Stockholms universitet

Skattningsmetoder för binär data: En simuleringsstudie

Greta Olsson Lööf

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2017:7
Matematisk statistik
Juni 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2017:7**
http://www.math.su.se/matstat

Skattningsmetoder för binär data: En simuleringsstudie

Greta Olsson Lööf*

Juni 2017

Sammanfattning

Logistic regression is useful when analyzing binary data, where in the analysis one can use the approximations taken from analyses of large samples. These approximations are not as precise when working with small samples. This thesis aims to test alternative estimation methods - maximum likelihood estimation and exact conditional inference using statistical software SAS and R, and see if the methods are suitable for this type of analysis. A simulation study was constructed containing two models - one with an independent binary variable and one with two independent binary variables. Comparisons were made based on the true parameter values in the models from which the data was generated, and the parameter estimates given from each reviewed method.

The thesis finds some support for the initial theory that when using exact conditional inference the estimated parameters are less biased than when using the maximum likelihood for analysis of small samples. However, more research is needed to try this on non-simulated data and test its applicability.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: gretaolsson999@hotmail.com. Handledare: Jan-Olov Persson.

Förord

Jag vill tacka min handledare Jan-Olov Persson för vägledning i detta kandidatarbete. Stödet från min sambo Sebastian Bergendorf har varit fantastiskt under arbetets gång. Jag tackar min nära vän, Josefine Öderyd, som lånade ut svårtillgänglig litteratur åt mig. Avslutningsvis vill jag påtala min tacksamhet till familj och vänner.

Contents

1 Inledning **5**
- 1.1 Bakgrund . . . 5
- 1.2 Syfte och Metod . . . 5
  - 1.2.1 Problemformulering . . . 5
  - 1.2.2 Syfte . . . 6
- 1.3 Metod . . . 6
  - 1.3.1 Simuleringsstudien och dess logistiska regressionsmodeller 7
  - 1.3.2 Simulerad data och dess egenskaper . . . 8
  - 1.3.3 Utvärdering av simuleringen . . . 8
  - 1.3.4 Avgränsningar . . . 9

2 Teori **10**
- 2.1 Logistisk regression . . . 10
- 2.2 Mjukvarupaket som används i uppsatsen . . . 12
  - 2.2.1 ExLog i SAS . . . 12
  - 2.2.2 Elrm i R . . . 13
  - 2.2.3 MLE i R . . . 13

3 Resultat och Analys **13**
- 3.1 Systematisk skillnad mellan medelvärden och sanna parametrar . 14
- 3.2 Modell 1 . . . 14
  - 3.2.1 Modell 1, stickprovsstorlek 10 . . . 14
  - 3.2.2 Modell 1, stickprovsstorlek 20 . . . 17
  - 3.2.3 Modell 1, stickprovsstorlek 50 . . . 20
  - 3.2.4 Modell 1, stickprovsstorlek 100 . . . 22
  - 3.2.5 Modell 1 - Systematisk skillnad mellan medelvärden och sanna parametrar . . . 24
- 3.3 Modell 2 . . . 24
  - 3.3.1 Modell 2, stickprovsstorlek 10 . . . 25
  - 3.3.2 Modell 2, stickprovsstorlek 20 . . . 27
  - 3.3.3 Modell 2, stickprovsstorlek 50 . . . 28
  - 3.3.4 Modell 2, stickprovsstorlek 100 . . . 31
  - 3.3.5 Modell 2 - Systematisk skillnad mellan medelvärden och sanna parametrar . . . 33
- 3.4 Komparativa skillnader och likheter mellan modellerna . . . 33

4 Diskussion **34**
- 4.1 Diskussion av simuleringen . . . 34
- 4.2 Studien i en större kontext . . . 36

5 Referenser **37**

6 Appendix **38**
6.1 Odds och Oddskvot 38
6.2 Exemplifiering av exakt betingad inferens 39
6.3 Kompletterande figurer och tabeller för Modell 1 44
6.4 Kompletterande figurer och tabeller för Modell 2 52

1 Inledning

1.1 Bakgrund

Logistiska regressionsmodeller har blivit populära vid undersökning av kvantitativa forskningsområden inom många vetenskapliga ämnen. Medan Maximum Likelihood-Skattningar ofta genererar träffsäkra resultat vid analys av stora stickprov, besitter dessa skattningar inte samma egenskaper vid analys av små stickprov. Eftersom forskaren inte alltid har möjlighet att samla in mer data, väcker det en intressant och viktig fråga - hur kan rimliga parameterskattingar tas fram i en logistisk modell vid analys av ett litet stickprov?

Logistisk regression är en användbar metod vid analys av data där responsvariabeln är binär och alltså bara kan anta två värden. För att skatta parametervärdena i en logistisk regressionsekvation används ofta Maximum Likelihood-Skattning. Genom att maximera den obetingade Likelihoodfunktionen får man fram parameterskattningar som antas vara konsistenta, asymptotiskt effektiva och asymptotiskt normalfördelade. Antagandet om normalitet tillåter beräkning av konfidensintervall och utförande av statistiska tester så som till exempel Wald-tester, Likelihoodkvot-tester eller Score-tester[1]. Dessa approximationer kan inte antas med samma säkerhet när stickprov med en liten mängd observationer analyseras, och studier har påvisat bias i parameterskattningarna och där dess storhet beror på stickprovets storlek och struktur.[2]

Denna uppsats undersöker två alternativ till skattningsmetoden Maximum Likelihood-Skattning vid analys av data där responsvariabeln är binär - exakt betingad inferens ($ExLog_{SAS}$) och approximativ betingad inferens ($Elrm_R$). Dessa undersöks genom simuleringar i mjukvaruprogrammen SAS respektive R. Maximum Likelihood-Skattning (MLE_R) används som en kontrollmetod och referenspunkt för de alternativa skattningsmetoderna och utförs i mjukvaruprogrammet R.

1.2 Syfte och Metod

I denna del presenteras uppsatsens problemformulering och syfte för läsaren. Vidare introduceras läsaren för den metod som rapportförfattaren har använt sig utav, beskriven i detalj.

1.2.1 Problemformulering

Vid analys av data där responsvariabeln är binär beskrivs ofta dess relation med förklarande variabler med en logistisk regressionsekvation enligt modellen nedan

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_j x_j,$$

där $\pi_i = Pr(Y_i = 1|X_1 = x_1, X_2 = x_2 ... X_j = x_j)$, är sannolikheten att Y_i antar värdet 1 givet en uppsättning förklarande variabler och där α, $\beta_1, ..., \beta_j$

är regressionskoefficienter [1].

För att skatta parametervärdena i en logistisk regressionsekvation används ofta Maximum Likelihood-Skattning (MLE) där man genom att maximera den obetingade Likelihoodfunktionen får fram parameterskattningar som vid analys av stora stickprov antas vara asymptotiskt icke-biased, alltså att $E[\hat{\beta}^{MLE}] \approx \beta$. Long (1997) har bl.a. visat att vid analys av små stickprov är MLE-skattningarna bias. Därför undersöks i denna rapport olika skattningsmetoder vid olika stickprovsstorlekar, simulerade från två olika modeller, för att klargöra hur deras parameterskattningar förhåller sig till varandra och till de sanna parametervärdena.

1.2.2 Syfte

Uppsatsen ämnar undersöka, med hjälp av simuleringsstudier, huruvida det föreligger någon systematisk skillnad mellan skattningsmetoderna $ExLog_{SAS}$, två alternativ av $Elrm_R$ som betecknas A och B (se även avsnitt 2.2.2) och MLE_R, och de sanna parametervärdena i de modeller från vilka datan är simulerad. Studien undersöker, för stickprovsstorlekar $10, 20, 50$ och 100, $\boldsymbol{\beta}-$skattningarna från de olika skattningsmetoderna vid analys av data simulerad från två modeller beskrivna nedan. Studien strävar efter att klargöra och utvärdera skillnader mellan metoderna och undersöka huruvida det råder någon form av skevhet i metoderna - detta är uppsatsens ambition.

Två olika modeller används för att undersöka problemformuleringen. Vi betecknar ekv 1 som Modell 1 och ekv 2 som Modell 2, vilka kan beskrivas som följande

$$\text{Modell 1} \qquad log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_1 = -1 + 2x_1, \tag{1}$$

och

$$\text{Modell 2} \qquad log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 = 0.2 + 1.5x_2 - 1x_2, \tag{2}$$

där de sanna parametervärdena för α och β_1 i Modell 1 är -1 respektive 2, och de sanna parametervärdena för α, β_1 och β_2 i Modell 2 är 0.2, 1.5 respektive -1. Modellerna beskrivs närmare i nedanstående avsnitt.

1.3 Metod

Nedan förklaras hur simuleringsstudien har utförts och hur utvärderingen av resultaten av simuleringen görs för skattningsmetoderna $ExLog_{SAS}$, A-, $B-Elrm_R$ och MLE_R.

För att utvärdera om det råder någon form av skevhet i metoderna simuleras 1000 stickprov från Modell 1 och 2 i stickprovsstorlek $10, 20, 50$ och 100. Från de olika skattningsmetoderna $ExLog_{SAS}$, A-, B-$Elrm_R$ och MLE_R fås $\boldsymbol{\beta}-$skattningar

som sedan jämförs med varandra och parametervärdena i Modell 1 - $\beta_1 = 2$ och Modell 2 - $\beta_1 = 1.5$ och $\beta_2 = -1$.

Resultaten kommer att ge insikt om hur skevheten i parameterskattingarna förhåller sig till stickprovsstorleken för de olika skattningsmetoderna. Vidare beskrivs i detta metodavsnitt simuleringsstudien och övriga metodval i närmare detalj.

1.3.1 Simuleringsstudien och dess logistiska regressionsmodeller

De stickprov som ska analyseras med hjälp av de olika skattningsmetoderna är genererade från två logistiska regressionsmodeller, benämnda Modell 1 och Modell 2.

För Modell 1 låter vi Y vara en responsvariabel av dikotom karaktär och X_1 vara en förklarande variabl av samma karaktär, $Y \sim Be(\pi)$ och $X_1 \sim Be(1/2)$. Låt sedan $\pi = Pr(Y = 1|X_1 = x_1)$. Vi har då följande logistiska regressionsmodell:

$$logit[\pi] = log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_1, \tag{3}$$

där

$$\pi(x) = \frac{exp(\alpha + \beta_1 x_1)}{1 + exp(\alpha + \beta_1 x_1)},$$

och de sanna parametervärdena för α respektive β_1 är -1 respektive 2. Vi har då att

- $Pr(Y = 1|X_1 = 0) = 0.27$,
- $Pr(Y = 1|X_1 = 1) = 0.73$,

där sannolikheterna avrundats till två decimaler.

För Modell 2 låter vi Y vara en responsvariabel av dikotom karaktär och X_1, X_2 vara förklarande variabler av samma karaktär, $Y \sim Be(\pi)$ och $X_i \sim Be(1/2)$, för $i = 1, 2$. Låt sedan $\pi = Pr(Y = 1|X_1 = x_1, X_2 = x_2)$. Vi har då följande logistiska regressionsmodell:

$$logit[\pi] = log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2, \tag{4}$$

där

$$\pi(x) = \frac{exp(\alpha + \beta_1 x_1 + \beta_2 x_2)}{1 + exp(\alpha + \beta_1 x_1 + \beta_2 x_2)},$$

och de sanna parametervärdena för α, β_1 respektive β_2 är 0.2, 1.5 respektive -1. Vi har då att

- $Pr(Y = 1|X_1 = 0, X_2 = 0) = 0.55$,
- $Pr(Y = 1|X_1 = 1, X_2 = 0) = 0.85$,
- $Pr(Y = 1|X_1 = 0, X_2 = 1) = 0.31$,
- $Pr(Y = 1|X_1 = 1, X_2 = 1) = 0.67$,

där sannolikheterna avrundats till två decimaler.

Från de simulerade stickproven skattas $\boldsymbol{\beta}-$parametrnarna med hjälp av de olika skattningsmetoderna $ExLog_{SAS}$, A- och B-$Elrm_R$ och MLE_R, där den förstnämnda metoden utförs i mjukvaruprogrammet SAS och de resterande utförs i mjukvaruprogrammet R. Dessa redovisas mer utförligt i avsnitt 2.2.

1.3.2 Simulerad data och dess egenskaper

Vid analys av simulerad data kan man göra undersökningar under kontrollerade förhållande. De skattade $\boldsymbol{\beta}-$parametrarna jämförs från de olika skattningsmetoderna $ExLog_{SAS}$, A-, B-$Elrm_R$ och MLE_R med de sanna parametervärdena i de modeller stickproven har simulerats från. I linje med uppsatsens syfte kan alltså simuleringsstudier hjälpa till att undersöka stickprov av olika storlekar som är simulerade från samma modell och för olika skattningsmetoder. Då uppsatsen grundar sig i en simuleringsstudie har antalet stickprov kunnat väljas relativt fritt. I denna studie simuleras 1000 stickprov för varje stickprovsstorlek, med andra ord görs 1000 iterationer av en stickprovsstorlek i själva simuleringen. Att arbeta med simulerad data kan underlätta arbetsprocessen stort för den som undersöker skattningsmetoder m.m. Kostnaden för modifikationer av data är mindre för fiktiv data än för reell insamlad data - om man snabbt behöver mer underlag för en resultatdel och analys av denna, så kan man tillgodogöra sig detta relativt enkelt med simulerad data.

Simulerad data har alltså inte samma restriktioner som reell data vilket kan ge en studie möjlighet att visa på signifikanta resultat snarare än bara ett udda exempel. Med det sagt, kan endast generella trender påvisas med simulerad data, då det inte alltid kan framställa den komplexitet som kan återfinnas i reell data.

1.3.3 Utvärdering av simuleringen

I denna studie har data simulerats från två modeller, 3 och 4. Initialt kommer skattningsmetoderna jämföras inom varje modell, för att sedan jämföras mellan modellerna. Parameterskattningarna som fås från metoderna kommer att undersökas i histogram och tabeller som redovisar median, medelvärde, minimum och maximum från varje skattningsmetod och stickprovsstorlek.

Sedan utförs ett så kallat "1-sample t-test" för att undersöka noll-hypotesen att medelvärdet av alla parameterskattningar från en metod är lika med det

sanna värdet för $\beta-$parametern ifråga. Först undersöks det histogram som hör till skattningsmetoden och stickprovsstorleken för att avgöra om ett tvåsidigt "1-sample t-test" ska utföras - testet görs om histogramet antyder en normalfördelning av alla de skattade $\boldsymbol{\beta}-$parametrarna från skattningsmetoden. Även om histogrammen ger en ganska tydlig bild av utfallet så redovisas resultaten ändå för tydlighets skull.

Testet kommer att utvärdera om det föreligger någon signifikant skillnad mellan medelvärdet av alla parameterskattningar som har tagits fram från varje skattningsmetod och stickprovsstorlek. På samma sätt testas även det sanna värdet på den undersökta $\beta-$parametern i modellen från vilken stickproven har simulerats från. Noll-hypotesen förkastas om p-värdet för testet är mindre än den valda signifikansnivån på 0.05. Då antas alternativhypotesen, vilken säger att det föreligger en signifikant skillnad mellan medelvärdet från alla skattade parametervärden och det sanna parametervärdet. Detta är ett lämpligt test eftersom vi har $\beta-$skattningar från cirka 1000 stickprov för varje skattningsmetod och stickprovsstorlek och vi vill undersöka hurivida skillnaden beror på slumpmässiga eller systematiskt effekter.

Den externa validiteten av denna studie är begränsad - studien hade kunnat utföras på flera olika sätt med fler alternativa metoder än $Elrm_R$, $ExLog_{SAS}$ och MLE_R - och med en annan datastruktur. Resultaten är således inte direkt överförbara till alla tänkbara situationer, utan bör initialt bedömas i liknande kontexter som i denna studie. [3]

1.3.4 Avgränsningar

Data simuleras från två olika modeller (se ekv 3 och ekv 4). Detta för att kunna jämföra de skattade $\boldsymbol{\beta}-$parametrarna vi får från de olika skattningsmetoder som undersöks - $ExLog_{SAS}$, A-, B-$Elrm_R$ och MLE_R.

I denna studie gjordes en avgränsning att undersöka fyra olika stickprovsstorlekar - $10, 20, 50$ och 100. Detta för att se hur resultaten från de olika skattningsmetoderna ändras vid analys av små (storlek 10 och 20), medelstora (storlek 50) och stora (storlek 100) stickprov. Efter initiala test bedömdes dessa storlekar som lämpliga för att tydligt påvisa skillnader.

Resultatet från en viss skattningsmetod kan bero på datastrukturen i stickprovet som analyseras. Denna uppsats är avgränsad till att endast undersöka data som har simulerats från två modeller, där den binära responsvariabeln i Modell 1 beror av en binär variabel, och i Modell 2 beror av två binära variabler. Dessa modeller redovisas i avsnitt 1.3.1, och resultaten är således inte direkt överförbara till alla tänkbara situationer, utan bör initialt bedömas i liknande kontexter som i denna studie. Det var ett medvetet val att inte ta med en kontinuerlig förklarande variabel i de undersökta modellerna då tidigare studier har visat på att $ExLog_{SAS}$ inte alltid är applicerbart. Detta beror på att den betingade fördelningen för de tillräckliga statistikorna, vilka behövs för att utföra denna sakttningsmetod (se avsnitt 2.1), ofta degenererar [4], vilket skulle försvåra analysen.

För skattningsmetoderna A- respektive B-$Elrm_R$ genereras Markov-kedjor[1] för att få fram parameterskattningar från ett stickprov. Här finns det en begränsning på hur långa Markov-kedjor som kan göras i ett statistiskt mjukvaruprogram, eftersom det virtuella minnet i datorn är begränsad, denna begränsning diskuteras närmare i diskussionen i denna uppsats.

2 Teori

Detta avsnitt baseras på teori från Agresti (2002) [1], Mehta & Patel (1995)[5] och Hirji (2006) [6].

2.1 Logistisk regression

Logistisk regression är en analysmetod som lämpar sig väl för att förklara relationen mellan en eller flera förklarande variabler och en binär responsvariabel.

Antag att Y är en binär responsvariabel som beror av en eller flera förklarande variabler $X = (X_1, ..., X_n)$ som antar värdena $\boldsymbol{x} = (x_1, x_2, ..., x_n)$.

Låt sedan π_j vara sannolikheten att $Y_j = 1$ och $(1 - \pi_j)$ vara sannolikheten att $Y_j = 0$. Vi har då en logistisk regressionsmodell som förklarar relationen mellan π_j och $\boldsymbol{x}$ som följande

$$log\left(\frac{\pi_j}{1-\pi_j}\right) = \alpha + \boldsymbol{x}'\boldsymbol{\beta} \tag{5}$$

där α och $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ är de okända parametrarna och där uttryckets vänsterled kallas *logit* eller *log-odds* [2].

