.5	0.05	-0.001	40-50% borttittande bland fall, 20% borttittande bland kontroller, Negativt samspel mellan x_1 och x_2, $\text{Korr}(x_1, x_2) = -0.1$
IV	-0.02	-0.02	2.5	0.1	0.001	60-70% borttittande bland fall, 20% borttittande bland kontroller, Höga hastigheter bland fall, $\text{Korr}(x_1, x_2) = -0.1$
V	11	-0.2	2.5	0.05	0.001	40-50% borttittande bland fall, 20% borttittande bland kontroller, $\text{Korr}(x_1, x_2) = -0.5$

5.1.1 Scenario I

Scenario I är ett grundscenario som är tänkt att efterlikna verkligheten, där förekomsten av bortittande bland fall är ungefär dubbelt så hög som bland kontroller. Korrelationen mellan borttittande och hastighet är negativ, vilket innebär att borttittande förekommer mindre frekvent i samband med höga hastigheter. Detta kan antas stämma med verkligheten, då en förare troligen är mer uppmärksam på trafiken vid hög färdhastighet.

Till vänster i Figur 5.1.1 kan vi se hur simuleringskombinationer, när antalet fall är 100-200, ger en klar minskning i varians för vår metod. Varians minskningen gäller i synnerhet för parameterna β_1 och β_3. För den bästa situationen minskar variansskattningen med upp till 40%. För 30 fall med 8 kontroller per fall, kan vi se ett negativt värde för majoriteten av parametrarna. Det betyder att den skattade variansen för vår metod är högre. Linjen med de cirkelformade punkterna visar att variansskattningen för β_2 inte vinner så mycket på sammanvägningen. Vi kan också se att för ett givet antal fall avtar den relativa förbättringen med vår metod med ett ökande antal kontroller per fall. Till höger i Figur 5.1.1 kan vi till exempel se för simuleringskombinationen med 200 fall och 1 kontroll, att variansskattningen minskar i nästan 100% av simuleringarna för β_1 och β_3. Tabell 5.1.2 visar att skattningen av riskminskningen för vår metod är närmare referensvärdet än delstudien med slumpmässiga kontroller i 64% av simuleringarna i den mest fördelaktiga situationen och 53% i den minst fördelaktiga situationen. Vår metod presterar alltså bättre, även om förbättringen inte är

stor.



Figur 5.1.1: Scenario I. Till vänster: Ett genomsnitt av den relativa skattade variansminskningen mellan skattningarna med vår metod och skattningarna från de båda delstudierna. Till höger: Andel simuleringar där variansen blev mindre med vår metod.

Tabell 5.1.2: Scenario I. Tabellen visar andel simuleringar där avståndet är kortare till den verkliga riskminskningen då sammanvägning använts. Endast det minsta och det största resultatet visas.

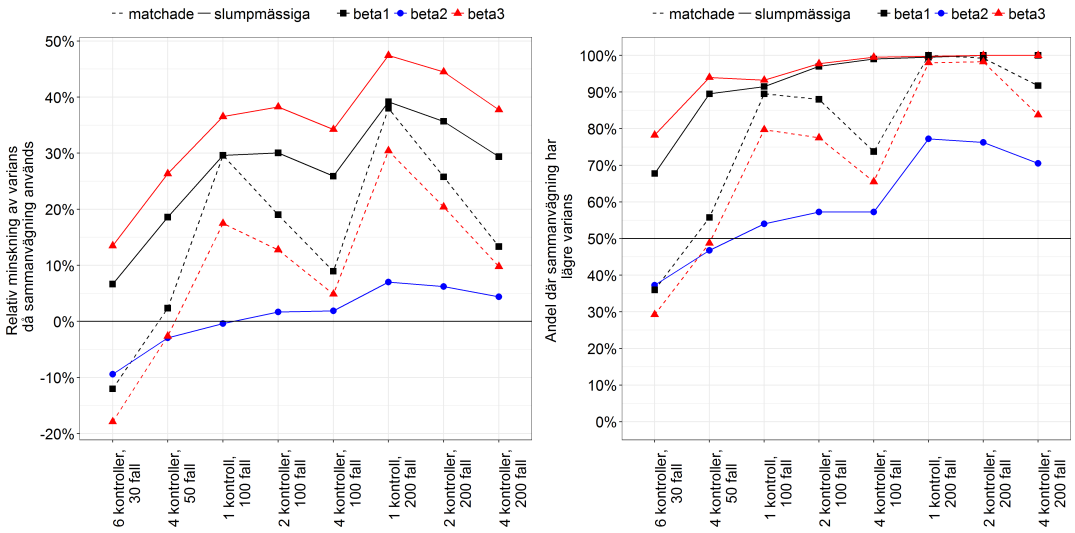
	Andel	**Simuleringskombination**
Minsta	53%	8 kontroller, 100 fall
Största	64%	4 kontroller, 200 fall

5.1.2 Scenario II

Scenario II är i grunden samma som scenario I, med skillnaden att korrelationen mellan de förklarande variablerna bortittande och hastighet är noll. Det innebär att oavsett hur fort föraren kör, kommer föraren att vara lika benägen att titta bort.

Till vänster i Figur 5.1.2 kan vi se hur variansskattningen av β_1 samt β_3, för alla simuleringskombinationer, blir lägre för vår metod jämfört mot delstudien med slumpmässiga kontroller. När antalet fall är 100 eller högre, ger vår metod en bättre skattning än delstudien med matchade kontroller. Speciellt för simuleringskombinationen med 200 fall och 1 kontroll per fall. Där minskar variansen i β_1 och β_3 med cirka 30%. Ett ökat antal kontroller per fall ger även här, precis som scenario I, en mindre fördel för vår metod. Till höger i Figur 5.1.2 kan vi se att den skattade variansen med sammanvägda parametrarskattningar är mindre än den skattade variansen för de två delstudierna i en majoritet av simuleringarna, när antalet fall överstiger 50.

Resultaten för riskminskningsmetoden visas i Tabell 5.1.3. I simuleringskombinationen med 100 fall och 1 kontroll per fall är vår metod bättre än delstudien med slumpmässiga kontroller i 69% av simuleringarna.



Figur 5.1.2: Scenario II. Till vänster: Ett genomsnitt av den relativa skattade variansminskningen mellan skattningarna med vår metod och skattningarna från de båda delstudierna. Till höger: Andel simuleringar där variansen blev mindre med vår metod.

Tabell 5.1.3: Scenario II. Tabellen visar andel simuleringar där avståndet är kortare till den verkliga riskminskningen då sammanvägning använts. Endast det minsta och det största resultatet visas.

	Andel	**Simuleringskombination**
Minsta	57%	4 kontroller, 50 fall
Största	69%	1 kontroll, 100 fall

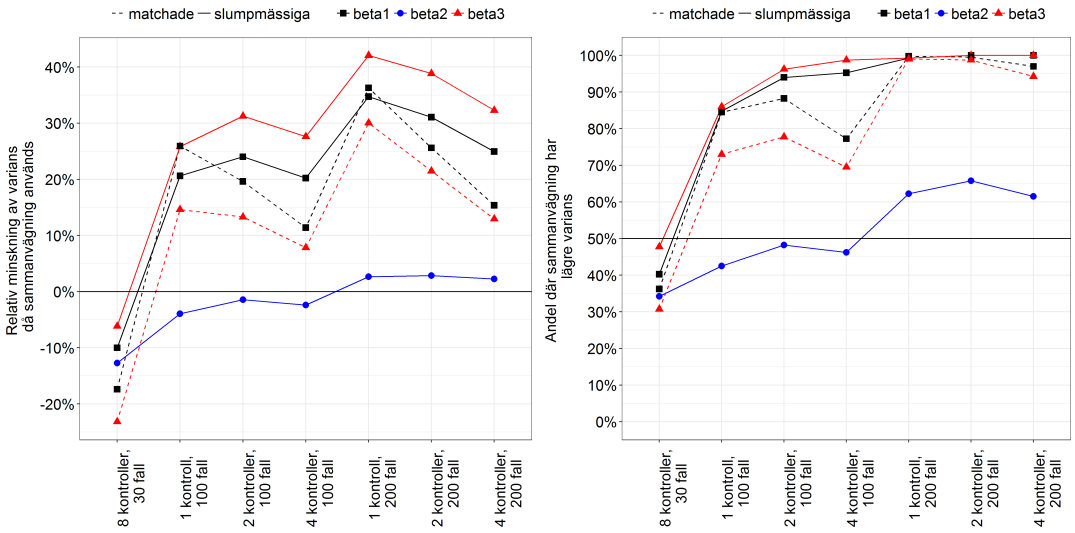
5.1.3 Scenario III

Scenario III liknar återigen scenario I och II, men här är samspelskoefficienten, alltså β_3, negativ. Det ska tolkas som att den relativa riskökningen med att titta bort är lägre vid höga hastigheter än vid låga. Detta innebär dock inte att det är ofarligt att titta bort, då β_1 och β_2, som hör till effekten av borttittande respektive hastighet, båda är positiva.

Resultaten liknar de i scenario I-II, och i Figur 5.1.3 kan vi utläsa en klar förbättring med vår metod, när antalet fall är 100 eller mer. Vi ser dock åter att vår metod inte nämnvärt förbättrar skattningen av β_2, men resultatet har ändå generellt en positiv trend i samband med fler antal fall.

För simuleringskombinationerna med 200 fall och 1-2 kontroller per fall ser vi, till höger i Figur 5.1.3, att variansskattningen är lägre i över 90% av simuleringarna för vår metod.

Resultaten för riskminskningsmetoden visas i Tabell 5.1.4. I simuleringskombinationen med 200 fall och 1 kontroll per fall är vår metod bättre än delstudien med slumpmässiga kontroller i 67% av simuleringarna.



Figur 5.1.3: Scenario III. Till vänster: Ett genomsnitt av den relativa skattade variansminskningen mellan skattningarna med vår metod och skattningarna från de båda delstudierna. Till höger: Andel simuleringar där variansen blev mindre med vår metod.

Tabell 5.1.4: Scenario III. Tabellen visar andel simuleringar där avståndet är kortare till den verkliga riskminskningen då sammanvägning använts. Endast det minsta och det största resultatet visas.

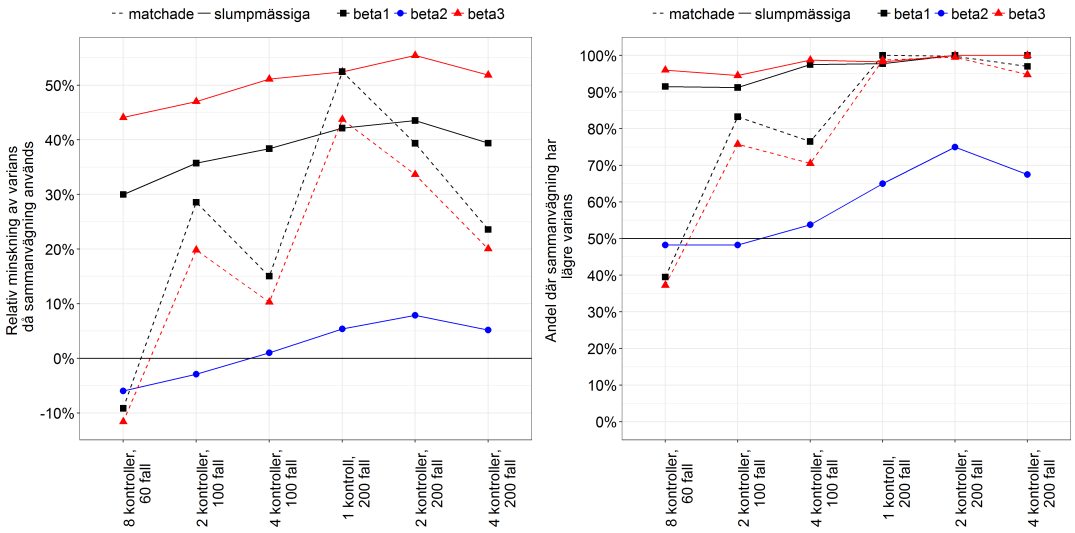
	Andel	**Simuleringskombination**
Minsta	58%	8 kontroller, 30 fall
Största	67%	1 kontroll, 200 fall

5.1.4 Scenario IV

Scenario IV innebär ett starkt samband mellan borttittande och krock, samt hastighet och krock. Risken för krock är alltså stor om föraren kör fort samt tittar bort. Förhoppningen är att det starka sambandet ska framhäva sammanvägningens förmåga att ta till vara på fördelarna hos de två delstudierna.