Notera att sannolikheten för $Y_j = 1$ är en siffra mellan 0 och 1, medan log-transformationen[3] inte har en begränsning

$$0 \leq \pi_j \geq 1, \qquad\qquad -\infty < log\left(\frac{\pi_j}{1-\pi_j}\right) < \infty.$$

Ett tillvägagångasätt för att skatta $\alpha-$ och $\boldsymbol{\beta}-$parametrarna är att maximera *ekv.* 5 med avseende på dessa parametrar. För att maximera log-Likelihood-funktionen kan man derivera ekvationen med avseende på de okända parametrarna och sedan sätta de partiella derivatorna till noll, vilket ger ett ekvationssystem. Parameterskattningarna som tas fram genom denna metod kallas Maximum-Likelihood-skattningar och antas vara konsistenta, asymptotiskt effektiva och asymptotiskt normalfördelade. Asymptotiska statistikor, som exempelvis Likelihood-kvot, Score-test och Wald-statistikor, kan användas för att utföra diverse hypotestester. För små stickprov har studier visat på att dessa MLE-skattningar har en bias vid analys av små stickprov, vilket diskuterades i avsnitt 1.1. Vi redovisar nedan ett alternativ till MLE-skattningar kallad exakt betingad inferens.

[1] förklaras närmre i 2.2.2

[2] För vidare teori om odds, se appendix.

[3] I denna uppsats använder vi oss av den naturliga logaritmen.

Betingad exakt inferens grundar sig i att ta fram den betingade fördelningen för de parametrar man är intresserad av genom att betinga på de så kallade skräp-parametrarna [4]. Denna fördelning kallas för exakt betingad fördelning.

Låt oss anta att vi är intresserade av att skatta $\boldsymbol{\beta}$ och låter α vara en skräp-parameter. Istället för att skatta α från ovanstående obetingade Likelihood-funktion 5, kan vi eliminera α genom att betinga på det observerade värdet av dess tillräckliga statistika $\boldsymbol{m} = \sum_j^n y_j$. Den betingade Likelihoodfunktionen fås då som

$$Pr(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n | m) = \frac{exp(\sum_{j=1}^n y_j \boldsymbol{x}_j' \boldsymbol{\beta})}{\sum_R (exp \sum_{j=1}^n y_j \boldsymbol{x}_j' \boldsymbol{\beta})} \tag{6}$$

där den yttre summeringen i nämnaren i *ekv.* 6 är över mängden

$$R = \left\{ (y_i, y_2, ..., y_n) : \sum_{j=1}^n y_j = m \right\}.$$

Från den betingade Likelihood-funktionen har vi att vektorn av effektiva statistikor för $\boldsymbol{\beta}$ är

$$\boldsymbol{t} = \sum_{j=1}^n y_j \boldsymbol{x} \tag{7}$$

och har fördelning

$$Pr(T_1 = t_1, T_2 = t_2, ..., T_p = t_p) = \frac{c(\boldsymbol{t}) e^{\boldsymbol{t}' \boldsymbol{\beta}}}{\sum_{\boldsymbol{u}} c(\boldsymbol{u}) e^{\boldsymbol{u}' \boldsymbol{\beta}}} \tag{8}$$

där

$$c(\boldsymbol{t}) = |S(\boldsymbol{t})|$$

$$S(\boldsymbol{t}) = \left\{ (y_1, y_2, ..., y_n) : \sum_{j=1}^n y_j = m, \sum_{j=1}^n y_j x_{ij} = t_i, i = 1, 2, ...p \right\}$$

och där $|S|$ betecknar antalet distinkta element i mängden S, och summan i nämnaren över alla $\boldsymbol{u}$ för vilka $c(\boldsymbol{u}) \leq 1$. Alltså, $c(\boldsymbol{t})$ är antalet binära sekvenser på formen $(y_1, y_2, ..., y_n)$ sådana att $\sum_j y_j x_{ij} = t_i$, för $i = 1, 2, ...p$.

Exakt slutledning om $\boldsymbol{\beta}$ kräver beräkning av koefficienter såsom $c(\boldsymbol{t})$, där vissa av de tillräckliga statistikorna är fixa vid sina observerade värden och där

[4] En skräp-parameter (nuisance parameter) är en parameter som inte är av omedelbart intresse, men som vi måste ta hänsyn till i analysen av de parametrar som är av intresse.

andra varierar över deras tillåtna intervall. Detta blir fort beräkningsmässigt tungt och är därför en metod som man kan applicera på väldigt enkla exempel eller med hjälp av algoritmer implementerade i statistiska mjukvaruprogram. Skattningsmetoderna som undersöks i denna studie där exakt betingad inferens implementeras förklaras i avsnitt 2.2.1 och 2.2.2.

Ett förenklat exempel finns med i appendix, se avsnitt 6.2, för att förtydliga hur denna fördelning tas fram.

2.2 Mjukvarupaket som används i uppsatsen

Två olika statistiska mjukvaruprogram har använts för att applicera de undersökta skattningsmetoderna $ExLog_{SAS}$, $A-$, $B-Elrm_R$ och MLE_R - SAS och R. Den data som har analyserats har simulerats från R (för en närmare beskrivning av datasimulering se avsnitt 1.3.1). Skattningsmetoden $ExLog$ har utförts i SAS medan $A-$ & $B-Elrm$ och MLE har utförts i R. För vidare fördjupning om jämförelser av olika metoder för exakt slutledning i olika statistiska mjukvaruprogram läses med fördel bl.a. Oster (2002) [7] och Oster (2003) [8].

2.2.1 ExLog i SAS

I mjukvaruprogrammet SAS kan man utföra ExLog för beroende variabler av dikotom karaktär, en metod som ingår i den s.k. $LOGISTIC$-proceduren[5].

Skattningsmetoden undersöker sannolikheten att få responsvektorn i stickprovet, med avseende på alla 2^n möjliga responsvektorer, där n är antalet observationer i stickprovet [7]. Detta tillvägagångssätt blir beräkningsmässigt tungt när n ökar. Exempelvis fås 2^{30} möjliga responsvektorer vid analys av 30 observationer - alltså över en miljard möjliga responsvektorer.

En algoritm, utvecklad av Hirji, Mehta och Patel, kallad "The Multivariate shift algorithm", tillsammans med en nätverksalgoritm, beskriven av Mehta, Patel och Senchaudun (1992) är implementerad i proceduren för att snabbt generera och räkna möjliga responsvektorer.

"The Multivariate shift algoritm" är en metod som tar fram och räknar möjliga $\boldsymbol{y}$-vektorer för större problem. Den baserar sig på en rekursiv relation mellan den tillräckliga statistikan och möjliga y-vektorer. Stegen av denna algoritm illustreras i avsnitt 6.2 i appendix. Nätverksalgoritmen bygger kopplingar för varje parameter som betingas bort i ekv 6 för att på så sätt indentifiera och reducera antalet möjliga $\boldsymbol{y}-$vektorer. Detta möjliggör att dra en exakt betingad inferens och ta fram $\boldsymbol{\beta}-$skattningar med denna metod. Tillsammans med ett exempel illustreras "The Multivariate shift algorithm" i figur 9 i avsnitt 6.2. Derr (2009) [9] beskriver ytterligare fördjupat hur denna skattningsmetod fungerar.

[5] LOGISTIC-proceduren i SAS anpassar linjära logistiska regressionsmodeller för diskret responsdata.

2.2.2 Elrm i R

För att uföra exakt logistisk regression i R används funktionen $Elrm_R$.

Funktionen implementerar en modifikation av "Markov-kedja Monte Carlo", förkortat MCMC, framtagen av Forster (2003) för att approximera exakt betingad slutledning för logistiska regressionsmodeller. MCMC tar fram en beroende sekvens av möjliga tillräckliga statistikor för β-parametern av intresse. Slutledningen grundar sig i fördelningen av de tillräckliga statistikorna för parametrarna av intresse betingat på de kvarvarande skräp-parametrarna. [10, 11, 12]

I denna studie används två alternativ av denna skattningsmetod, vilka skiljer sig på två sätt. För alternativ A har standardvärden för skattningsmetoden $Elrm_R$ valts ut. I alternativ B utförs fler Markov-kedje-iterationer som rör sig med mindre och fler tagna steg. Allmänt för denna metod gäller att värdet för hur Markov-kedjan rör sig, r, måste förhålla sig till restriktionen att det måste vara ett jämnt heltal som är mindre eller lika med längden av responsvektorn. I denna rapport kommer r att hållas konstant för de två skattningsmetoderna A- och $B - Elrm_R$. Zamar, McNeney & Graham (2007) [10] presenterar en mer fördjupad bild av A- och $B - Elrm_R$.

2.2.3 MLE i R

Skattningsmetoden för så kallad vanlig logistisk regression utförs i R av en funktion kallad glm. Denna är utformad för att utföra generaliserade linjära modeller på bl.a. binär data, den typ av data som simuleras i denna studie. Likt de ovannämnda skattningsmetoderna specificeras olika funktionsparametrar där parametern "family" är en beskrivning av felfördelningen och länkfunktionen som ska användas vid skattning av de intressanta parametrarna. I denna studie sätts denna lika med binomial, vilket talar om för R att utföra logistisk regression och skatta parametrarna med MLE_R.

3 Resultat och Analys

I simuleringen har olika skattningsmetoder - $ExLog_{SAS}$, $A-$, $B - Elrm_R$ och MLE_R, använts för att skatta $\boldsymbol{\beta}-$parametrar från två logistiska regressionsmodeller vid analys av stickprov i olika storlekar - $10, 20, 50$ och 100. För varje stickprovsstorlek har 1000 stickprov simulerats för vilka $\boldsymbol{\beta}-$skattningarna har varit av intresse i denna studie. Båda modellerna beskriver relationen mellan en binär responsvariabel, och en respektive två förklarande variabler av samma karaktär.

Avsnittet inleds med en presentation och analys av resultat från

$$\text{Modell 1} \qquad log\left(\frac{\pi_i}{1-\pi_1}\right) = \alpha + \beta_1 x_1, \tag{9}$$

där det sanna värdet för α respektive β_1 är -1 och 2, och där $X_1 \sim Be(1/2)$.

Sedan presenteras och analyseras resultaten från

$$\text{Modell 2} \qquad log\left(\frac{\pi_i}{1-\pi_1}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2, \tag{10}$$

där det sanna värdet för α, β_1 respektive β_2 är $0.2, 1.5$ och -1, och där $X_j \sim Be(1/2)$, för $j = 1, 2$. $\boldsymbol{\beta}-$skattningarna undersökas för att se om de är väntevärdesriktiga, dvs. om medelvärdet av alla skattningar för $\boldsymbol{\beta}-$skattningarna är ungefär lika med de sanna parametervärdena. Avsnittet avslutas med en jämförelse av de olika modellerna.

3.1 Systematisk skillnad mellan medelvärden och sanna parametrar

För att utvärdera om det föreligger någon signifikant skillnad mellan medelvärdet av alla $\boldsymbol{\beta}-$skattningarna framtagna vid analys av 1000 stickprov med de sanna parametervärdena i Modell 1 och Modell 2, utförs ett så kallat "1-sample t-test". I testet ställs noll-hypotesen - att medelvärdet av alla $\boldsymbol{\beta}-$skattningarna är lika med de sanna värdena, mot den tvåsidiga alternativhypotesen - att medelvärdet av alla $\boldsymbol{\beta}-$skattningarna inte är lika med de sanna parametervärdena. Noll-hypotesen förkastas vid signifikansnivå 0.05.

I de histogram där det tydligt kan utläsas att $\boldsymbol{\beta}-$skattningarna inte kan antas vara normalfördelade utförs testet endast för de övriga skattningsmetoderna. P-värdena från respektive test redovisas i tabeller under respektive modell och stickprovsstorlek, tillsammans med ett konstaterande huruvida noll-hypotesen förkastas eller inte.

3.2 Modell 1

I detta avsnitt presenteras och analyseras β_1-skattningar framtagna av de undersökta skattningsmetoderna - $ExLog_{SAS}$, $A-$, $B - Elrm_R$ och MLE_R, vid analys av data simulerad från Modell 1 (se 9). I denna simuleringsstudie har skattningsmetoderna skattat β_1 från analys av 1000 stickprov av storlek $10, 20, 50$ respektive 100, skattningarna har sedan jämförts med det sanna $\beta-$värdet 2 i denna modell. Nedan presenteras resultat och analys av skattningsmetoderna i stigande ordning av stickprovsstorlek.

3.2.1 Modell 1, stickprovsstorlek 10

Tabell 1 sammanfattar skattningsmetodernas β_1-skattningar.

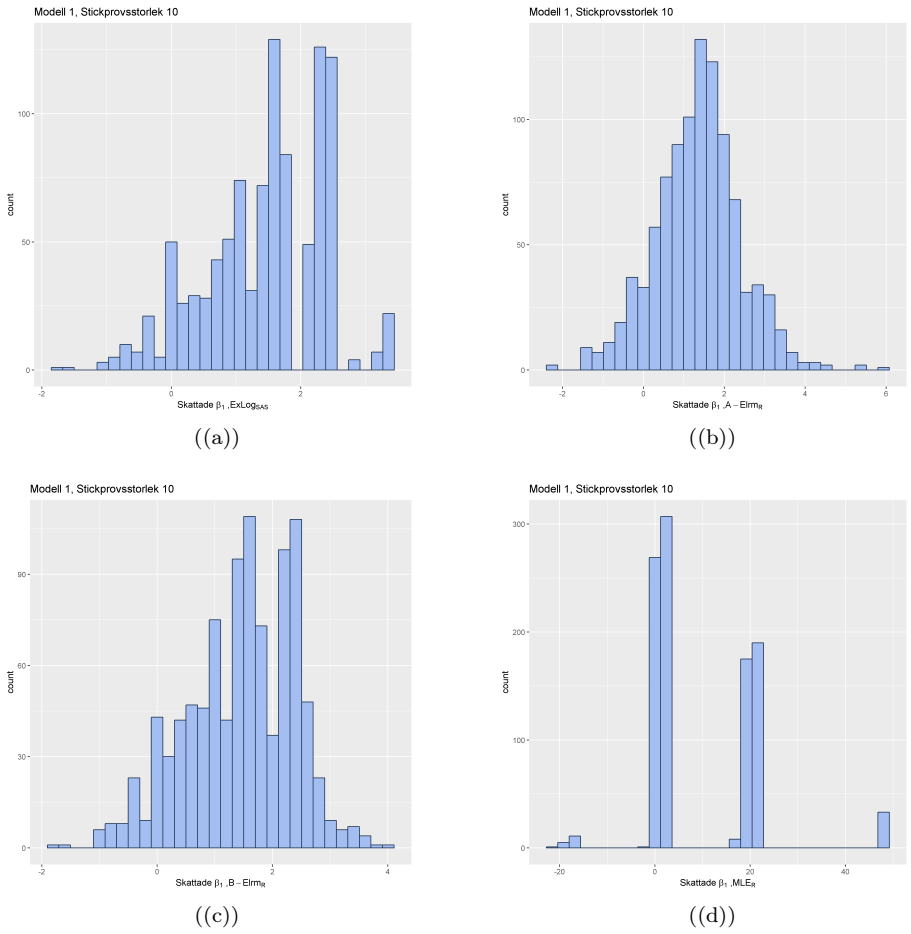
Modell 1, Stickprovsstorlek 10

	Min	1a Kvartil	Median	Medelvärde	3e Kvartil	Max
$ExLog_{SAS}$	-1.7714	0.8814	1.5930	1.4484	2.3316	3.3540
$A - Elrm_R$	-2.3628	0.6861	1.3821	1.3415	1.9667	5.8362
$B - Elrm_R$	-1.8057	0.8634	1.5382	1.4521	2.2137	4.0060
MLE_R	-20.482	1.099	2.708	9.700	20.259	49.132

Table 1: Sammanfattning av β_1-skattningar från 1000 stickprov som analyserats med respektive skattningsmetod vid stickprovsstorlek 10.

Av tabell 1 kan man utläsa att medelvärdet av β_1-skattningarna från metoderna $ExLog_{SAS}$, A- och $B - Elrm_R$ ligger under det sanna parametervärdet. Detta medan medelvärdet från MLE_R ligger långt ovanför det sanna parametervärdet, som är lika med 2. Ingen av de undersökta skattningsmetoderna tycks skatta väntevärdesriktiga parametrar.

I nedanstående figurer visar histogram fördelningen över β_1-skattningarna från de fyra olika skattningsmetoderna vid analys av 1000 stickprov.



((a)) ((b))

((c)) ((d))

Notera att skalan på axlarna skiljer sig åt mellan graferna.

Figure 1: Histogram över β_1-skattningarna från skattningsmetod ((a)) $ExLog_{SAS}$, ((b)) $A - Elrm_R$, ((c)) $B - Elrm_R$ och ((d)) MLE_R, vid stickprovsstorlek 10.

Av figur 1(a) - 1(c) ovan noteras att fördelningen över β_1-skattningarna efterliknar en normalfördelning, där A- och $B - Elrm_R$ visar på en mer jämn fördelning än $ExLog_{SAS}$. Dessa tre figurer visar också på en skevhet i fördelningen - skattningsmetoderna $Exlog_{SAS}$, A- och $B - Elrm_R$ tycks systematiskt skatta β_1 lägre än det sanna parametervärdet 2 i modellen från vilken datan är simulerad.

Från figur 1(d) noteras att β_1-skattningarna framtagna av metoden MLE_R kan delas in i fyra intervall, antingen fås en β_1-skattning som är

- < -10,

- $\in (-10, 10)$,
- $\in (10, 30)$,
- > 40.

De stickprov som är framtagna från Modell 1 (se ekv. 9) har fyra möjliga observationskombinationer, där (y, x_1) kan anta värdena $(0,0), (1,0), (0,1)$ eller $(1,1)$. Vid vidare analys av β_1-skattningarna framtagna av MLE_R visade det sig att om en eller flera observationskombinationer saknades i stickprovet skattades β_1-parametern till ett värde utanför det intervall där det sanna parametervärdet ligger. Denna undersökning sammanfattas i tabellen nedan.

Intervall		Saknade observationskombinationer	Antal stickprov
	$<$-10	(0,0) eller (1,1)	17
$>$-10,	$<$ 10	-	577
$>$ 10,	$<$ 30	(1,0) eller (0,1)	373
$>$ 40		(1,0) och (0,1)	33

Table 2: Modell 1, Intervall för β_1-skattningar framtagna av MLE_R vid stickprovsstorlek 10.