Figur 5.1.4 visar hur variansskattningen i simuleringskombinationen med 60 fall och 8 kontroller per fall inte blir lägre med vår metod. Men när antalet fall är 100 eller högre ses en stor förbättring. För simuleringskombinationen med 200 fall och en kontroll per fall, blir variansskattningen för β_1 50% lägre med vår metod jämfört med delstudien med matchade kontroller. Detsamma gäller för vår skattning av β_3, jämfört med delstudien med slumpmässiga kontroller. Vi ser återigen en trend med minskad effektivitet av vår metod med ökande antal kontroller per fall. Båda graferna i Figur 5.1.4 visar på en fördel för vår metod när sambanden mellan både hastighet och borttittande och risken för krock är starka.

Resultatet för den skattade riskminskningen då borttittande elimineras ser vi i Tabell 5.1.5. Andelen är minst 69%, vilket innebär att vår metod skattar riskminskningen bättre i majoriteten av simuleringarna för alla simuleringskombinationer.



Figur 5.1.4: Scenario IV. Till vänster: Ett genomsnitt av den relativa skattade variansminskningen mellan skattningarna med vår metod och skattningarna från de båda delstudierna. Till höger: Andel simuleringar där variansen blev mindre med vår metod.

Tabell 5.1.5: Scenario IV. Tabellen visar andel simuleringar där avståndet är kortare till den verkliga riskminskningen då sammanvägning använts. Endast det minsta och det största resultatet visas.

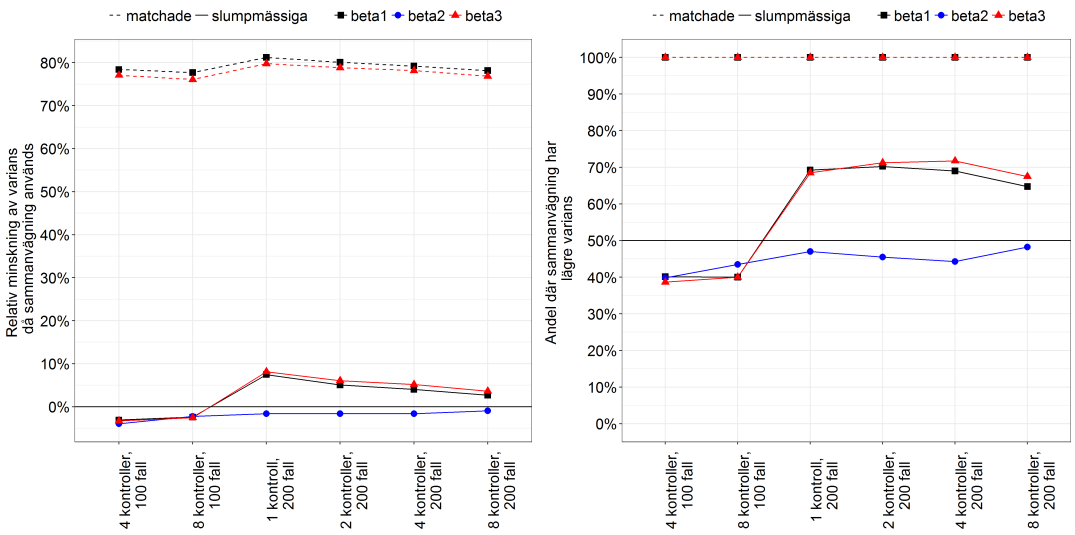
	Andel	**Simuleringskombination**
Minsta	69%	4 kontroller, 100 fall
Största	72%	2 kontroller, 100 fall

5.1.5 Scenario V

Scenario V simulerar en situation då borttittande och hastighet har en mycket stark negativ korrelation. Denna korrelation resulterar i en kraftig minskning av borttittande i höga hastigheter och vice versa. Det är känt att höga korrelationer försvårar skattningar av effekter i regressionsmodeller. Därför är det intressant att se hur vår metod presterar i denna situation.

Resultaten i scenario V skiljer sig mest från de övriga. Till vänster i Figur 5.1.5 ser vi hur variansskattningen är 80% lägre för vår metod än motsvarande från delstudien med matchade kontroller. Vi ser däremot att vår metod endast får en 10% lägre variansskattning jämfört med delstudien med slumpmässiga kontroller för β_1 och β_3. Grafen till höger Figur 5.1.5 speglar resultaten från den vänstra grafen och jämfört med delstudien med matchade kontroller är vår metod bättre i 100% av simuleringarna.

Resultatet för den skattade riskminskningen då borttittande elimineras ser vi i Tabell 5.1.6. En marginell förbättring kan ses, där vår metod är bättre i 52% av simuleringarna i den minst fördelaktiga simuleringskombinationen och 57% i den mest fördelaktiga.



Figur 5.1.5: Scenario V. Till vänster: Ett genomsnitt av den relativa skattade variansminskningen mellan skattningarna med vår metod och skattningarna från de båda delstudierna. Till höger: Andel simuleringar där variansen blev mindre med vår metod.

Tabell 5.1.6: Scenario V. Tabellen visar andel simuleringar där avståndet är kortare till den verkliga riskminskningen då sammanvägning använts. Endast det minsta och det största resultatet visas.

	Andel	**Simuleringskombination**
Minsta	52%	4 kontroller, 100 fall
Största	57%	1 kontroll, 200 fall

6 Diskussion

Syftet med projektet var att undersöka om parameterskattningar i en fall-kontrollstudie kunde förbättras med en metod som sammanväger skattningar från en delstudie med slumpmässiga kontroller och en med matchade kontroller. Projektet har med framgång genomförts och våra resultat visar på en förbättring i ett flertal situationer.

6.1 Resultatdiskussion

Den sammanvägda metoden presterade som bäst i våra simuleringsscenarion när antalet fall och slumpmässiga kontroller var runt 200 och när 1 kontroll matchades per fall. I denna kombination sänktes variansskattningen med upp till 50%. Men även när antalet fall var 100, och i vissa scenarion så lågt som 50, kan förbättringar utläsas (se Figur 5.1.1, 5.1.2, 5.1.3, 5.1.4 och 5.1.5). För lägre antal fall ger vår simuleringar ett sämre resultat för vår metod.

Scenario I-III gav likvärdiga resultat (se Figur 5.1.1, 5.1.2 och 5.1.3). De likvärdiga resultaten är inte förvånande då variationen i datastrukturen för dessa tre scenarion inte var stor. Samtidigt visar det på att vår metod ger en stabil skattning vid små ändringar av datastrukturen och att resultaten kan anses pålitliga. Vi kan även se att förbättringen för skattningen av β_2 (hastighet) inte är lika stor som för de andra parametrarna. Denna marginella förbättring var förväntad då vår metod nästan uteslutande får information om β_2 från delstudien med slumpmässiga kontroller och inte från delstudien med matchade kontroller. Endast samspelet från den matchade delstudien kan inverka på sammanvägning av β_2.

I scenario IV som simulerade starka orsakssamband mellan krock och borttittande respektive hastighet, kan vi se den tydligaste förbättringen av sammanvägningen. Då det finns tydliga kopplingar mellan de förklarande variablerna (hastighet och borttittande) och responsvariabeln (krock) blir sambanden lättare att upptäcka och därmed blir skattningarna bättre

i de båda delstudierna. Vi tror att en bra skattning i båda delstudierna resulterar i en ännu bättre sammanvägd skattning. En idé med att ha matchade kontroller är att öka precisionen i skattningen av β_1 (borttittande) och en fördel med att ha slumpmässiga kontroller är att man får en skattning av β_2. Den sammanvägda skattningen tar tillvara på båda dessa fördelar.

Scenario V simulerade en stark negativ korrelation mellan bortittande och hastighet. Här ses också det mest avvikande resultatet, men fortfarande ett positivt sådant. Intressant i scenario V är att vår metod ger en tydlig förbättring jämfört mot delstudien med matchade kontroller (se Figur 5.1.5). Förbättringen kan bero på att fallen generellt har höga hastigheter och kommer matchas ihop med kontroller som också har höga hastigheter. Kontroller med höga hastigheter har generellt inget borttittande på grund av den höga negativa korrelationen och det försvårar skattningen. Det vill säga, kombinationen att ett fall inte tittar bort ($x_1 = 0$) när en kontroll tittar bort ($x_1 = 1$) är mycket osannolikt. Denna kombination behöver jämföras med kombinationen när fall tittar bort och när en kontroll inte tittar bort för att skatta log-oddset. I de andra scenariorna såg vi en omvänd trend, där sammanvägningen hade en tydlig förbättring gentemot delstudien med slumpmässiga kontroller.

I samtliga scenarion, och i synnerhet i scenario IV, ser vi att den skattade riskminskningen med sammanvägda parametrar är närmare den simulerade verkliga riskminskningen i mer än 50% av alla simuleringar. Vi vill dock påpeka att trovärdigheten på resultaten med just denna riskminskningsmetod kan ifrågasättas. Eftersom skattningarna av riskminskningen låg väldigt nära det verkliga referensvärdet för båda metoderna och skillnaden mellan avstånden inte var särskilt stora. Alldeles för små tal kan leda till stora numeriska fel.

Studerar vi de vänstra graferna i figurerna i kapitel 5 och speciellt simuleringskombinationerna med 200 antal fall, så ser vi att vår metod tappar i effektivitet i relation till referensmetoderna med ett ökande antal kontroller per fall. Förklaringen till detta är att för ett fixt antal fall, ökar vetskapen om de förklarande variablernas fördelning när antal kontroller ökar. För ett stort antal kontroller blir parameterskattningarna i de båda delstudierna väldigt nära verkligheten för just denna population vilket medför att sammanvägningen inte kan förbättra skattningarna. Vi kan också se i figurerna att vår metod presterar sämre än de båda delstudierna med ett litet antal fall. Sammanvägningsmetoden bygger på att en kovariansmatris skattas med hjälp av bootstrap. Fallen återsamplas separat och det är känt att bootstrap fungerar dåligt med ett litet stickprov. Med återsampling är endast ungefär $2/3$ av de unika observationerna med i det nya stickprovet. Ett litet antal fall, som till exempel 50, finns det endast ungefär 33 unika fall kvar i bootstrapstickprovet. Dessutom bygger variansskattningen i (2.4.4) på att vi har stora stickprov, vilket kan göra att variansskattningen blir dålig och gör att sammanvägningsmetoden kanske i själva verket har presterat bättre men att variansskattningen är dåligt skattad.

6.2 Avgränsningar och möjligheter för framtida forskning

I denna rapport har vi fokuserat på två förklarande variabler, dels för vår egen förståelse och dels för att enkelt kunna presentera resultaten. Vi vill framhäva att teorin och själva metoden med sammanvägda parameterskattningar håller för en studie med ett godtyckligt antal förklarande variabler och ett godtyckligt antal matchade variabler. Sammanvägningsmetoden begränsas heller inte till fall-kontrollstudier. Den kan även appliceras på godtyckligt antal parameteruppsättningar från andra studier, speciellt med *godtyckliga linjärkombinationer* (se avsnitt 3.1).

En större simuleringsstudie med flera antal kombinationer av antal fall och kontroller kunde ha varit av intresse för att belysa metodens styrkor och svagheter. Högre variation på scenariorna kan vara intressant att studera för att undersöka i vilka situationer som metoden ger de bästa resultaten. Det kunde också varit av intresse att inkludera fler kovariater i modellen för olika scenarion, för att se hur bra vår metod förhåller sig till det.