Av figurerna och tabellerna ovan dras slutsatsen att vid analys av stickprov av storlek 10 simulerad från Modell 1 skattas β_1-parametern systematiskt lägre än det sanna parametervärdet av skattningsmetod $ExLog_{SAS}$, A- och $B - Elrm_R$. Skattningsmetoden MLE_R påverkades av vilka observationskombinationer stickprovet innehöll och gav i fall där observationskombinationer saknades β_1-skattningar som låg långt ifrån det sanna parametervärdet. Studien påvisade resultat att de undersökta metoderna inte är lämpliga för analys av stickprov av storlek 10 simulerad från Modell 1.

3.2.2 Modell 1, stickprovsstorlek 20

I tabell 3 presenteras en kort sammanfattning av skattningsmetodernas β_1-skattningar vid stickprovsstorlek 20.

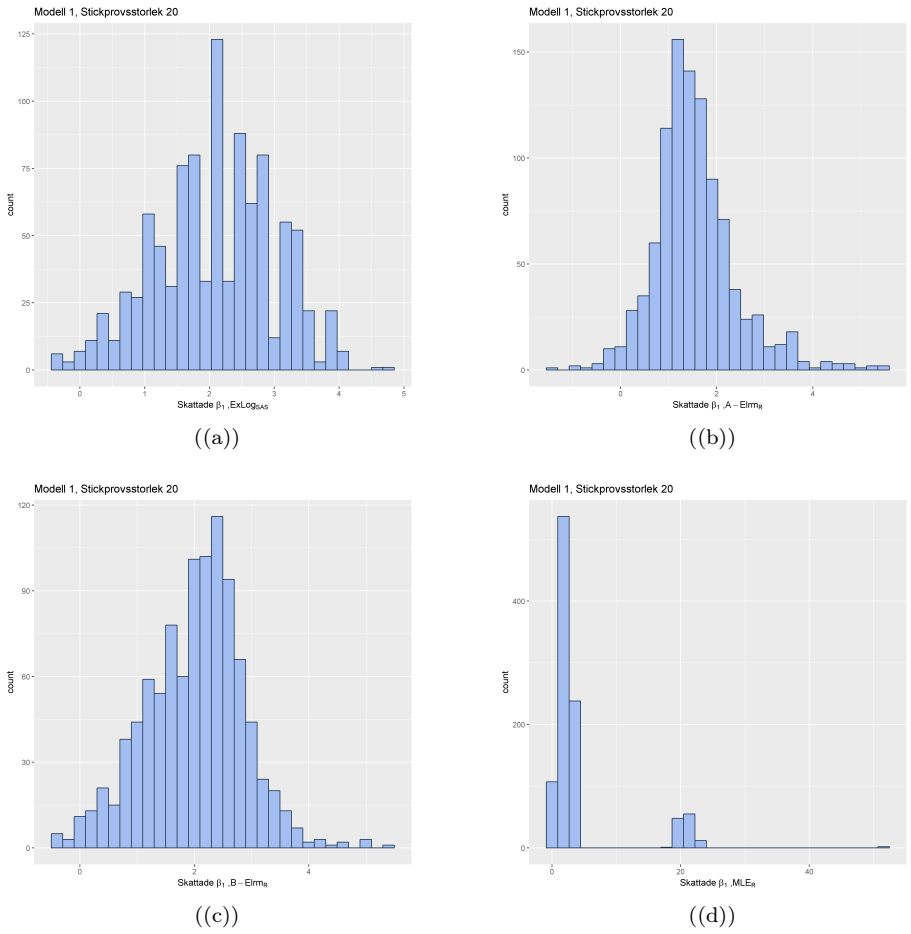
Modell 1, Stickprovsstorlek 20

	Min	1a Kvartil	Median	Medelvärde	3e Kvartil	Max
$ExLog_{SAS}$	-0.3855	1.5089	2.0985	2.0756	2.7773	4.7300
$A - Elrm_R$	-1.509	1.049	1.444	1.551	1.950	5.388
$B - Elrm_R$	-0.4295	1.4460	2.0916	2.0071	2.5367	5.3616
MLE_R	-0.4055	1.5992	2.2336	4.3387	3.0445	51.1321

Table 3: Sammanfattning av β_1-skattningar från 1000 stickprov som analyserats med respektive metod vid stickprovsstorlek 20.

I tabell 3 noteras att medelvärdet av β_1-skattningarna från metoderna $ExLog_{SAS}$ och $B - Elrm_R$ ligger nära det sanna parametervärdet 2 och kan alltså antas vara väntevärdesriktiga. Differensen mellan medelvärdet av alla β_1-skattningar från $A-Elrm_R$ har minskat vid underökning av stickprovsstorlek 20 i jämförelse med den tidigare undersökta stickprovsstorleken, men medelvärdet ligger fortfarande under det sanna parametervärdet. Medelvärdet från MLE_R ligger ovanför det sanna parametervärdet och vi noterar att den maximala β_1-skattningen för denna metod ligger på ett värde precis över 51 - långt ifrån det sanna parametervärdet 2. De framtagna β_1-skattningarna för metoderna $A - Elrm_R$ och MLE_R kan inte anses vara väntevärdesriktiga.

I nedanstående figurer visar histogram fördelningen över β_1-skattningarna från de olika skattningsmetoderna vid analys av 1000 stickprov.



((a)) ((b))

((c)) ((d))

Notera att skalan på axlarna skiljer sig åt mellan graferna.

Figure 2: Histogram över β_1-skattningarna från skattningsmetod ((a)) $ExLog_{SAS}$, ((b)) $A - Elrm_R$, ((c)) $B - Elrm_R$ och ((d)) MLE_R, vid stickprovsstorlek 20.

Av figur 2(a) - 2(c) ovan noteras att fördelningen över β_1-skattningarna kan liknas vid en normalfördelning, där 2(b) indikerar en skevhet i skattningarna från $A - Elrm_R$. Detta är ett resultat som påminner om det från stickprovsstorlek 10.

Från histogrammet över β_1-skattningarna framtagna av MLE_R, se figur 2(d), noteras att skattningarna kan delas in i tre intervall enligt följande

- < 10,
- $\in (10, 30)$,

- > 40.

De stickprov som är framtagna från Modell 1 (se ekv. 9), har som sagt fyra möjliga observationskombinationer - (y, x_1) kan anta värdena $(0,0), (1,0), (0,1)$ eller $(1,1)$. Av noggrann undersökning av β_1-skattningarna framtagna av MLE_R visade det sig att om en eller flera observationskombinationer saknades i stickprovet så skattades β_1-parametern till ett värde utanför det intervall där det sanna parametervärdet ligger. Denna undersökning sammanfattas i tabellen nedan.

Intervall	Saknade observationskombinationer	Antal stickprov
<10	-	882
>10, < 30	(1,0) eller (0,1)	116
> 40	(1,0) och (0,1)	2

Table 4: Modell 1, Intervall för β_1-skattningar framtagna av MLE_R vid stickprovsstorlek 20.

Av resultaten ovan dras slutsatsen att vid analys av stickprov av storlek 20 simulerad från Modell 1 är $ExLog_{SAS}$ och $B - Elrm_R$ lämpliga skattningsmetoder. Metoderna ger väntevärdesriktiga skattningar som är normalfördelade kring det sanna parametervärdet.

Liknas resultaten från stickprovsstorlek 10, skattades även här β_1-parametrarna av $A - Elrm_R$ systematiskt lägre än det sanna värdet och MLE_R påverkades av vilka observationskombinationer stickprovet innehöll.

3.2.3 Modell 1, stickprovsstorlek 50

Nedan undersöks tabell 5 som sammanfattar metodernas β_1-skattningar vid stickprovsstorlek 50.

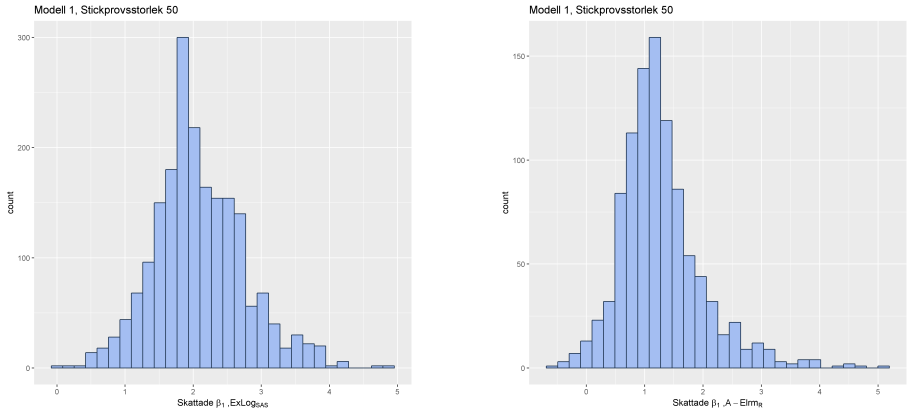
Modell 1, Stickprovsstorlek 50

	Min	1a Kvartil	Median	Medelvärde	3e Kvartil	Max
$ExLog_{SAS}$	0.0306	1.6601	2.0414	2.0898	2.4735	4.8988
$A - Elrm_R$	-0.5624	0.8346	1.1599	1.2794	1.5843	5.1197
$B - Elrm_R$	0.1056	1.3969	1.7327	1.8060	2.1446	4.5373
MLE_R	0.1907	1.5870	2.0794	2.1046	2.5390	4.8771

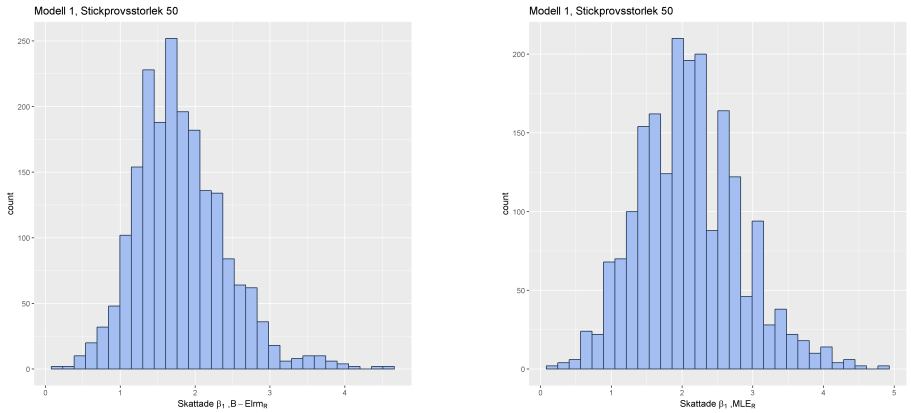
Table 5: Sammanfattning av β_1-skattningar från 1000 stickprov som analyserats med respektive metod vid stickprovsstorlek 50.

Av tabell 5 ser vi att medelvärdet av alla β_1-skattningarna från $ExLog_{SAS}$ och MLE_R ligger nära det sanna parametervärdet - metoderna skattar väntevärdesriktiga parametrar, medan det för A- och $B - Elrm_R$ ligger under det sanna värdet - metoderna skattar inte väntevärdesriktiga parametrar.

I nedanstående figurer visar histogram fördelningen över β_1-skattningarna från de olika skattningsmetoderna vid analys av 1000 stickprov.



((a)) ((b))



((c)) ((d))

Notera att skalan på axlarna skiljer sig åt mellan graferna.

Figure 3: Histogram över β_1-skattningarna från skattningsmetod ((a)) $ExLog_{SAS}$, ((b)) $A - Elrm_R$, ((c)) $B - Elrm_R$ och ((d)) MLE_R, vid stickprovsstorlek 50.

Av figur 3(a) - 3(d) ovan noteras att fördelningen över β_1-skattningarna kan liknas vid en normalfördelning, där figur 3(b) indikerar en skevhet i skattningarna från $A - Elrm_R$. Återigen fås resultat som ligger i linje med simuleringarna för andra stickprovsstorlekar, specifikt 10 och 20. Figur 3(c) visar på en tendens till skevhet i skattningarna från $B - Elrm_R$.

Vid analys av stickprovsstorlek 50 simulerad från Modell 1 anses $ExLog_{SAS}$

och MLE_R vara lämpade skattningsmetoder. Metoderna ger väntevärdesriktiga skattningar som är ungefärligt normalfördelade kring den sanna parametern i Modell 1. A- och $B - Elrm_R$ skattar β_1-parametern systematiskt lägre än det sanna värdet, där tendenserna för skevhet är större hos $A - Elrm_R$ än $B - Elrm_R$. Dessa metoder anses inte vara lämpliga alternativ vid analys av data analyserad i detta avsnitt.

3.2.4 Modell 1, stickprovsstorlek 100

Tabell 6 presenterar en kort sammanfattning av metodernas β_1-skattningar vid stickprovsstorlek 100.

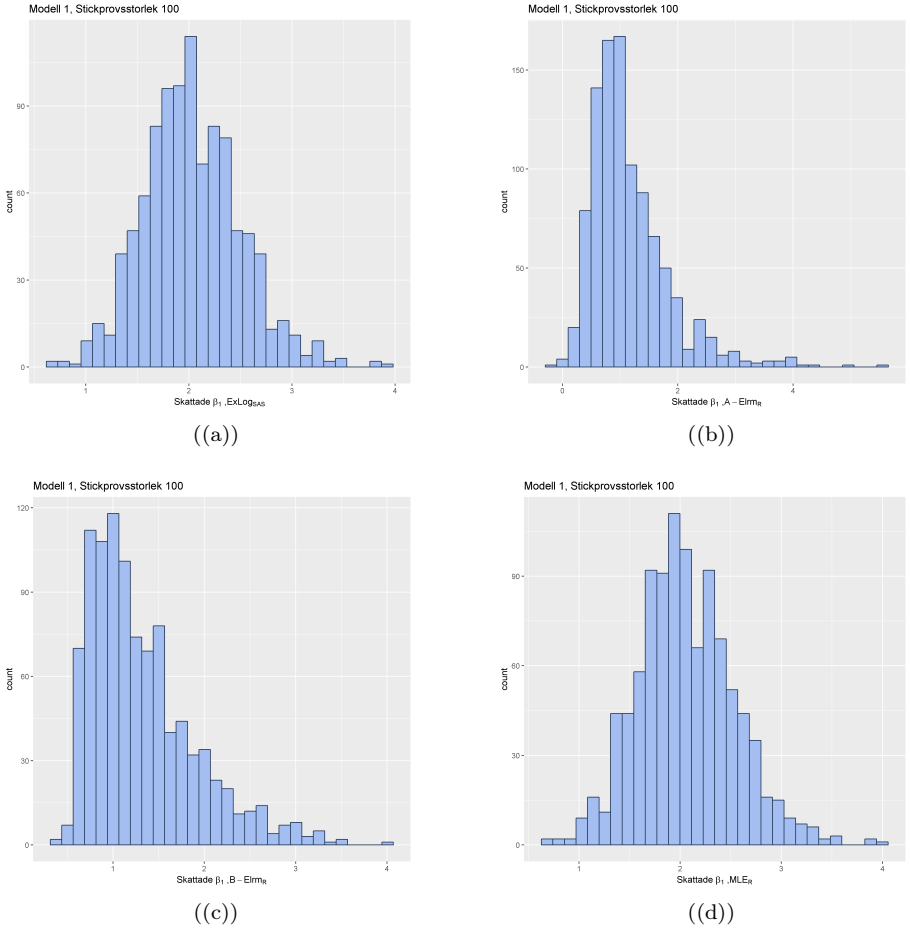
Modell 1, Stickprovsstorlek 100

	Min	1a Kvartil	Median	Medelvärde	3e Kvartil	Max
$ExLog_{SAS}$	0.640	1.703	1.991	2.036	2.359	3.892
$A - Elrm_R$	-0.2486	0.6976	1.0002	1.1520	1.4326	5.4988
$B - Elrm_R$	0.4202	0.8750	1.1619	1.3161	1.5978	4.0507
MLE_R	0.6466	1.7223	2.0149	2.0605	2.3889	3.9587

Table 6: Sammanfattning av β_1-skattningar från 1000 stickprov som analyserats med respektive metod vid stickprovsstorlek 100.

Av tabell 6 ser vi att medelvärdet av β_1-skattningarna från $ExLog_{SAS}$ och MLE_R ligger nära det sanna parametervärdet. Medelvärdena för A- och $B - Elrm_R$ ligger för denna stickprovsstorlek under det sanna värdet, där diskrepansen har ökat i och med att stickprovsstorleken har ökat.

I nedanstående figurer visar histogram fördelningen över β_1-skattningarna från de olika skattningsmetoderna vid analys av 1000 stickprov.



Notera att skalan på axlarna skiljer sig åt mellan graferna.

Figure 4: Histogram över β_1-skattningarna från skattningsmetod ((a)) $ExLog_{SAS}$, ((b)) $A - Elrm_R$, ((c)) $B - Elrm_R$ och ((d)) MLE_R, vid stickprovsstorlek 100.

Av figur 4(a) och 4(d) noteras att fördelningen över β_1-skattningarna från $ExLog_{SAS}$ och MLE_R påminner tydligt om en normalfördelning. Från figur 4(b) och 4(c) ser vi att β_1-skattningarna från A- och $B - Elrm_R$ är skevt fördelade kring det sanna parametervärdet 2.

Vid analys av stickprovsstorlek 100 av data simulerad från Modell 1, tyder resultaten på att $ExLog_{SAS}$ och MLE_R är lämpliga skattningsmetoder. Metoderna ger väntevärdesriktiga skattningar - inget systematiskt fel i skattningarna kunde upptäckas från simuleringen. A- och $B - Elrm_R$ skattar β_1-parametern systematiskt lägre än det sanna värdet och resultaten från stu-

dien talar alltså för att dessa skattningmetoder inte är lämpade för analys av den datastruktur och stickprovsstorlek som har analyserats i detta avsnitt, se Modell 1 (9).

3.2.5 Modell 1 - Systematisk skillnad mellan medelvärden och sanna parametrar

Modell 1, Systematisk skillnad mellan medelvärden och sanna parametrar

Metod	p-värde	Förkasta/Förkasta inte	Stickprovsstorlek
$ExLog_{SAS}$	$< 2.2e-16$	Förkasta H_0	10
$A-Elrm_R$	$< 2.2e-16$	Förkasta H_0	10
$B-Elrm_R$	$< 2.2e-16$	Förkasta H_0	10
$ExLog_{SAS}$	0.009396	Förkasta H_0	20
$A-Elrm_R$	$< 2.2e-16$	Förkasta H_0	20
$B-Elrm_R$	0.7942	Förkasta inte H_0	20
$ExLog_{SAS}$	$2.747e-05$	Förkasta H_0	50
$A-Elrm_R$	$< 2.2e-16$	Förkasta H_0	50
$B-Elrm_R$	$< 2.2e-16$	Förkasta H_0	50
MLE_R	$4.825e-06$	Förkasta H_0	50
$ExLog_{SAS}$	0.01772	Förkasta H_0	100
$A-Elrm_R$	$< 2.2e-16$	Förkasta H_0	100
$B-Elrm_R$	$< 2.2e-16$	Förkasta H_0	100
MLE_R	$7.435e-05$	Förkasta H_0	100

Table 7: P-värden från "1-sample t-test", stickprovsstorlek 10, 20, 50 & 100.

Av tabell 7 noteras att nollhypotesen inte kan förkastas på signifikansnivån 0.05 för skattninngsmetoden $B-Elrm_R$ vid stickprovsstorlek 20, detta var alltså den enda metod och stickprovsstorlek där en signifikant skillnad mellan medelvärdet av alla β_1-skattningar och det sanna parametervärdet inte kunde påvisas. För övriga metoder och stickprovsstorlekar förkastas nollhypotesen och påvisar att skillnaden inte är slumpmässig utan snarare systematisk.

3.3 Modell 2

I detta avsnitt presenteras och analyseras $\boldsymbol{\beta}-$skattningarna, $\boldsymbol{\beta} = \beta_1, \beta_2$, skattade av metoderna - $ExLog_{SAS}$, $A-$, B-$Elrm_R$ och MLE_R, vid analys av data simulerad från Modell 2, se ekv. 10 ovan. I denna simuleringsstudie har skattningsmetoderna skattat β_1 och β_2 från analys av 1000 stickprov av storlek $10, 20, 50$ respektive 100. Skattningarna har sedan jämförts med de sanna $\boldsymbol{\beta}-$värdena i denna modell, $\beta_1 = 1.5$ och $\beta_2 = -1$. Nedan presenteras resultat och analys av skattningsmetoderna i stickprovsstorleksordning.

3.3.1 Modell 2, stickprovsstorlek 10

Vi inleder detta avsnitt med att undersöka tabell 8 som presenterar en sammanfattning av skattningsmetodernas $\boldsymbol{\beta}-$skattningar.

Modell 2, Stickprovsstorlek 10

	Min	1a Kvartil	Median	Medelvärde	3e Kvartil	Max
$ExLog_{SAS}$						
$\hat{\beta}_1$	-2.8134	-0.5180	0.3889	0.3037	1.1584	2.9202
$\hat{\beta}_2$	-2.7060	-0.8959	0.0000	-0.0509	0.7949	2.8225
$A - Elrm_R$						
$\hat{\beta}_1$	-2.5331	0.2146	0.8876	0.8297	1.4674	4.6564
$\hat{\beta}_2$	-4.0970	-1.3013	-0.6573	-0.5938	0.1313	2.3823
$B - Elrm_R$						
$\hat{\beta}_1$	-2.1912	0.2299	0.8624	0.7961	1.4340	2.9874
$\hat{\beta}_2$	-3.18261	-1.24979	-0.59470	-0.54287	0.07572	2.41241
MLE_R						
$\hat{\beta}_1$	-48.5111	0.4399	2.7403	12.2010	20.2592	49.6483
$\hat{\beta}_2$	-49.970	-19.702	-1.419	-8.572	0.000	48.742

Table 8: Sammanfattning av $\boldsymbol{\beta}-$skattningar från 1000 stickprov som analyserats med respektive metod vid stickprovsstorlek 10.

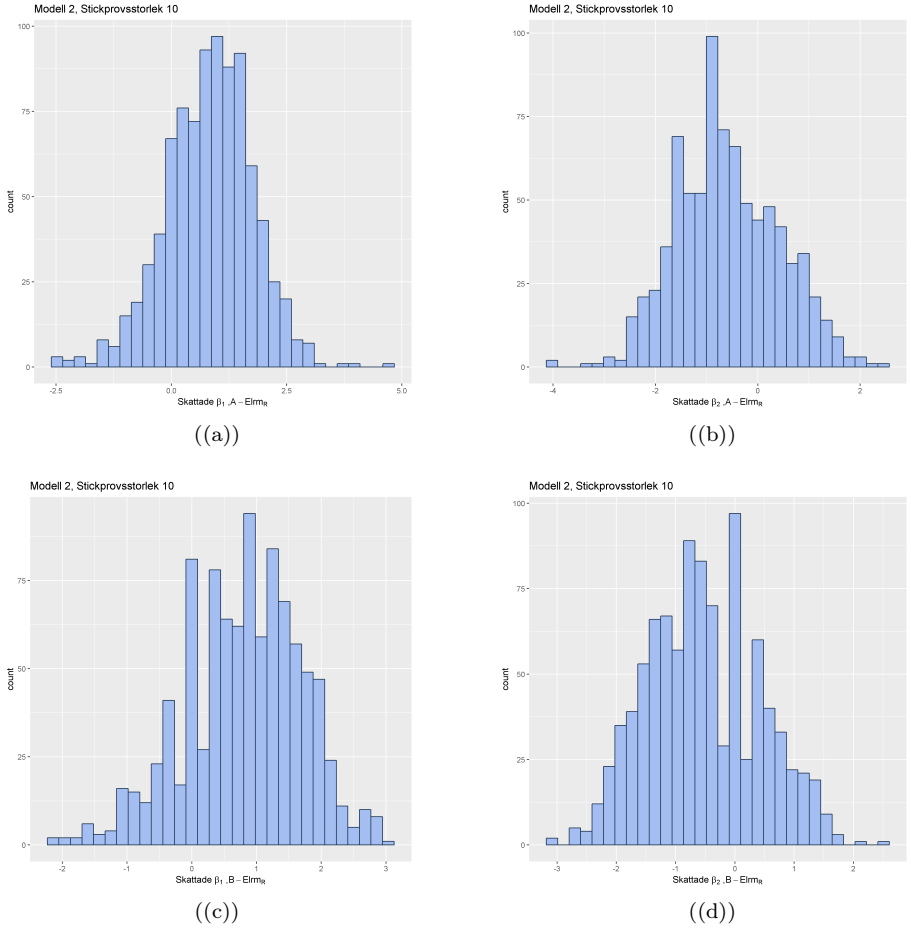
Av tabell 8 kan vi undersöka om skattningarna från de olika skattningsmetoderna är väntevärdesriktiga, dvs. om medelvärdet av alla skattningar av β_1 respektive β_2 är ungefär lika med 1.5 respektive -1. Från tabellen ser vi att medelvärdet av $\boldsymbol{\beta}-$skattningarna inte kan sägas ligga nära de sanna parametervärdena för någon av de undersökta skattningsmetoderna - ingen metod tycks skatta väntevärdesriktiga skattningar.

Gemensamt för alla skatttningsmetoder är att de för vissa stickprov misslyckades ta fram skattningar för $\boldsymbol{\beta}-$parametern. Vid en närmare undersökning av parameterskattningarna visade det sig att flera stickprov var perfekt eller delvis perfekt separererade vilket innebär att en eller flera förklarande variabler förklarar värdet på responsvariabeln perfekt. Exempelvis om vi har ett stickprov där för varje $x_1 = 1$ är $y = 1$. Har vi ett stickprov som är väldigt litet med binära variabler, som i detta fall, är detta vanligt förekommande. Det som händer när vi försöker skatta parametrarna med MLE_R är att MLE-skattningen framför (i exemplet) X_1 inte existerar. Detta resulterade i saknade skattade $\boldsymbol{\beta}-$värden eller i skattningar som var väldigt långt ifrån de sanna värdena med tillhörande stora standardfel.

Av figur 18(c) - 18(d) i appendix, fås liknande resultat för MLE_R som vid analys av data från Modell 1 vid små stickprov, skattningarna lägger sig inom vissa intervall. Vilka intervall beror återigen på vilka observationskombinationer som saknas. Av histogrammen för skattningarna av metoden $ExLog_{SAS}$, se figur

18(a) och 18(a) i appendix, skattar metoden β_1 respektive β_2 systematiskt för lågt respektive för högt. Metoderna $ExLog_{SAS}$ och MLE_R anses därför inte vara lämpliga skattningsmetoder för analys av data med en binär responsvariabel och två förklarande variabler av samma karaktär vid stickprovsstorlek 10.

I nedanstående figurer visar histogram fördelningen över $\boldsymbol{\beta}-$skattningarna från skattningsmetoderna A- och $B - Elrm_R$ vid analys av 1000 stickprov.



((a)) ((b))

((c)) ((d))

Notera att skalan på axlarna skiljer sig åt mellan graferna.

Figure 5: Histogram över $\boldsymbol{\beta}-$skattningarna från skattningsmetod $A - Elrm_R$, figur ((a)) & ((b)), och $B - Elrm_R$, figur ((c)) & ((d)), vid stickprovsstorlek 10.

Av figurerna ovan noteras att fördelningen över $\boldsymbol{\beta}-$skattningarna från A- och $B - Elrm_R$ kan liknas vid en normalfördelning dock inte runt de sanna parametervärdena 1.5 respektive -1. Resultaten från denna studie kan inte

fastställa vilken skattningsmetod som lämpar sig bäst vid analys av data likt den simulerad från Modell 2 vid stickprovstorlek 10. Resultaten visar dock att skattningarna från metoderna A- och $B - Elrm_R$ tycks ha mindre påvisbara systematiska fel i jämförelse med $ExLog_{SAS}$ och MLE_R.

3.3.2 Modell 2, stickprovsstorlek 20

Vid analys av stickprov av storlek 20 simulerade från Modell 2 misslyckades $A - Elrm_R$ att skatta $\boldsymbol{\beta}-$parametrarna. Detta grundar sig i att metoden inte kunde utföra en fullständig betingad inferens då Markov-kedjan var för liten [10]. Denna metod anses därför inte vara lämplig för analys av data simulerad från Modell 2 vid stickprovssorlek 20.

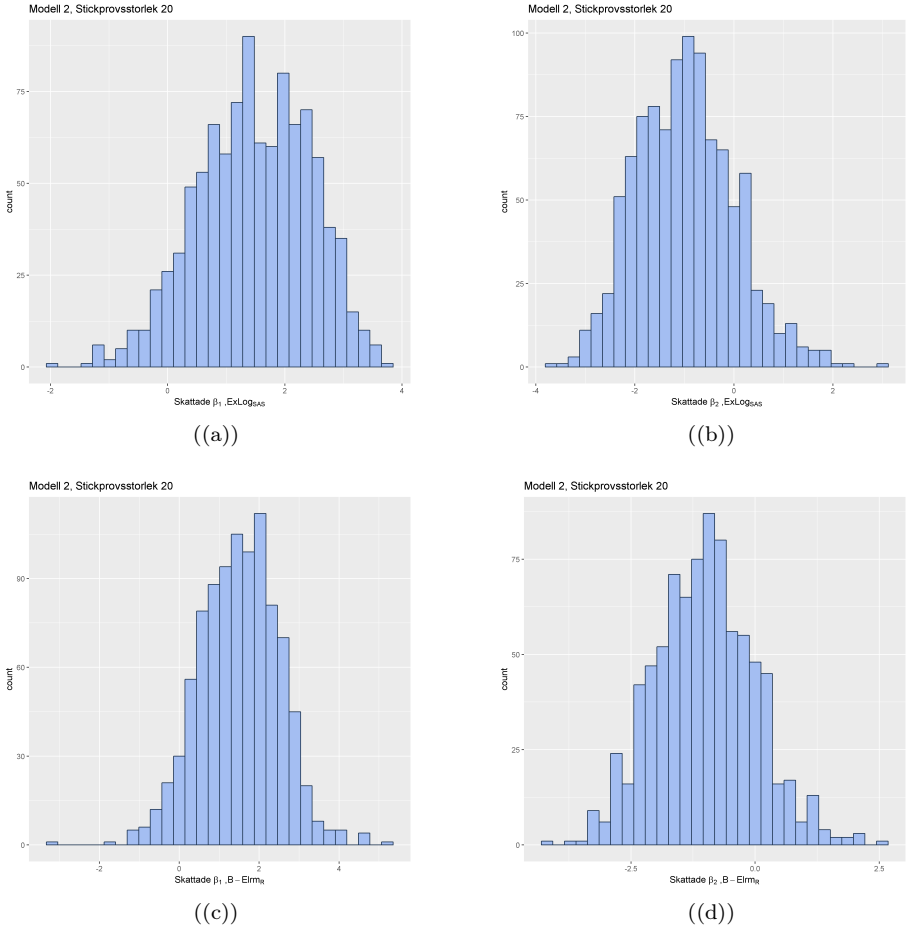
Tabell 9 presenterar en sammanfattning av $\boldsymbol{\beta}-$skattningar från resterande skattningsmetoder - $ExLog_{SAS}$, $B - Elrm_R$ och MLE_R.

Modell 2, Stickprovsstorlek 20

	Min	1a Kvartil	Median	Medelvärde	3e Kvartil	Max
$ExLog_{SAS}$						
$\hat{\beta}_1$	-2.0698	0.7946	1.4785	1.4765	2.2250	3.6506
$\hat{\beta}_2$	-3.7429	-1.7004	-0.9865	-0.9757	-0.3380	2.9526
$B - Elrm_R$						
$\hat{\beta}_1$	-3.2004	0.7978	1.5152	1.5000	2.1729	5.1911
$\hat{\beta}_2$	-4.2125	-1.7122	-1.0093	-1.0165	-0.3489	2.5365
MLE_R						
$\hat{\beta}_1$	-2.3110	0.8905	1.7108	4.7315	2.7181	50.1387
$\hat{\beta}_2$	-50.2835	-2.0796	-1.1484	-3.0200	-0.3804	39.5524

Table 9: Sammanfattning av $\boldsymbol{\beta}-$skattningar från 1000 stickprov som analyserats med respektive metod vid stickprovsstorlek 20.

Av tabell 9 ser vi att skattningarna som är gjorda med $ExLog_{SAS}$ och $B - Erlm_R$ tycks vara väntevärdesriktiga, medan skattningarna från MLE_R inte är det. Histogram över skattningarna från MLE_R, se appendix 20(a) och 20(b) visar, likt undersökning av samma stickprovsstorlek i Modell 1, på att skattningarna hamnar i olika intervall. Efter vidare undersökning visades även detta bero på att en eller flera observationskombinar saknades i stickprovet. Vi går därför vidare med att undersöka histogram framtagna för $ExLog_{SAS}$ och $B - Elrm_R$ vid analys av 1000 stickprov.



Notera att skalan på axlarna skiljer sig åt mellan graferna.

Figure 6: Histogram över $\boldsymbol{\beta}-$skattningarna från skattningsmetod $ExLog_{SAS}$, figur ((a)) & ((b)), och $B-Elrm_R$, figur ((c)) & ((d)), vid stickprovsstorlek 20.

Av figurerna ovan noteras att fördelningen över $\boldsymbol{\beta}-$skattningarna från $ExLog_{SAS}$ och $B-Elrm_R$ kan liknas vid en normalfördelning runt de sanna parametervärdena 1.5 respektive -1. Resultaten från simuleringsstudien tyder alltså på att $ExLog_{SAS}$ och $B-Elrm_R$ är lämpliga skattningsmetoder för analays av data likt den simulerad från Modell 2 vid stickprovsstorlek 20.

3.3.3 Modell 2, stickprovsstorlek 50

Likt resultaten från stickprovsstorlek 20 misslyckades $A-Elrm_R$ att skatta $\boldsymbol{\beta}-$parametrarna. Anledningen till detta handlade också i detta fall om för små

Markov-kedjor [10]. Denna metod anses därför inte vara lämplig för analys av data simulerad från Modell 2 vid stickprovssorlek 50.

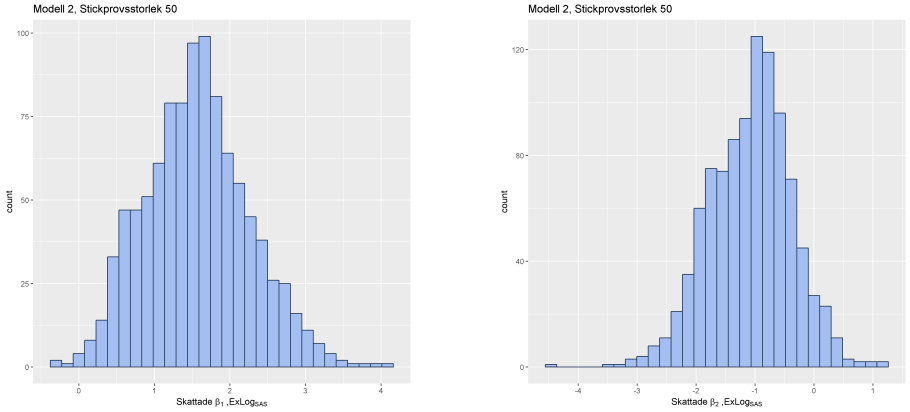
Nedan ser vi tabell 10, denna presenterar en sammanfattning av $\boldsymbol{\beta}-$skattningar från $ExLog_{SAS}$, $B-Elrm_R$ och MLE_R.

Modell 2, Stickprovsstorlek 50

	Min	1a Kvartil	Median	Medelvärde	3e Kvartil	Max
$ExLog_{SAS}$						
$\hat{\beta}_1$	-0.359	1.099	1.552	1.566	1.995	4.030
$\hat{\beta}_2$	-4.3768	-1.5560	-1.0266	-1.0803	-0.6321	1.2510
$B-Elrm_R$						
$\hat{\beta}_1$	-0.4424	1.0630	1.5103	1.5111	1.9205	3.7911
$\hat{\beta}_2$	-3.3814	-1.5329	-1.0206	-1.0546	-0.5898	2.4202
MLE_R						
$\hat{\beta}_1$	-0.3736	1.1471	1.6244	1.6728	2.0869	21.2001
$\hat{\beta}_2$	-40.279	-1.625	-1.072	-1.180	-0.658	1.311

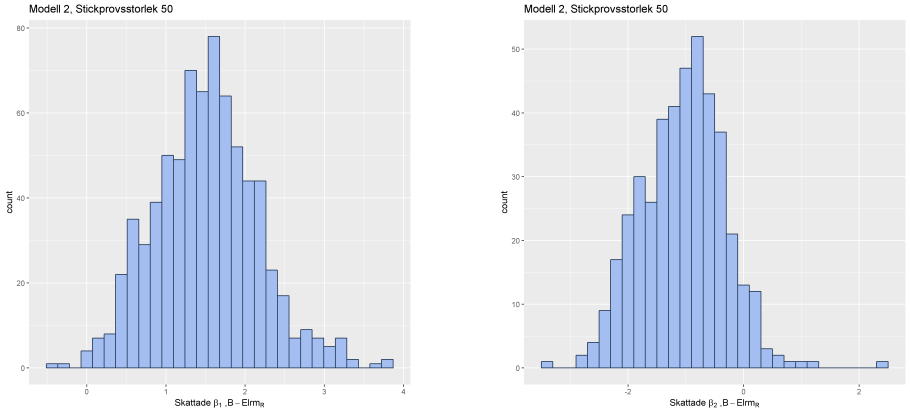
Table 10: Sammanfattning av $\boldsymbol{\beta}-$skattningar från 1000 stickprov som analyserats med respektive metod vid stickprovsstorlek 50.