En annan undersökning som kunde ha genomförts är hur bra variansskattningarna för $\boldsymbol{\beta}$-parametrarna är. Detta kunde ha jämförts med empiriska skattningar på varianserna, det vill säga genom att plocka ut flera stickprov ur samma population och jämföra de olika $\boldsymbol{\beta}$-skattningarna från dessa stickprov. En jämförelse mellan vår metod och referensmetoderna baserad på empiriska varianser istället för skattade varianser hade även gett pålitligare resultat. Detta eftersom de skattade varianserna kräver stora stickprov för att ge en god

approximation av den sanna variansen. Man skulle även kunna gå vidare med att undersöka normaliteten för de sammanvägda parameterskattningarna, och studera giltigheten av konfidensintervall och statistiska test baserade på approximativ normalitet hos parameterskattningarna.

Då antalet fall varierade för en viss simuleringskombination, kunde det varit bättre att ha skapat en funktion som simulerade en population med ett givet antal fall. Det hade även varit möjligt att simulera fixa antal slumpmässiga kontroller och matchade kontroller direkt från respektive fördelning, istället för att skapa en stor population från vilken kontroller valdes ut. Detta hade underlättat jämförelserna i av simuleringsresultaten och lett till bättre precision i medelvärdesskattningar. Det hade även kunnat göra simuleringarna snabbare och minskat de numeriska felen.

För fortsatta studier av vår metod, samt för att underlätta användadet av en sådan sammanvägning i praktiken, skulle det vara lämpligt att göra ett `R`-paket med implementation av vår metod. Bootstrap skulle eventuellt kunna ersättas med *jackknife*-metoden, som är mer beräkningseffektiv då bootstrap är tidskrävande [13, s. 141-150].

6.3 Kommentar angående rapportens relation till forskningslitteraturen

Under projektets senare del hittade vi ett antal publikationer med liknande idéer om att kombinera parameterskattningar, i fall-kontrollstudier som väger samman resultat från studier med matchade respektive slumpmässigt valda kontroller ([2], [3]). Vi vill dock framhäva att vårt arbete har varit fristående från dessa tidigare publikationer. Vi har också, som nämns i kapitel 1, delvis analyserat andra saker än vad som har gjorts i dessa. Bland annat har vi i större utsträckning hur faktorer som stickprovsstorlek och korrelation mellan variabler påverkar nyttan av sammanvägningen. I mån av tid hade det varit intressant att jämföra vår metod med andra metoder som föreslagits i litteraturen, samt att jämföra resultaten mellan dessa.

Referenser

[1] R. Doll, A. B. Hill, "A study of the aetiology of carcinoma of the lung", *British Medical Journal*, vol. 2, nr. 4797, 1952, ss. 1271-1286.

[2] V. Moreno, et al., "Combined Analysis of Matched and Unmatched Case-Control Studies: Comparison of Risk Estimates from Different Studies", *American Journal of Epidemiology*, vol. 143, nr. 3, 1996.

[3] S. le Cessie, et al., "Combining Matched and Unmatched Control Groups in Case-Control Studies", *American Journal of Epidemiology*, vol. 168, nr. 10, 2008.

[4] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

[5] D. R. Cox, H. Keogh, *Case-Control Studies*, Cambridge: Cambridge University Press, 2014.

[6] D. R. Cox, "The regression analysis of binary sequences", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 20, nr. 2, 1958, ss. 215-242.

[7] A. Agresti, *Categorical Data Analysis*, 2nd edition, Hoboken: Wiley, 2007.

[8] R. L. Prentice, R. Pyke, "Logistic Disease Incidence Models and Case-Control Studies", *Biometrika*, vol. 66, nr. 3, 1979, ss. 403-411.

[9] N. Breslow, "Covariance Adjustment of Relative-Risk Estimates in Matched Studies", *Biometrics*, Vol. 38, No. 3, Special Issue: Analysis of Covariance, 1982, ss. 661-672.

[10] J. O. Rawlings, S. G. Pantula, D. A. Dickey, *Applied Regression Analysis: A Research Tool*, 2nd Edition, Berlin: Springer-Verlag New York, 1998.

[11] R. A. Johnson, D. W. Wichern, *Applied Multivariate Statistical Analysis*, 3rd edition, Upper Saddle River: Prentice-Hall, 1992.

[12] D. C. Lay, *Linear Algebra and its Applications*, 4th edition, London: Pearson Education, 2012.

[13] B. Efron, R. J. Tibshirani, *An Introduction to the Bootstrap*, London: Chapman & Hall/CRC, 1993.

[14] T. Therneau (2015). *A Package for Survival Analysis in S*, version 2.38, URL: http://CRAN.R-project.org/package=survival

[15] W. N. Venables, B. D. Ripley, (2002) *Modern Applied Statistics with S*, Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

[16] D. Wuertz, Y. Chalabi with contribution from M. Miklovic, C. Boudt, P. Chausse and others (2016). fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling. R package version 3010.82.1. https://CRAN.R-project.org/package=fGarch