Av tabell 10 ser vi att skattningarna som är gjorda med $ExLog_{SAS}$, $B-Erlm_R$ och MLE_R tycks vara väntevärdesriktiga, dessa undersöks närmare i histogrammen nedan.



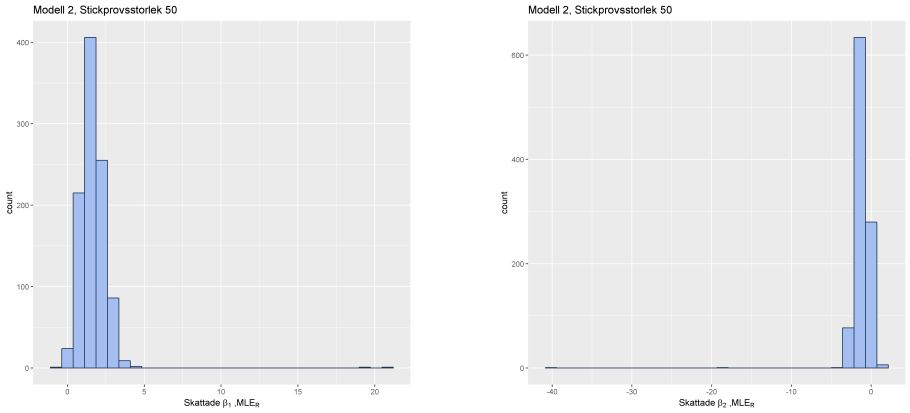
((a))

((b))



((c))

((d))



((e))

((f))

Notera att skalan på axlarna skiljer sig åt mellan graferna.

Figure 7: Histogram över $\boldsymbol{\beta}-$skattningarna från skattningsmetod $ExLog_{SAS}$, figur ((a)) & ((b)), $B - Elrm_R$, figur ((c)) & ((d)), och MLE_R, figur ((e)) & ((f)), vid stickprovsstorlek 50.

Hur $\boldsymbol{\beta}-$skattningarna är fördelade enligt ovan figurer för $ExLog_{SAS}$ och $B-Elrm_R$ kan likställas med en normalfördelning runt de sanna parametervärdena 1.5 respektive -1. Av figur 7(e) och 7(f) noteras att MLE_R skattar extrema värden för $\boldsymbol{\beta}-$parametrarna. Efter vidare undersökning visade det sig att dessa extremvärden estimerades för två stycken stickprov som båda saknade två respektive tre observationskombinationer.

Resultaten från ovanstående studie talar för att $ExLog_{SAS}$, $B-Elrm_R$ och MLE_R är lämpliga skattningsmetoder för analys av data likt den simulerad från Modell 2, där diskrepansen mellan medelvärdet av alla $\boldsymbol{\beta}-$skattningar är något större för MLE_R än för de övriga två metoderna.

3.3.4 Modell 2, stickprovsstorlek 100

Likt resultaten från stickprovsstorlek 20 och 50 misslyckades $A-Elrm_R$ att skatta $\boldsymbol{\beta}-$parametrarna vid stickprovsstorlek 100, på grund av för små Markovkedjor [10]. Detta inträffade nu även för stickprovsstorlek 100 med $B-Elrm_R$ som skattningsmetod. Dessa metoder anses därför inte vara lämpliga för analys av data simulerad från Modell 2 vid stickprovssorlek 100.

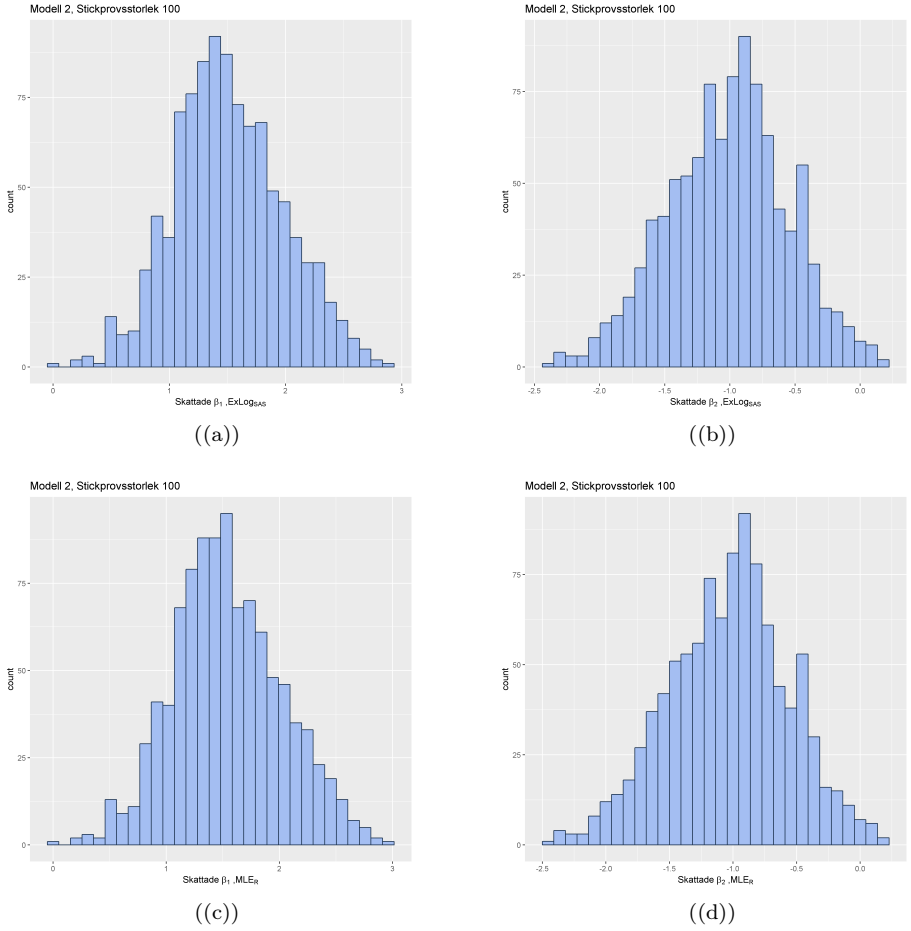
Vi fortsätter detta avsnitt med att undersöka tabell 11 som presenterar en sammanfattning av $\boldsymbol{\beta}-$skattningar från $ExLog_{SAS}$ och MLE_R.

Modell 2, Stickprovsstorlek 100

	Min	1a Kvartil	Median	Medelvärde	3e Kvartil	Max
$ExLog_{SAS}$						
$\hat{\beta}_1$	0.006087	1.194475	1.481450	1.509570	1.815100	2.889900
$\hat{\beta}_2$	-2.4063	-1.3264	-0.9808	-1.0138	-0.7077	0.1669
MLE_R						
$\hat{\beta}_1$	0.006214	1.220040	1.513500	1.543311	1.856018	2.968102
$\hat{\beta}_2$	-2.4659	-1.3534	-1.0020	-1.0354	-0.7223	0.1703

Table 11: Sammanfattning av $\boldsymbol{\beta}-$skattningar från 1000 stickprov som analyserats med respektive metod vid stickprovsstorlek 100.

Av tabell 11 ser vi att skattningarna som är gjorda med $ExLog_{SAS}$ och MLE_R tycks vara väntevärdesriktiga - tillhörande histogram undersöks närmare.



((a)) ((b))

((c)) ((d))

Notera att skalan på axlarna skiljer sig åt mellan graferna.

Figure 8: Histogram över $\boldsymbol{\beta}-$skattningarna från skattningsmetod $ExLog_{SAS}$, figur ((a)) & ((b)), och MLE_R, figur ((c)) & ((d)), vid stickprovsstorlek 100.

Av histogrammen ovan kan man utläsa att fördelningen över $\boldsymbol{\beta}-$skattningarna från $ExLog_{SAS}$ och MLE_R kan liknas vid en normalfördelning runt de sanna parametervärdena 1.5 respektive -1.

Resultaten påvisar att både $ExLog_{SAS}$ och MLE_R är metoder som är lämpliga för analays av data likt den simulerad från Modell 2 vid stickprovsstorlek 100.

3.3.5 Modell 2 - Systematisk skillnad mellan medelvärden och sanna parametrar

Modell 2, Systematisk skillnad mellan medelvärden och sanna parametrar

Metod	p-värde	Förkasta/Förkasta inte	Stickprovsstorlek
$A - Elrm_R$, β_1	$< 2.2e-16$	Förkasta H_0	10
$A - Elrm_R$, β_2	$< 2.2e-16$	Förkasta H_0	10
$B - Elrm_R$, β_1	$< 2.2e-16$	Förkasta H_0	10
$B - Elrm_R$, β_2	$< 2.2e-16$	Förkasta H_0	10
$ExLog_{SAS}$, β_1	0.4395	Förkasta inte H_0	20
$ExLog_{SAS}$, β_2	0.437	Förkasta inte H_0	20
$B - Elrm_R$, β_1	0.9995	Förkasta inte H_0	20
$B - Elrm_R$, β_2	0.6328	Förkasta inte H_0	20
$ExLog_{SAS}$, β_1	0.002712	Förkasta H_0	50
$ExLog_{SAS}$, β_2	0.0003243	Förkasta H_0	50
$B - Elrm_R$, β_1	0.6458	Förkasta inte H_0	50
$B - Elrm_R$, β_2	0.1183	Förkasta inte H_0	50
MLE_R , β_1	$9.81e-07$	Förkasta H_0	50
MLE_R , β_2	0.0002396	Förkasta H_0	50
$ExLog_{SAS}$, β_1	0.5196	Förkasta inte H_0	100
$ExLog_{SAS}$, β_2	0.3445	Förkasta inte H_0	100
MLE_R , β_1	0.004558	Förkasta H_0	100
MLE_R , β_2	0.01789	Förkasta H_0	100

Table 12: P-värden från "1-sample t-test", stickprovsstorlek 10, 20, 50 & 100.

Av tabell 12 ser vi att nollhypotesen inte kan förkastas för skattningsmetoderna $ExLog_{SAS}$, vid stickprovsstorlek 20 och 100, och $B-Elrm_R$, vid stickprovsstorlek 20 och 50. Testet kan alltså inte konstatera någon signifikant skillnad mellan medelvärdet av alla $\boldsymbol{\beta}-$parameterar från dessa två metoder och stickprovsstorlekar och det sanna parametervärdet. Resultaten talar därför inte heller emot de resultat som har diskuterats i avsnitten ovan - att $ExLog_{SAS}$ och $B-Elrm_R$ är lämpliga skattningsmetoder för stickprovsstorlek 20 och 100 respektive 20 och 50 i Modell 2.

För övriga skattningsmetoder och stickprovsstorlekar påvisar testen att skillnaden mellan alla $\boldsymbol{\beta}-$skattningarna och de sanna parametervärdena i Modell 2 snarare beror på något systematiskt än något slumpmässigt fel.

3.4 Komparativa skillnader och likheter mellan modellerna

Vid analys av den minsta stickprovsstorleken 10 visade resultaten på att skattningsmetoderna systematiskt skattade modellparametrarna fel i både Modell 1 och 2. Av studien tyder resultaten på att vid analys av stickprov, av storlek 10

och struktur likt de strukturer som har analyserats i denna uppsats, är ingen av de undersökta skattningsmetoderna - $ExLog_{SAS}$, A-, $B - Elrm_R$ och MLE_R, lämpliga.

För stickprovsstorlek 20 indikterade resultaten från simuleringsstudien att $ExLog_{SAS}$ och $B - Elrm_R$ var lämpliga skattningsmetoder för data från Modell 1 och 2 - det kunde inte påvisas någon eller endast en förhållandevis liten systematisk skevhet vid skattning av $\boldsymbol{\beta}-$parametrarna. Skattningsmetoden $A - ELrm_R$ skattade modellparametern systematiskt för lågt i Modell 1 och mysslyckades att skatta modellparametrarna i Modell 2 helt. I de undersökta modellerna berodde skattningarna från MLE_R på om, och i sådant fall vilka, observationskombinationer saknades i det stickprovs som analyserades. Dessa metoder anses därför inte vara lämpliga för analys av stickprov av storlek 20 av struktur likt de som har undersökts i denna uppsats.

Analys av stickprovsstorlek 50 visade att både $ExLog_{SAS}$ och MLE_R var lämpliga skattningsmetoder för data simulerad från de två undersökta modellerna, båda skattade väntevärdesriktiga $\boldsymbol{\beta}-$parametrar. Resultaten för A- och $B - Elrm_R$ visade på att dessa metoder inte är lämpade för analys av data från Modell 1 - de skattade modellparametern systematiskt lägre än det sanna parametervärdet. Vid analys av data simulerad från Modell 2, skiljde resultaten sig åt för de två ovannämnda skattningsmetoderna. $A - Elrm_R$ anses inte vara lämplig då den misslyckades med att ta fram $\boldsymbol{\beta}-$skattningar, medan mycket tyder på att $B - Elrm_R$ lämpar sig väl för analys av liknande problem då metoden skattade väntevärdesriktiga skattningar där ingen systematisk skevhet kunde påvisas.

Vid analys av den största stickprovsstorleken 100, fås liknande resultat för de undersökta modellerna som för stickprovsstorlek 50. Det som skiljer resultaten åt är att skattningsmetoden $B - Elrm_R$ inte visade sig vara lämpad för analys av data simulerad från Modell 2 då den misslyckades med att ta fram skattningar för modellparametrarna utan här föredras istället skattningsmetoderna $ExLog_{SAS}$ och MLE_R.

4 Diskussion

I denna del analyseras och diskuteras resultaten från del 3. Tolkningar och tankar inför vidare forskning på ämnet tas också upp.

4.1 Diskussion av simuleringen

Gemensamt för analysen av de båda modellerna, 9 och 10, är att i och med att stickprovsstorleken ökar närmar sig parameterskattningarna från $ExLog_{SAS}$ och MLE_R de sanna parametervärdena. Skillnaden mellan de båda metodena minskar i och med att stickprovsstorleken ökar. Detta medan resultaten från A- och $B - Elrm_R$ visar att de skattade parametervärdena vid analys av små stickprov lämpar sig bättre än vid analys av stora stickprov. Vid analys av små stickprov av datastrukturer studerad i denna uppsats visar studien generella

tendenser till att av de metoder som har undersökts i denna uppsats är $B - Elrm_R$ är mest lämplig.

Enligt en studie av Nemes et al. (2009) [2] tyder simuleringsresultaten på att MLE_R överskattar effekter vid analys av små stickprov, likt de resultat vi har fått i denna studie, en bias som beror på stickprovets storlek och struktur. I denna studie har endast två datastrukturer undersökts, Modell 1 där den binära responsvariabeln beror på en binär variabel, och Modell 2, där den binära responsvariabeln beror på två av varandra oberoende binära variabler. Vid små stickprov av dessa strukturer förekommer stickprov för vilka en eller fler observationskombinationer saknas, vilket resulterar i skattningar från MLE_R som ligger långt ifrån de sanna modellparametrarna. Detta indikerar att man generellt behöver större stickprov för analys av liknande problem. Resultat från liknande simuleringsstudier har också påvisat samma tendenser, och att analys med hjälp av MLE_R av data där responsvariabel beror av diskreta variabler, likt Modell 1 och 2, generellt behöver större stickprov än vid analys av data där responsvariabeln beror av kontinuerliga variabler eller variabler som är starkt korrelerade [2]. Detta är dock något som ligger utanför studiens tillämpningsområde men det är ett viktigt resultat som bör tas i beaktning av framtida forskning på området.

De två skattningsmetoderna A- och $B-Elrm_R$ är två alternativ av en funktion implementerad i R. I A återfinns standardvärdena för funktionen och i B en utökad iterationsgräns av Markov-kedjan (MC), se avsnitt 2.2.2. I linje med uppsatsens syfte är det viktigt att jämföra analys av olika stickprovsstorlekar vid olika iterationsgränser för MC. Hur stora MC som kan sparas ned i datorns virtuella minne är begränsat då RAM-minnet på datorn kan vara otillräckligt - under studiens gång visade sig detta vara en förhindrande omständighet. Att applicera skattningsmetod A- respektive $B - Elrm_R$ på ett stickprov tog cirka 3-9 sekunder respektive 15-50 sekunder beroende på den modell och stickprovsstorlek som analyserades. Här erhålls alltså ett resultat från funktionen Elrm i R med alternativ A ungefär fem gånger snabbare än för alternativ B. Denna skillnad kanske inte spelar så stor roll vid analys av ett eller endast ett fåtal stickprov, men så fort man kommer upp i ett större antal, likt denna simuleringsstudie, blir detta en faktor att väga in. Det går också att testa skattningsmetoden Elrm för större MC än vad det virtuella minnet tillåter. Då rekommenderas att man genererar MC i olika sekvenser och sparar ned dem separat för att på så sätt frigöra virtuellt minne och därigenom ta bort restriktionen för MC med avseende på datorns virtuella minne.

Skattningsmetoderna A- och $B-Elrm_R$ misslyckades att skatta $\boldsymbol{\beta}-$ parametrarna i Modell 2 vid stickprovsstorlek $20, 50, 100$ respektive 100. Detta då de inte lyckades utföra en fullständig inferens på grund av att iterationsgränsen för Markov-kedjan inte kunde uppnås. Detta gav resultat som indikerade att vid undersökning av större och mer komplicerade problem var dessa skattningsmetoder inte lämpliga. Här kan det vara viktigt att ta i beaktning att båda

metoderna är alternativ av en funktion implementerad i mjukvaruprogrammet R där forskaren själv kan sätta iterationsgränsen för Markov-kedjan. De misslyckade skattningarna kan indikera att iterationsgränsen för Markov-kedjan kanske bör utökas i och med att stickprovsstorleken ökar och datastrukturen blir mer komplicerad. Att utforska detta ytterligare är en rekommendation för framtida simuleringar inom forskningsområdet.

4.2 Studien i en större kontext

I denna uppsats har olika skattningsmetoder undersökts vid analys av data simulerad från två modeller i fyra olika stickprovsstorlekar för att undersöka om det finns några lämpliga alternativ till den mer vanligt tillämpade skattningsmetoden MLE.