A Simuleringskod

A.1 Huvudfunktion

```
# Simuleringsfunktion som tar fram resultat för ett specifikt scenario
# Indatan till denna funktion tas fram genom att använda pop_stats (se nedan) för att
# skapa ett visst scenario
# Funktionen returnerar relativ skillnad mellan variansskattningar från en matchad
    delstudie med sammanvägda parametrar och
# en slumpmässig delstudie med sammanvägda parametrar


sim_func <- function(N = 200000, mu = 70, sigma = 15, theta1 = -0.03, theta2 = -0.02,
              alpha = -11, beta_1 = 0.5, beta_2 = 0.05, beta_3 = 0.01, mult_cc = 4){

  A1 <- diag(1, nrow = 3)
  A2 <- rbind(c(1,0,0), c(0,0,1))
  A <- rbind(A1,A2)

  df <- pop_simulation2(N = N, mu = mu, sigma = sigma, theta1 = theta1, theta2 =
      theta2, alpha = alpha,
                      beta_1 = beta_1, beta_2 = beta_2, beta_3 = beta_3)

  case <- cases(df)
  control <- controls(df, all = FALSE, mult_control = mult_cc)
  cc_sample <- rbind(case, control)
  m_sample <- matched_df(df, mult_case = mult_cc)
  sigma <- boot_sigma(m_sample, cc_sample)
  coeff_1 <- est_parameters(cc_sample, matched = FALSE)
  coeff_2 <- est_parameters(m_sample, matched = TRUE)
  y <- beta_est(coeff_1[-1], coeff_2)
  beta_hat <- est_sbeta(sigma, y)

  beta_cov <- ginv(t(A) %*% ginv(sigma) %*% A)

  cov_omatchad <- cov_no_boot(cc_sample, matched = FALSE)[-1,-1]
  cov_matchad <- cov_no_boot(m_sample, matched = TRUE)

  det_o <- det(cov_omatchad)
  det_m <- det(cov_matchad)
  # Beräknar determinanten av de olika kovariansmatriserna
  det_beta_o<- det(beta_cov)
  det_beta_m<- det(beta_cov[c(1,3),c(1,3)])

  kvot_result_no_x1 <- kvot_no_x1(df, cc_sample, m_sample, sigma, beta_hat, coeff_1,
      coeff_2, alpha,
                                beta_1, beta_2, beta_3, N)

  # Beräknar de olika kvoterna och sätter dem i en vektor för att kunna lättare
      rapportera/använda dem
  result <- c(beta_cov[1,1]/cov_omatchad[1,1], beta_cov[2,2]/cov_omatchad[2,2], beta_
      cov[3,3]/cov_omatchad[3,3],
              beta_cov[1,1]/cov_matchad[1,1], beta_cov[3,3]/cov_matchad[2,2], det_beta
                  _o/det_o, det_beta_m/det_m,
              kvot_result_no_x1[1], kvot_result_no_x1[2], kvot_result_no_x1[3])

  names(result) <- c("beta1 kvot omatchad", "beta2 kvot omatchad", "beta3 kvot
      omatchad", "beta1 kvot matchad",
                   "beta3 kvot matchad", "det kvot omatchad", "det kvot matchad", "
                       kvot no x1 omatchad",
                   "kvot no x1 hopvägd", "kvot no x1 real")
```

```
  return(result)

}
```

A.2 Simulera data med rsnorm

```
pop_simulation <- function(N = 200000, mu = 70, sigma = 15, theta1 = -0.85, theta2 =
    -0.02,
                           alpha = -10.8, beta_1 = 0.5, beta_2 = 0.05, beta_3 = 0.01){

  library(fGarch)
  x2 <- abs(rsnorm(n = N,mean = mu,sd = sigma, xi = 2.5)) # simulering av x2 variabel

  px1 <- exp(theta1+theta2*x2)/(1+exp(theta1+theta2*x2)) # simulering av P(x1 = 1)

  x1 <- rbinom(n = N, size = 1, prob = px1) # bortittande dvs x1 = 1

  p <- exp(alpha+beta_1*x1+beta_2*x2+beta_3*x1*x2)/(1+exp(alpha+beta_1*x1+beta_2*x2+
      beta_3*x1*x2)) # simulering av P(y = 1)

  y = rbinom(n=N,size=1,prob=p) # krockar dvs y = 1

  df <- data.frame(y,x1,x2) # simuleringsdatan


  return(df)

}
```

A.3 Funktion som ger en summering av den simulerade datan

```
# funktion som kollar antal fall, korrelation etc för en viss data frame simulerade
    från pop_simulation.

pop_stats <- function(df){

  x1 <- df$x1
  x2 <- df$x2
  y <- df$y

  andel_x1 <- sum(x1)/length(x1) # Andel borttittande
  andel_y <- sum(y)/length(y) # Andel cases
  x1_x2_cor <- cor(x1,x2) # Korrelation mellan hastighet och borttittande
  y_x1_cor <- cor(y,x1) # Korrelation mellan olycka och borttitande
  y_x2_cor <- cor(y,x2) # Korrelation mellan olycka och hastighet
  max_hastighet <- max(x2) # Max hastighet
  min_hastighet <- min(x2) # Min hastighet
  max_hastighet_case <- max(df[df$y==1,]$x2)
  min_hastighet_case <- min(df[df$y==1,]$x2)
  andel_x1_cases <- sum(df$x1[y==1])/sum(df$y) # Andel borttittande bland cases
  andel_x1_controls <- sum(df$x1[y==0])/sum(df$y==0) # Andel borttitande bland
      controls

  stats <- data.frame(andel_x1,andel_y,andel_x1_cases,andel_x1_controls,x1_x2_cor,y_
      x1_cor,y_x2_cor,max_hastighet,min_hastighet,max_hastighet_case,min_hastighet_
      case)

  return(stats)
```

```
}
```

A.4 Funktion som tar ut fallen ur en population

```
# funktion som tar ut fallen i en data frame från pop_simulation

cases <- function(df){
  return(df[df$y == 1,])
}
```

A.5 Funktion som tar ut kontroller ur en population

```
# funktion som plockar ut kontrollpopulationen eller
# slumpmässiga kontroller till en delstudie med slumpmässiga kontroller
# mult_control anger hur många kontroller det är per fall.

controls <- function(df, all = TRUE, mult_control = 2){
  n_controls <- mult_control*sum(df$y)

  if(all == TRUE){
    return(df[df$y == 0,])
  }else{
    control <- df[sample(which(df$y == 0), n_controls, replace = FALSE),]
    return(control)
  }
}
```