En tydlig avgränsning i denna studie var möjligheterna att köra algoritmer för samtliga skattningsmetoder i enbart ett mjukvaruprogram. Exempelvis kan inte SAS köra simulering för $Elrm_R$ på ett tillfredställande vis. Detta kan skapa problem om man vill ha så likartade simuleringar som möjligt - att analysera resultat från två olika mjukvaruprogram handlar också om att sammanställa resultaten på ett sätt så att den enkelt kan jämföras - något som är tidskrävande och visade sig vara en utmaning under arbetets gång.

De avgränsningar som gjordes var i mångt och mycket tillräckliga. Något som saknades var en analys av fler datastrukturer än de två som analyserades i denna studie. Detta hade kunnat ge ett större analysunderlag för jämförelsen mellan de olika skattningsmetoderna. Det finns liknande studier som har genomförts och som redovisar resultat av intresse för de simuleringar som har gjorts i denna studie. Oster [7], [8] undersöker och jämför exakta metoder som finns implementerade i olika mjukvarupaket.

Ett lämpligt område att fortsätta undersöka skulle vara att bredda jämförelsen till att inkludera "penalized maximum Likelihood method", en skattningsmetod som inte berörs av denna uppsats. Liknande studier har inkluderat denna skattningsmetod, vars resultat har påvisat att denna metod skulle kunna vara en lämplig skattningsmetod vid analys av små stickprov [13], [14].

5 Referenser

References

[1] Alan Agresti. *Categorical Data Analysis*. John Wiley & Sons Inc., 2002.

[2] Szilard Nemes, Junmei Milao Jonasson, Anna Genell & Gunnar Steineck. *Bias in odds ratios by logistic regression modelling and sample size*. BMC Medical Research Methodology, Vol. 9, No. 56 (2009).

[3] Bryman A & Bell E. *Business Research Methods*. OUP Oxford, 2011.

[4] Georg Heinze. *A comparative investigation of methods for logistic regression with separated or nearly separated data*. Stat Med, Vol. 30, No. 25(24), (2006), pp. 4216-26.

[5] Cyrus R Mehta & Nitin R Patel. *Exact Logistic Regression: Theory and Examples*. Statistics in Medicine, Vol. 14, No. 19 (Oct., 1995), pp. 2143-2160.

[6] Karim F Hirji. *Exact analysis of discrete data*. Chapman & Hall/CRC, 2006.

[7] Oster, Robert A. *An Examination of Statistical Software Packages for Categorical Data Analysis Using Exact Methods*. The American Statistician, Vol. 56, No. 3 (Aug., 2002), pp. 235-246.

[8] Oster, Robert A. *An Examination of Statistical Software Packages for Categorical Data Analysis Using Exact Methods – Part II*. The American Statistician, Vol. 57, No. 3 (Jan., 2003), pp. 201–213.

[9] Robert E Derr. *Performing Exact Logistic Regression with the SAS System — Revised 2009*. [*SAS Institute Inc., Cary, NC*].

[10] David Zamar, Brad McNeney & Jinko Graham. *elrm: Software Implementing Exact-like Inference for Logistic Regression Models*. Journal of Statistical Software, Vol. 21, No. 3 (Oct., 2007).

[11] Jordan M I, Doucet A, de Freitas N, Andrieu C *An Introduction to MCMC for Machine Learning* Machine Learning, Vol. 50, (2003), pp. 5–43.

[12] Mehta CR, Patel NR, Senchaudhuri P. *Efficient Monte Carlo Methods for Conditional Logistic Regression*. Journal of The American Statistical Association, Vol. 95, No.449 (Mar., 2000), pp. 99–108.

[13] Heinze G, Schemper M. *A solution to the problem of separation in logistic regression*. Statistics in Medicine, Vol. 21, No. 16 (Aug., 2002), pp. 2409–2419.

[14] Bull S, Mak C, Greenwod CMT. *A modified score function estimator for multinomial logistic regression in small samples*. Computational Statistics and Data Analysis, Vol. 39, No. 1 (Mar., 2002), pp. 57-74.

[15] Paul D. Allison. *Logistic Regression Using SAS: Theory and Application.* SAS Institute, (1999).

6 Appendix

6.1 Odds och Oddskvot

Sannolikhet tolkas ofta som det "naturliga" sättet att kvantifiera chansen att en händelse inträffar, och tillskrivs en siffra mellan 0 och 1, där 0 innebär att händelsen med säkerhet inte inträffar och 1 innebär att händelsen med säkerhet inträffar. Det finns alternativa sätt att framställa chanserna att en händelse inträffar, en av dem - the odds - kan påstås ha samma rättighet att kallas "naturlig".

The odds för en händelse är kvoten mellan det förväntade antalet gånger en händelse inträffar och det förväntade antalet gånger en händelse inte inträffar. Har vi en odds på 4 förväntar vi oss 4 gånger så många inträffade händelser som icke-inträffade händelser. En odds på $\frac{1}{5}$ innebär att vi förväntar oss en femtedels så många inträffade händelser som icke-inträffade händelser.

Relationen mellan odds och sannolikhet är enkel. Om vi låter π vara sannolikheten för att en händelse inträffar och O vara the odds att en händelse inträffar, då har vi:

$$O = \frac{\pi}{1-\pi} = \frac{\text{sannolikheten att en händelse inträffar}}{\text{sannolikheten att händelsen inte inträffar}} \tag{11}$$

eller uttryckt på ett annat sätt:

$$\pi = \frac{O}{1+O}.$$

Nedan ser vi den ovannämnda relationen illustrerad i en tabell.

Sannolikhet	Odds
0.1	0.11
0.2	0.25
0.3	0.43
0.4	0.67
0.5	1.00
0.6	1.50
0.7	2.33
0.8	4.00
0.9	9.00

Table 13: Relationen mellan sannolikhet och odds

Från tabellen ovan observerar vi att odds under 1 motsvarar sannolikheter under 0.5, medan odds större än 1 motsvarar sannolikheter större än 0.5. Lik-

som sannolikheterna har odds en undre gräns på 0, men till skillnad från sannolikheter har odds ingen övre gräns.

Hur kan vi använda oss av odds? Låt oss säga att sannolikheten att jag röstar i nästa riksdagsval är 0.40 och sannolikheten att en annan person röstar är 0.80. Då är det rimligt att kunna säga att den andra personens sannolikhet är dubbelt så stor som min. Men vad händer om sannolikheten att jag röstar är 0.80, det blir nu omöjligt för din sannolikhet att vara dubbelt så stor som min. Uttrycker vi chanserna för att jag ska rösta i termer av odds ser vi att en sannolikhet på 0.80 motsvarar en odds av $0.80/0.20 = 4$. Fördubblar vi den får vi en odds på 8, denna kan vi såklart konvertera till en sannolikhet igen och får då $8/(1+8) = 0.8889$.

Detta leder oss in på ett begrepp som kallas oddskvoten. Det är ett mått av relationen mellan två dikotomiska variabler. Oddskvoten är kvoten av två odds;

$$OR = \frac{O_1}{O_2}, \tag{12}$$

och används för att kvantifiera hur en variabel som man är intresserad av (att jag röstar i nästa val) förhåller sig till en annan variabel (att du röstar i nästa val). Återgår vi till exemplet om huruvida vi kommer att rösta i nästa riksdagsval eller inte, så är oddsen för att du röstar $8/4 = 1.5 = 50\%$ större än att jag röstar. Notera, beroende på hur vi ställer upp oddskvoten kan vi få en storhet som är större än 1 eller dess invers, vilket är mindre än 1. [15]

6.2 Exemplifiering av exakt betingad inferens

Nedanstående exempel av exakt betingad intferens är hämtat från Derr (2009) [9]. Målet i exakt betingad inferens är att bestämma hur sannolikt det observerade värdet $\boldsymbol{y_0}$ är med avseende på alla 2^n möjliga observerade $\boldsymbol{y} = (y_1, ..., y_n)'$. Ett tillvägagångssätt är att ta fram alla y-vektorer för vilka $\boldsymbol{y'X_0 = t_0}$ och beräkna antalet vektorer $\boldsymbol{y}$ gör vilka $\boldsymbol{y'X_1}$ är lika varje unikt $\boldsymbol{t_1}$, där $\boldsymbol{t_0}$ och $\boldsymbol{t_1}$ är tillräckliga statistikor för skräpparametrarna respektive parametrarna av intresse.

Antag att vi har följande stickprov, och att vi vill hitta den exakta betingade fördelningen av de tillräckliga statistikorna för X_1 betingat på dem för X_0.

Observation	y	x_0	x_1
1	0	1	1
2	1	1	1
3	0	1	2
4	1	1	0

Table 14

Vi har alltså observerat $\boldsymbol{y_0} = (0,1,0,1)'$, $\boldsymbol{X_0} = (1,1,1,1)'$ och $\boldsymbol{X_1} = (1,1,2,0)'$. Vi kan beräkna det observerade $\boldsymbol{t}$ som $(t_0, t_1) = 0 \times (1,1) + 1 \times$

$(1,1) + 0 \times (1,2) + 1 \times (1,0) = (2,1)$, så betingning utförs på $t_0 = 2$. Tabellen nedan är de 16 möjliga $\boldsymbol{y} = (y_1, y_2, y_3, y_4)$ vektorerna med respektive $\boldsymbol{t} = (t_0, t_1)$-värden:

Vektor	y_1	y_2	y_3	y_4	t_0	t_1
1	0	0	0	0	0	0
2	0	0	0	1	1	0
3	0	0	1	0	1	2
4	0	0	1	1	2	2
5	0	1	0	0	1	1
6	0	1	0	1	2	1
7	0	1	1	0	2	3
8	0	1	1	1	3	3
9	1	0	0	0	1	1
10	1	0	0	1	2	1
11	1	0	1	0	2	3
12	1	0	1	1	3	3
13	1	1	0	0	2	2
14	1	1	0	1	3	2
15	1	1	1	0	3	4
16	1	1	1	1	4	4

Table 15

Den betingade fördelningen härleds sedan från denna simultana fördelningen genom att ta ut varje vektor där $t_0 = 2$.

t_0	t_1	Frekvens	Sannolikhet
2	1	2	2/6
2	2	2	2/6
2	3	2	2/6
Totalt		6	1

Table 16

Att sedan ta fram den betingade fördelningen genom iteration över den simultana fördelningen är konceptuellt enkelt, men denna metod blir beräkningsmässigt tungt väldigt snabbt. Till exempel, med 30 observationer måste 2^{30} olika y-vektorer undersökas, vilket är mer än en miljard. En algoritm, kallad "The multivariate shift algoithm", framtagen av Hirji, Mehta, and Patel (1987) är en metod för att generera och räkna antalet y-vektorer vid analys av större problem. Algoritmen baseras på följande observation. Givet något $\boldsymbol{y} = (y_1, ..., y_n)'$ och $\boldsymbol{X} = (x_1, ..., x_n)'$, låt $\boldsymbol{y_{(i)}} = (y_1, ..., y_i)'$ och

$$X_{(i)} = (x_1, ..., x_i)' = \begin{bmatrix} x_{1,1} & ... & x_{1,p+q} \\ . & & . \\ . & & . \\ . & & . \\ x_{i,1} & ... & x_{1,p+q} \end{bmatrix}$$

vara de första i raderna i varje matris. De tillräckliga statistikorna av dessa i rader betecknas som $t'_{(i)} = y'_{(i)}X_{(i)}$. Den rekursiva relationen ger då: $t_{(i+1)} = t_{(i)} + y_{i+1}x_{i+1}$.

Figur 9 visar ett träddiagram där varje rad efter den första raden motsvarar en observation i, och varje nod i trädet är betcknad med siffror som representerar värdet av $\boldsymbol{t_{(i)}}$. Varje rad är numrerad, vilka representerar stegen i algoritmen. För att gå ner i grenarna, adderas y multiplicerat med nästa värde av (x_0, x_1) till det nuvarande värdet av (t_0, t_1), för $y = 0$ och 1. Ecempelvis, om vi börjar vid steg 0 med $t_{(0)} = (0, 0)$, så blr nästa värde av den vänstra grenen $t_{(0)} + yx_1 = (0, 0) + 0(1, 1) = (0, 0) = 00$ och $(0, 0) + 1(1, 1) = (1, 1) = 11$ för den högra grenen.

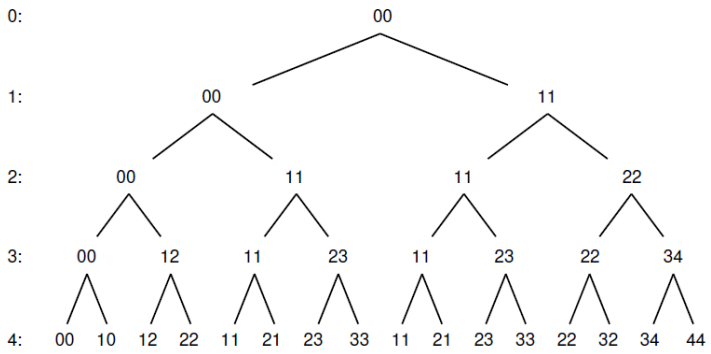


Figure 9: Stegen i "Multivariate Shift Algoritm" som används vid skattningsmetoden $ExLog_{SAS}$.

t_0	t_1	Frekvens	Sannolikhet
0	0	1	1/16
1	0	1	1/16
1	1	2	2/16
1	2	1	1/16
2	1	2	2/16
2	2	2	2/16
2	3	2	2/16
3	2	1	1/16
3	3	2	2/16
3	4	1	1/16
4	4	1	1/16
Totalt		16	1

Table 17

Tabell 17 visar fördelningen skapad från frekvenstabellen av $2^4 = 16$ möjliga $\boldsymbol{t}$-vektorer från sista steget i figur 9. Den betingade fördelningen för det observerade värdet $t_0 = 2$ är densamma som som den som togs fram med hjälp av tabell 16.

6.3 Kompletterande figurer och tabeller för Modell 1

Stickprovsstorlek 10

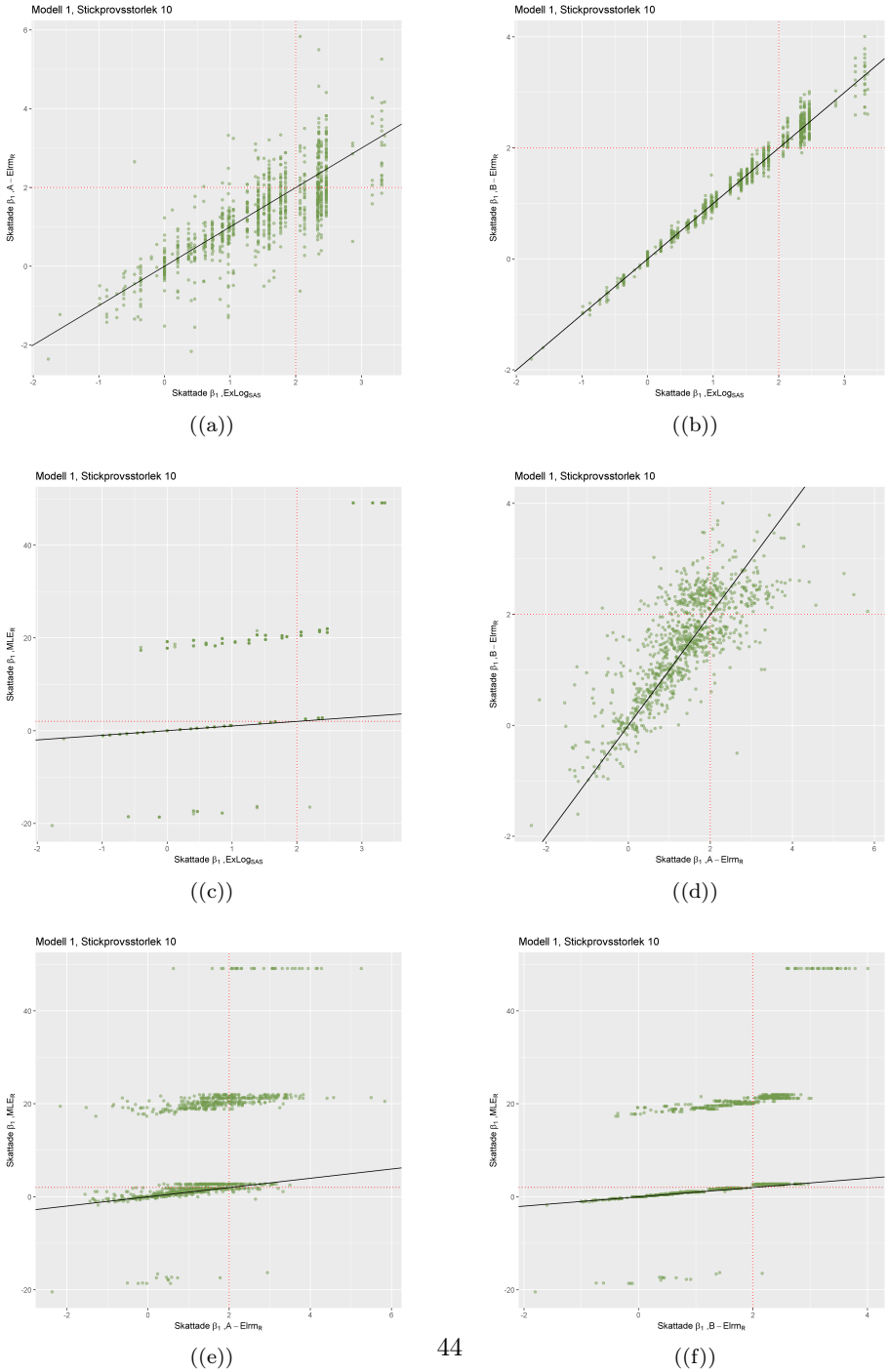
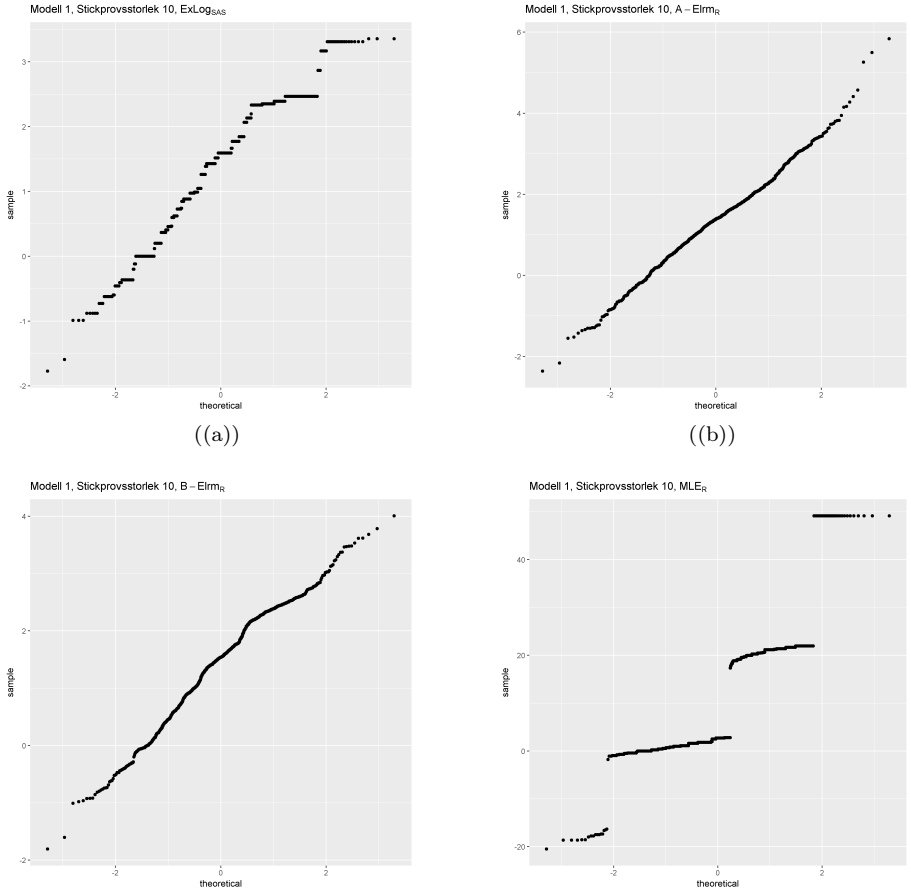


Figure 10: I ovanstående figurer plottas β_1-skattningrna från de olika skattningsmetoderna mot varandra.