A.6 Stickprov med matchade kontroller

```
# skapar att stickprov med matchade kontroller
# mult_case bestämmer hur många kontroller som ska matchas till varje fall
# hastigheten avrundas till heltal
# specifikt för ett fall väljs kontroller slumpmässigt som har samma hastighet
# eller de kontroller som ligger närmast i hastighet från kontrollpopulationen

matched_df <- function(df, mult_case = 1){

  all_control <- controls(df) # alla kontroller
  all_control$x2 <- round(all_control$x2) # avrundar alla hastigheterna för
      kontrollerna till heltal
  case <- cases(df) # alla cases
  case$x2 <- round(case$x2) # avrundar alla hastigheterna för fallen till heltal
  j <- 1

  m_sample <- data.frame(matrix(0,length(case$y)*(mult_case+1),3)) #data frame med
      matchade kontroller

  for (i in seq(1,length(case$y)*(mult_case+1),by = mult_case + 1)){

    control_match <- which(abs(all_control$x2 - case$x2[j]) == min(abs(all_control$x2
        - case$x2[j]), na.rm = TRUE))
    # Kollar vilka kontroller som har den lägsta hastighetsskillnaden jämfört med det
        j:e fallet
    k <- 1
    while(length(control_match) < mult_case){
      # ifall antalet kontroller som har den lägsta hastighetsskillnaden inte är mult
          _case många så kommer man in här,
      # ex om bara en kontrol har samma hastighet som fallet j och mult_case=2 så
          kommer control_match bara ha
```

```
      # längd 1 och vi behöver kolla vilka kontroller som har näst lägsta
          hastighetsskillnaden
      temp_control_match <- which(abs(all_control$x2 - case$x2[j]) == sort(unique(abs
          (all_control$x2 - case$x2[j])),partial=k+1)[k+1])
      # Kollar vilka kontroller som har den näst lägsta hastihetsskillnaden jämfört
          med j:e fallet
      if(length(temp_control_match) > mult_case - length(control_match)){
        # om vi nu hittat fler kontroller än vi behöver så kommer vi nu in här
        temp_control_match <- sample(temp_control_match, mult_case - length(control_
            match))
        # Här slumpar vi mult_case-length(control_match):st kontroller ifrån de näst
            lägsta eftersom att det bara är
        # så många som saknades för att control_match skulla få längden mult_case
      }

      control_match <- c(control_match, temp_control_match) # sätter ihop de
          kontroller vi nu fått fram


      k <- k + 1

    }

    if(length(control_match) > mult_case){
      # Vi behöver bara mult_case:st kontroller så ifall vi har fler så slumpar vi
          mult_case:st av dem
      control_match <- sample(control_match, mult_case)
    }


    m_sample[i,] <- case[j,]
    m_sample[(i+1):(i+mult_case),] <- all_control[control_match,]
    j <- j + 1

    all_control$x2[control_match] <- NA
    # sätter hastigheten på de kontrollerna vi valt till NA för att förhindra att de
        väljs igen

  }

  names(m_sample) <- c("y", "x1", "x2")

  m_id <- sort(rep(c(1:length(case$y)), mult_case + 1))
  # kolumnen med Par-id-nummer för att vi och programmen skall se vilka fall och
      kontroller som hör ihop

  m_sample <- cbind(m_sample, m_id) # Sätter in id kolumnen i matchade samplet

  return(m_sample)

}
```

A.7 Bootstrapskattning av kovariansmatris

```
# m_sample är stickprov med matchade kontroller
# cc_sample är stickprov med slumpmässigt valda kontroller
# B är antal bootstraps

boot_sigma <- function(m_sample, cc_sample, B = 500){

  index_m <- function(numb, mult_case){
```

```
# indexfunktion som gör att kontroller kommer efter fall
return(numb:(numb+mult_case))
}

library(survival)
case <- cases(cc_sample) # fallen
control <- controls(cc_sample, all = TRUE) # slumpmässigt valda kontrollerna

thetastar_m <- matrix(0, nrow = B, ncol = 5) # B:st olika beta-skattningar, skapar
    en tom matris för dessa

case_n <- length(case$y) # antalet fall
control_n <- length(control$y) # antalet slumpmässigt valda kontroller
mult_case <- control_n/case_n # kollar hur fördelningen fall:kontroller är (dvs 1:1
     eller 1:2 eller 1:3 ... etc)

m_id_b <- sort(rep(c(1:length(case$y)), mult_case + 1))
# kolumnen med Par-id-nummer för att vi och programmen skall se vilka fall och
    kontroller som hör ihop

for(i in 1:B){
  boot_bad <- TRUE # boolean som säger om bootstrapstickprovet är dåligt eller bra

  while(boot_bad){ # While loop för att fånga alla errors och varningar och göra om
       deras samplingar
    tryCatch({ # tryCatch kommandot fångar upp errors och warnings, om detta sker g
        örs bootstrapen om
      index_case <- sample(1:case_n, case_n, replace = TRUE)
      # slumpar fall rad-indexes med återställelse

      index_control <- sample((case_n+1):(case_n + control_n), control_n, replace =
           TRUE)
      # slumpar kontroll rad-indexes med återställelse, börjar i case_n+1 för att
          vi kommer lägga dem efter index_case

      subset_unmatch <- c(index_case, index_control)

      mm <- glm(y ~ x1 + x2 + x1*x2, 'binomial', data = cc_sample, subset = subset_
          unmatch)
      # passar vår generaliserade linjära model y ~ x1 + x2 + x1*x2 till
          stickprovet med slumpmässiga kontroller
      # för att skatta beta_1, beta_2 och beta_3 parametrarna i det omatchade
          samplet

      index_case <- (mult_case+1)*sort(index_case)-mult_case
      # Beräknar fram rad-indexes för fallen i stickprovet med matchade kontroller

      subset_match <- as.vector(sapply(index_case, FUN = index_m, mult_case = mult_
          case))
      # Vektor med fall-indexes och deras korresponderande kontroll-indexes

      boot_sample <- m_sample[subset_match,]
      # skapar bootstrapstickprovet vi kommer att köra betingad logistiska
          regression på
      row.names(boot_sample) <- 1:(dim(m_sample)[1])
      # sätter radnamnen i boot_sample till att vara 1 till och med antalet rader i
           det matchade samplet (dvs, 1,2,3,...)

      match_model <- clogit(y ~ x1 + x1:x2 + strata(m_id_b), data = boot_sample)
      # clogit skattar en logistisk regressions model på/till vårt givna stickprov
          med matchade kontroller
      # för att skatta beta_1 och beta_3 parametrarna
```

```
      thetastar_m[i,c(1,2,3)] <- mm$coefficients[-1] # [-1] för att ta bort
          intercept estimeringen
      thetastar_m[i,c(4,5)] <- match_model$coefficients # De skattade parametrarna
          från stickprovet med matchade kontroller
      boot_bad <- FALSE # om inget error eller warning har hänt så kommer boot_bad
          att bli och förbli FALSE
    }, error = function(e){ # Om ett error har uppstått (i antingen glm eller
        clogit) så kommer den att fångas här
      print(e) # printar ut error meddelandet så man ser vad det var som blev fel
      print(i)
      # printar även ut vilket nummer i bootstrappen man är på så att man kan se
          ifall man sitter helt fast eller ej
      return(boot_bad <- TRUE)
      # sätter boot_bad till TRUE om ett error(eller warning) har uppstått så att
          just den boot:en kan göras om
    }, warning = function(w){ # Om en warning har uppstått så kommer den att fångas
        här
      print(w) # printar ut warning meddelandet så man ser vad det var som blev fel
      print(i) # samma här som i error
      return(boot_bad <- TRUE) # samma här som för error
    }
    )


  }
 }
 return(cov(thetastar_m))
}
```

A.8 Parameterskattningar för de två delstudierna

```
# sample_df är stickprovet med antingen matchade eller slumpmässiga kontroller
# matched är en boolean som indikerar om stickprovet har matchade eller slumpmässiga
    kontroller
# funktionen retunerar skattningar på beta

est_parameters <- function(sample_df, matched = FALSE){

 if(matched == FALSE){

   mm1<-glm(y~x1+x2+x1*x2,'binomial', data = sample_df)

   return(mm1$coefficients)

 }else{

   match_model <- clogit(y ~ x1 + x1:x2 + strata(m_id), data = sample_df)

   return(match_model$coefficients)

 }

}
```

A.9 Kovariansmatriser för delstudier

```
# Kovariansmatrisen för beta i de två delstudierna
# sample_df är ett stickprov med antingen matchade eller slumpmässiga kontroller
```

```
# matched är en boolean som indikerar om stickprovet har matchade eller slumpmässiga
    kontroller

cov_no_boot <- function(sample_df, matched = FALSE){

  if(matched == FALSE){

    mm1<-glm(y~x1+x2+x1*x2,'binomial', data = sample_df)

    return(vcov(mm1))
  }else{

    match_model <- clogit(y ~ x1 + x1:x2 + strata(m_id), data = sample_df)

    return(vcov(match_model))

  }

}
```

A.10 Funktion som lägger ihop två vektorer till en vektor

```
beta_est <- function(coeff1, coeff2){ # Liten funktion för att sätta ihop två
    vektorer till en matris

  return(as.matrix(c(coeff1,coeff2)))

}
```

A.11 Skatta parametrar med hjälp av sammanvägning

```
# Funktion som sammanväger parametrar från två delstudier
# sigma är skattade kovariansmatrisen skattad från boot_sigma
# gamma är de skattade parametrarna från de två delstudierna

est_sbeta <- function(sigma, gamma){
  library(MASS)
  sigma_inv <- ginv(sigma) # inverterar sigma matrisen

  A1 <- diag(1, nrow = 3)
  A2 <- rbind(c(1,0,0), c(0,0,1))
  A <- rbind(A1,A2)

  D <- ginv(t(A) %*% sigma_inv %*% A) %*% t(A) %*% sigma_inv

  beta_hat <- D %*% gamma

  return(beta_hat)
}
```

A.12 Riskminskning då borttittande elimineras

```
# Funktionen returnerar tre värden
# Första är riskminskningsskattningen för en delstudie med slumpmässiga kontroller
# Andra är riskminskningsskattningen för sammanvägda skattade parametrar
# Tredje är det riktiga värdet för den specifika populationen

kvot_no_x1 <- function(df, cc_sample, m_sample, sigma, beta_hat, coeff_1, coeff_2,
    alpha,
```

```
                    beta_1, beta_2, beta_3, N){

  p_func <- function(alpha, beta_1, beta_2, beta_3, x1, x2){
    # Funktion för att beräkna sannolikheten för olycka (dvs sannolikheten för att y
        skall vara 1)
    return(exp(alpha+beta_1*x1_mod+beta_2*x2_mod+beta_3*x1_mod*x2_mod)/(1+exp(alpha+
        beta_1*x1_mod+beta_2*x2_mod+beta_3*x1_mod*x2_mod)))
  }

  alpha_func <- function(alpha,beta_1,beta_2,beta_3,x1,x2,y_antal){
    # Funktion för att användas vid beräkningen av det optimala alpha:t för att
        minimera nedanstående ekvation
    abs(sum(exp(alpha+beta_1*x1+beta_2*x2+beta_3*x1*x2)/(1+exp(alpha+beta_1*x1+beta_2
        *x2+beta_3*x1*x2))) - y_antal)
  }

  case <- cases(df)
  real_case_amount <- sum(case$y) # antal fall

  x1 <- df$x1
  x2 <- df$x2

  # modifierar ett sample så att man aldrig kollar bort, för att see hur mycket b
      ättre/säkrare det blir
  x1_mod <- 0
  x2_mod <- df$x2

  p <- p_func(alpha, beta_1, beta_2, beta_3, x1_mod, x2_mod) # Sannolikheten för
      olycka i det modifierade samplet

  y_mod = rbinom(n=N,size=1,prob=p)

  real_case_amount_mod <- sum(y_mod) # antal fall då datan är modifierad

  opt_alpha_om <- optimize(alpha_func, c(-100,100),beta_1=coeff_1[2],beta_2=coeff_
      1[3],beta_3=coeff_1[4],x1=x1,x2=x2,y_antal=real_case_amount)
  # Beräknar fram det optimala alpha:t för det slumpmässiga förhållandet och för det
      sammanvägda förhållandet
  opt_alpha_hop <- optimize(alpha_func, c(-100,100),beta_1=beta_hat[1],beta_2=beta_
      hat[2],beta_3=beta_hat[3],x1=x1,x2=x2,y_antal=real_case_amount)

  amount_hat_s <- sum(p_func(opt_alpha_om$minimum, coeff_1[2], coeff_1[3], coeff_
      1[4], x1_mod, x2_mod))
  # Skattar antalet y:n som vi får i det modifierade samplet i det omatchade förh
      ållandet och i det sammanvägda förhållandet
  amount_hat_sam <- sum(p_func(opt_alpha_hop$minimum, beta_hat[1], beta_hat[2], beta_
      hat[3], x1_mod, x2_mod))

  return(c(amount_hat_s/real_case_amount, amount_hat_sam/real_case_amount, real_case_
      amount_mod/real_case_amount))
}
```