((a)) ((b))

((c)) ((d))

Figure 11: QQ-Plot för ((a)) $ExLOG_{SAS}$, ((b)) $A - Elrm_R$, ((c)) $B - Elrm_R$ och ((d)) MLE_R, vid stickprovsstorlek 10, Modell 1.

Stickprovsstorlek 20

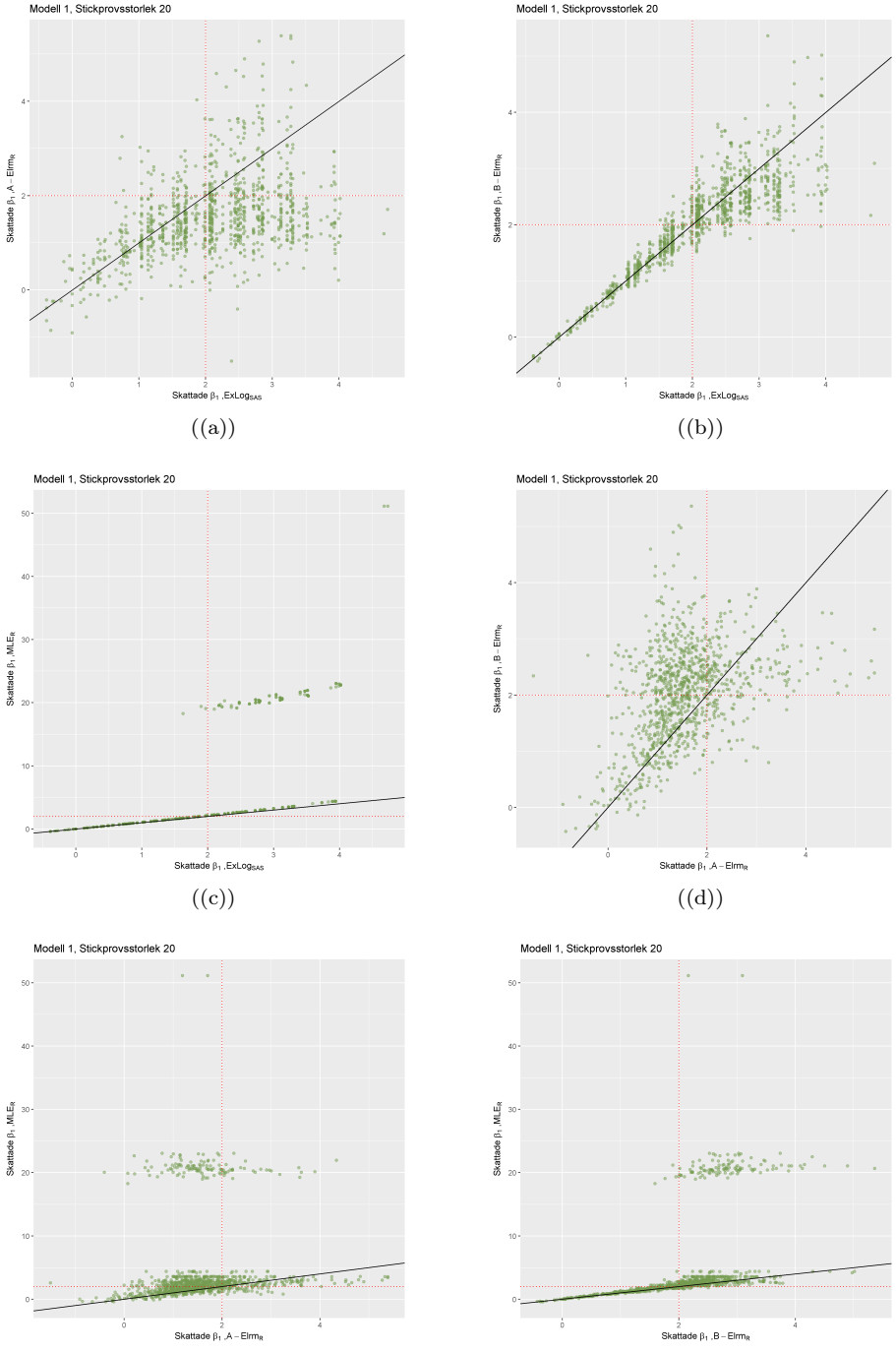
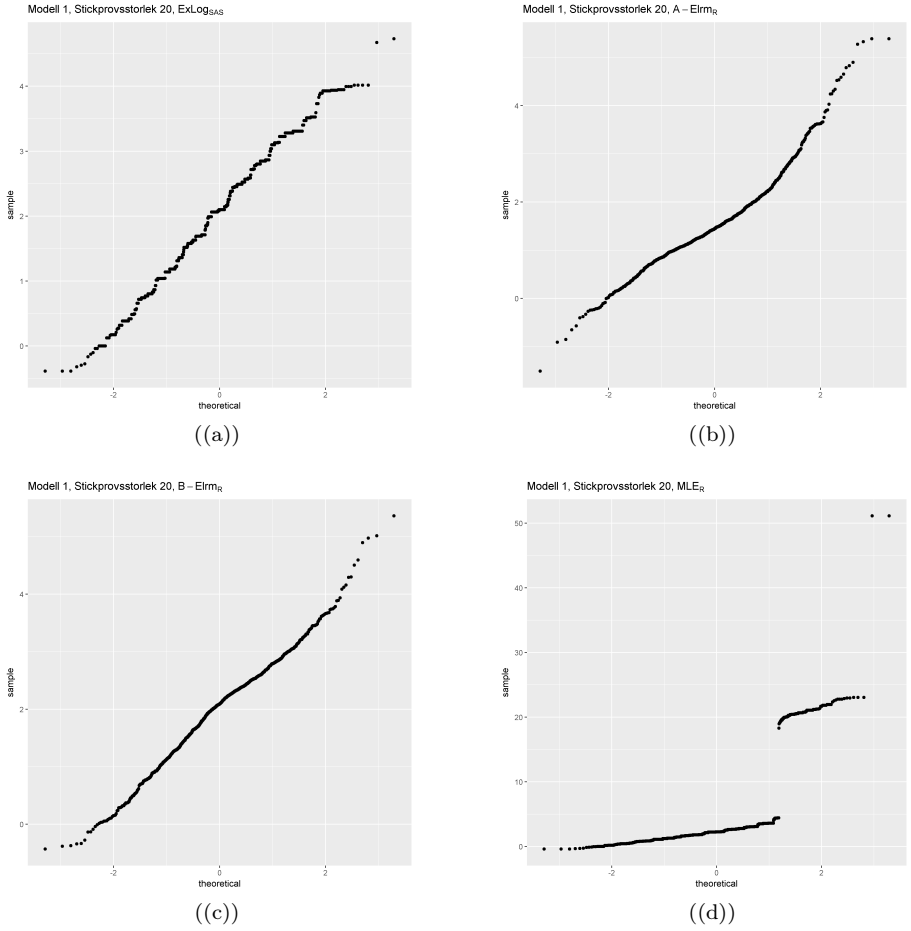


Figure 12: I ovanstående figurer plottas β_1-skattningrna från de olika skattningsmetoderna mot varandra vid stickprovsstorlek 20, Modell 1.



((a)) ((b)) ((c)) ((d))

Figure 13: QQ-Plot för ((a)) $ExLOG_{SAS}$, ((b)) $A - Elrm_R$, ((c)) $B - Elrm_R$ och ((d)) MLE_R, vid stickprovsstorlek 20, Modell 1.

Stickprovsstorlek 50

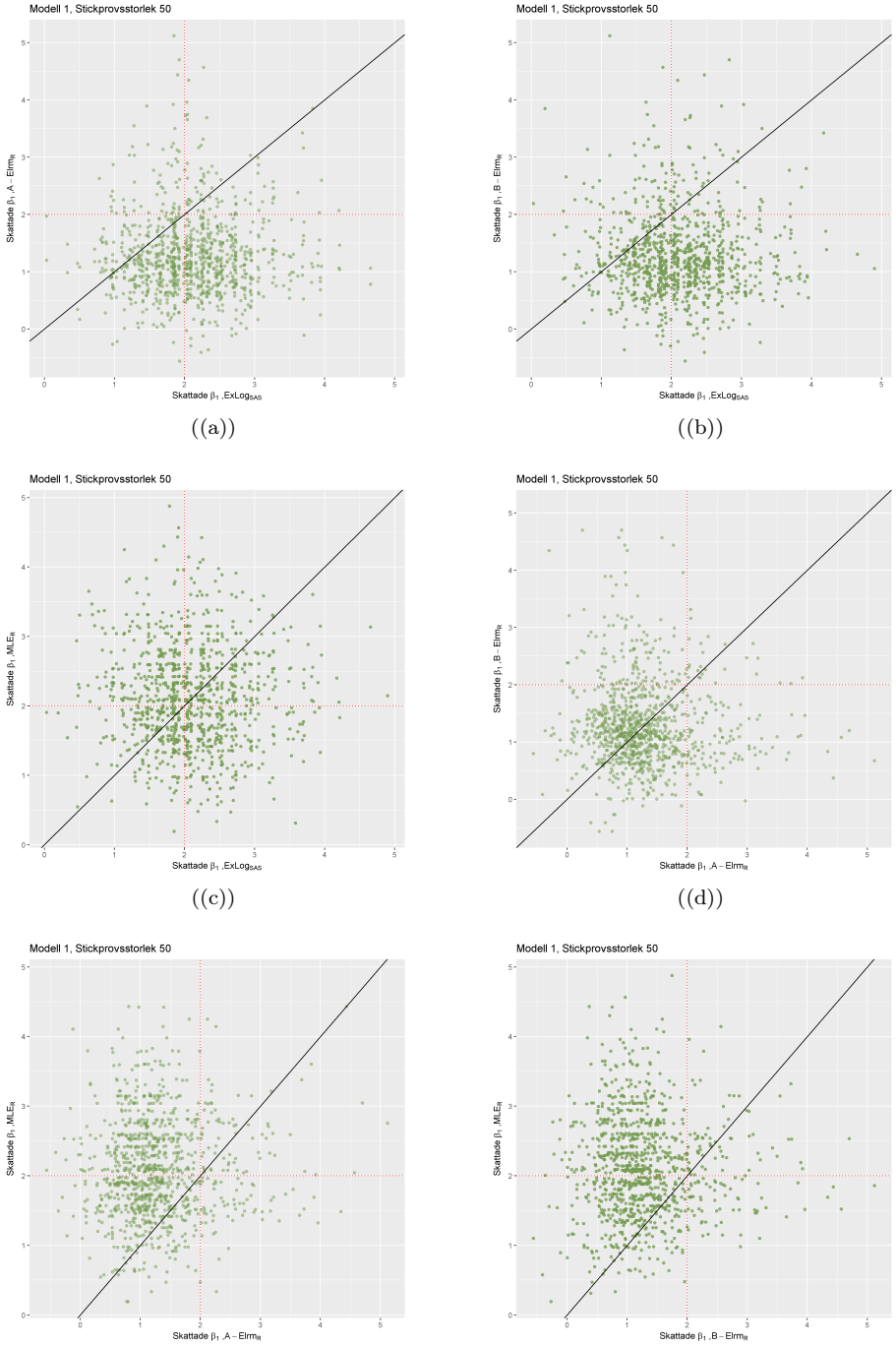
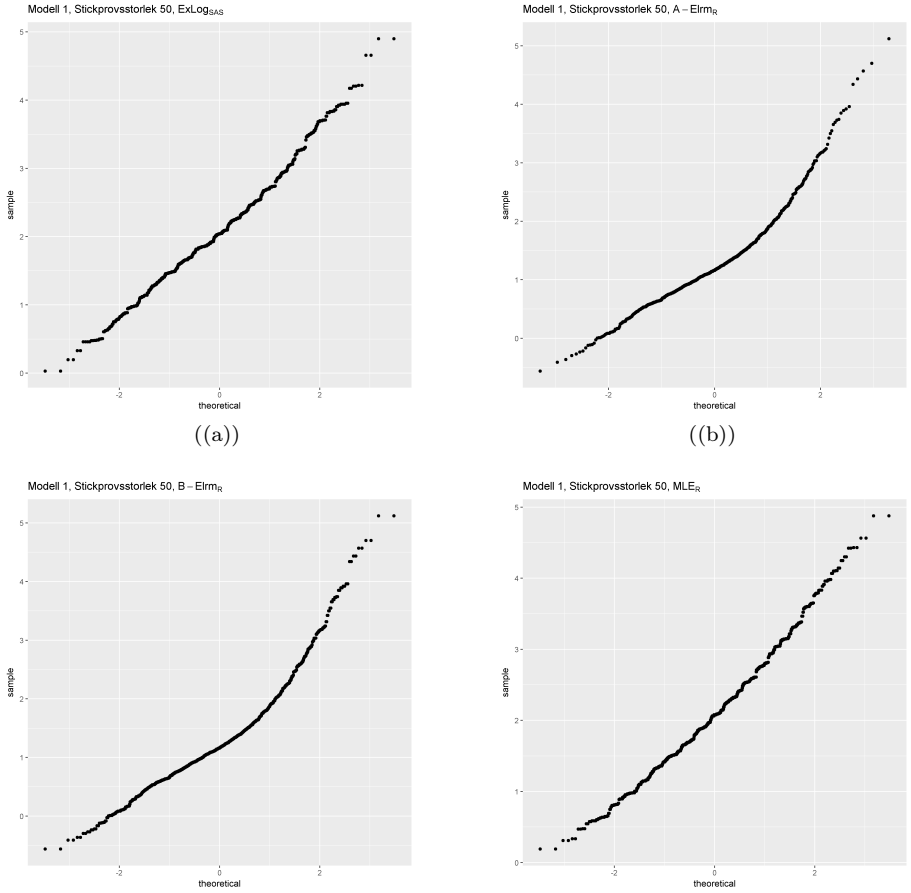


Figure 14: I ovanstående figurer plottas β_1-skattningrna från de olika skattningsmetoderna mot varandra vid stickprovsstorlek 50, Modell 1.



((a)) ((b)) ((c)) ((d))

Figure 15: QQ-Plot för ((a)) $ExLOG_{SAS}$, ((b)) $A - Elrm_R$, ((c)) $B - Elrm_R$ och ((d)) MLE_R, vid stickprovsstorlek 50, Modell 1.

Stickprovsstorlek 100

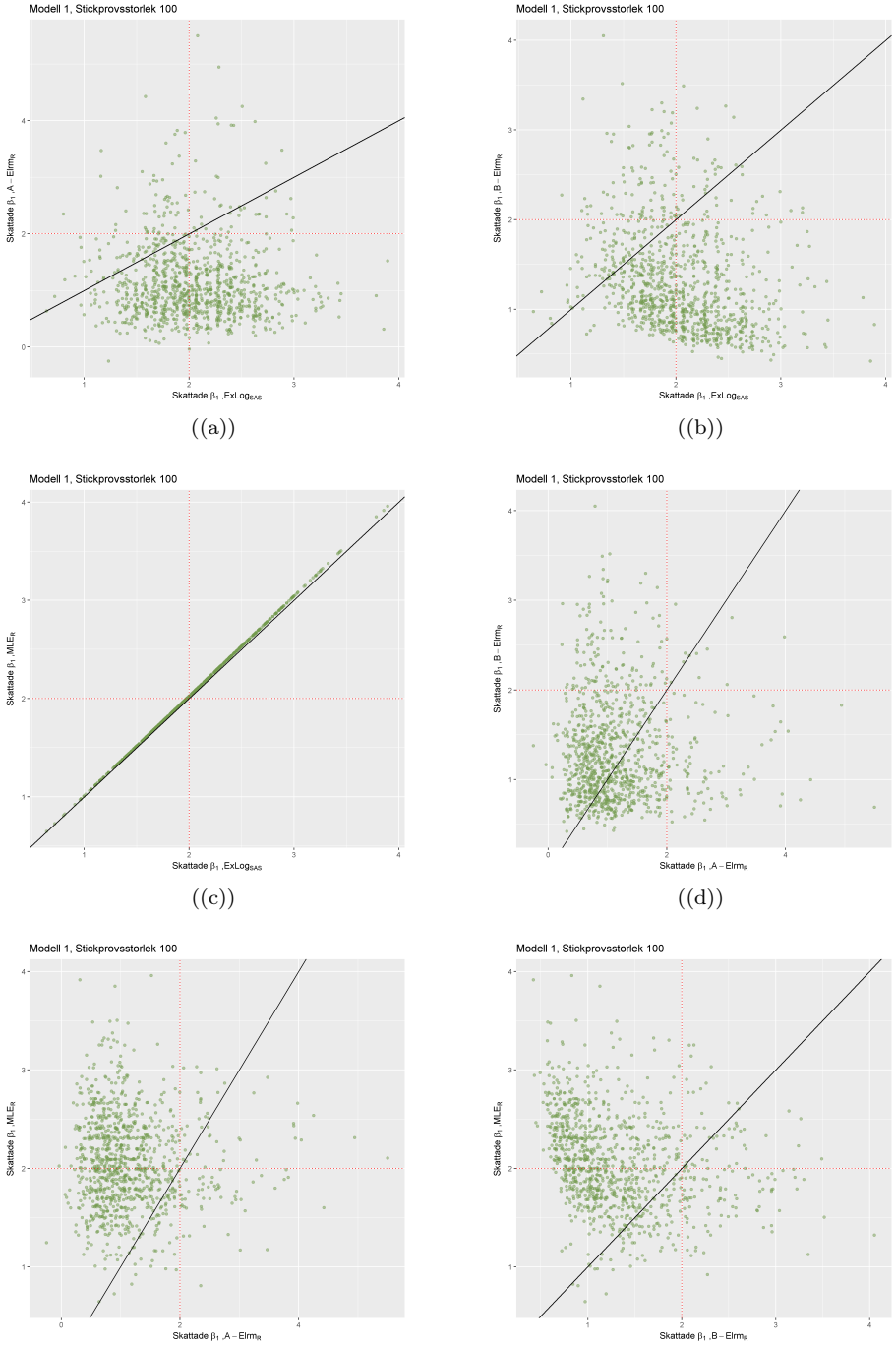
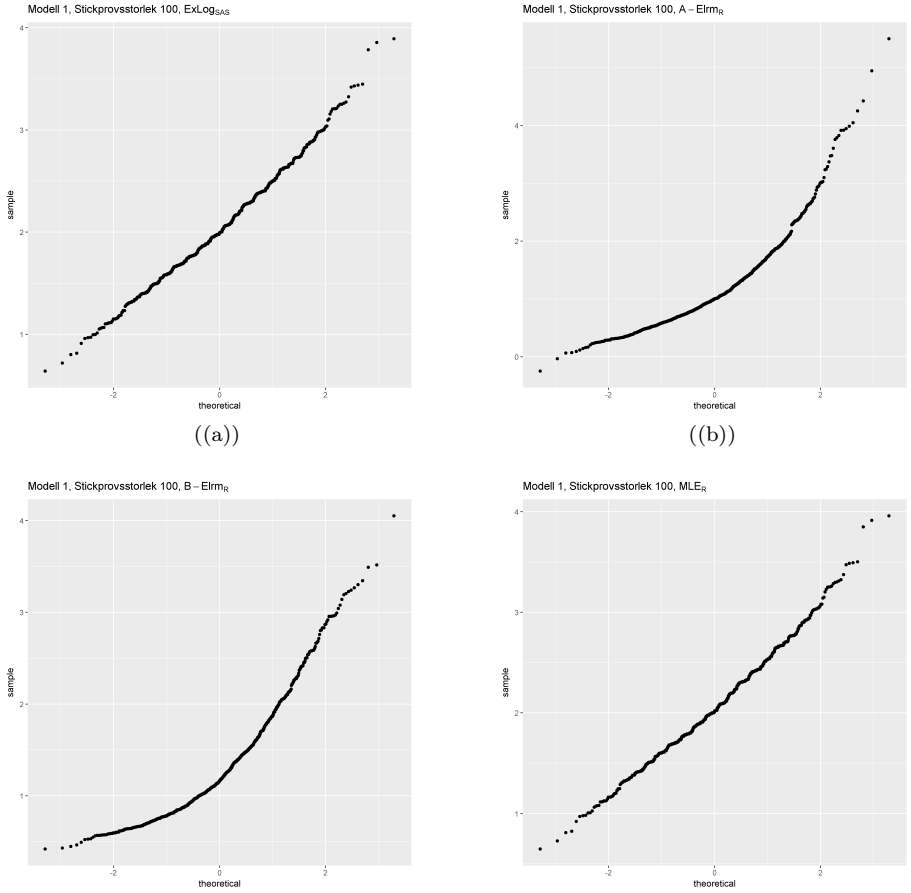


Figure 16: I ovanstående figurer plottas β_1-skattningrna från de olika skattningsmetoderna mot varandra vid stickprovsstorlek 100, Modell 1.

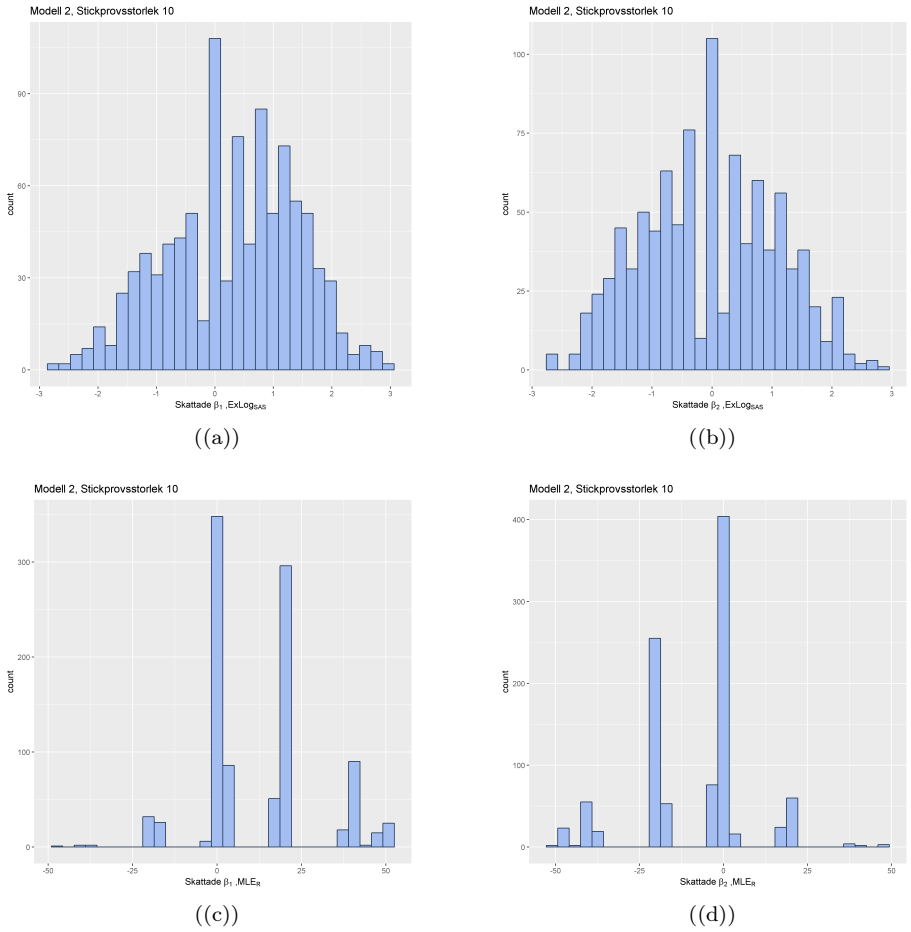


((a)) ((b)) ((c)) ((d))

Figure 17: QQ-Plot för ((a)) $ExLOG_{SAS}$, ((b)) $A - Elrm_R$, ((c)) $B - Elrm_R$ och ((d)) MLE_R, vid stickprovsstorlek 100, Modell 1.

6.4 Kompletterande figurer och tabeller för Modell 2

Stickprovsstorlek 10

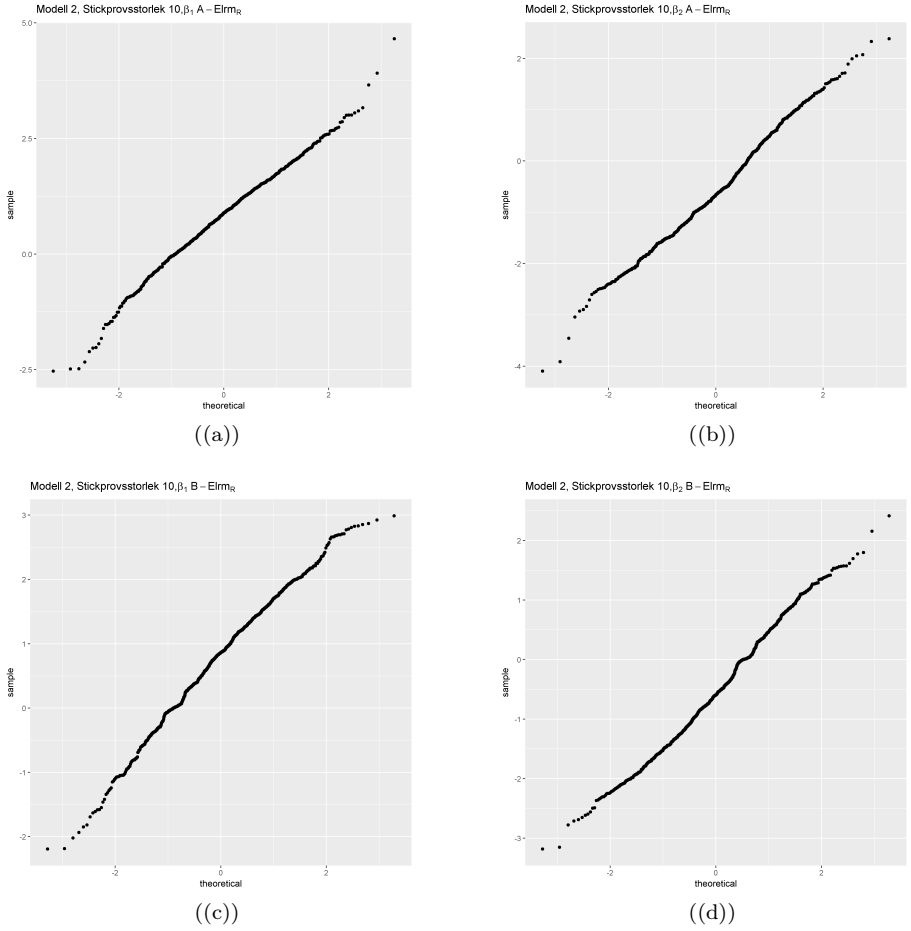


((a)) ((b))

((c)) ((d))

Notera att skalan på axlarna skiljer sig åt mellan graferna.

Figure 18: Histogram över $\boldsymbol{\beta}-$skattningarna från skattningsmetod $ExLog_{SAS}$, figur ((a)) & ((b)), och MLE_R, figur ((c)) & ((d)), vid stickprovsstorlek 10.

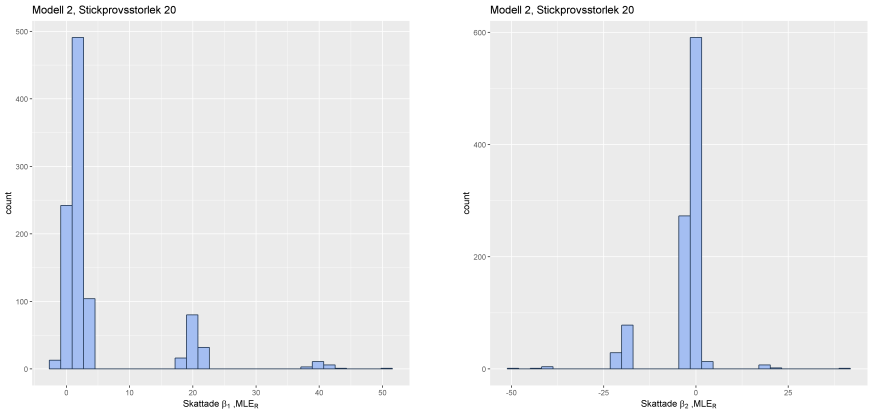


((a)) ((b))

((c)) ((d))

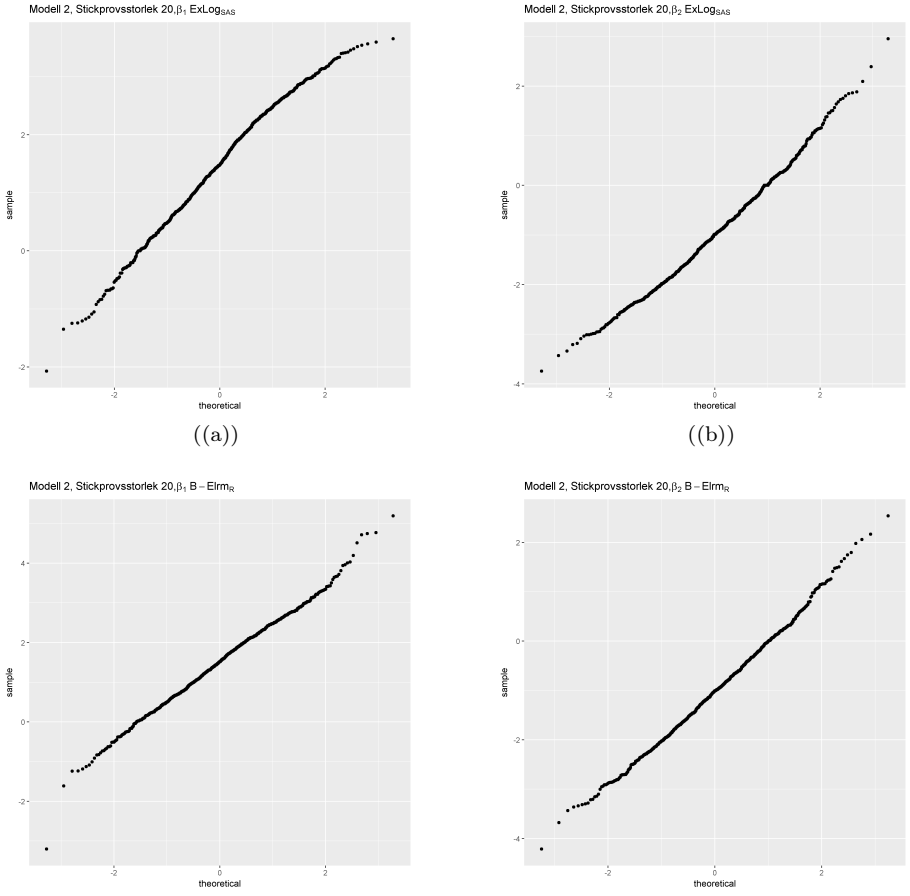
Figure 19: QQ-Plot för $A - Elrm_R$, figur ((a)) & ((b)), $B - Elrm_R$, figur ((c)) & ((d)), vid stickprovsstorlek 10, Modell 2.

Stickprovsstorlek 20



((a)) ((b))

Figure 20: Histogram över $\boldsymbol{\beta}-$skattningarna från skattningsmetod MLE_R, figur ((a)) & ((b)), vid stickprovsstorlek 20, Modell 2.

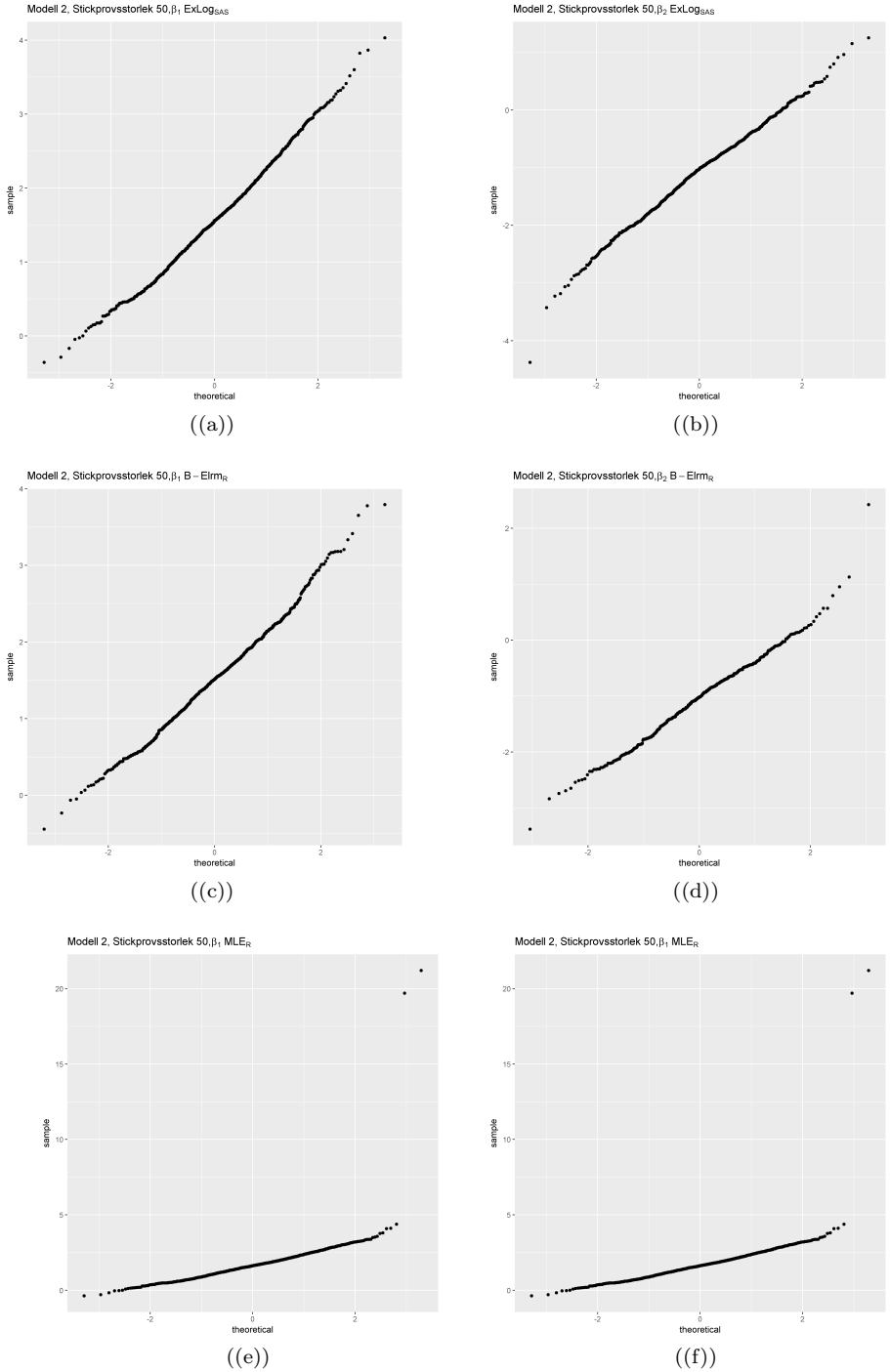


((a)) ((b))

((c)) ((d))

Figure 21: QQ-Plot för $ExLog_{SAS}$, figur ((a)) & ((b)), $B - Elrm_R$, figur ((c)) & ((d)), vid stickprovsstorlek 20, Modell 2.

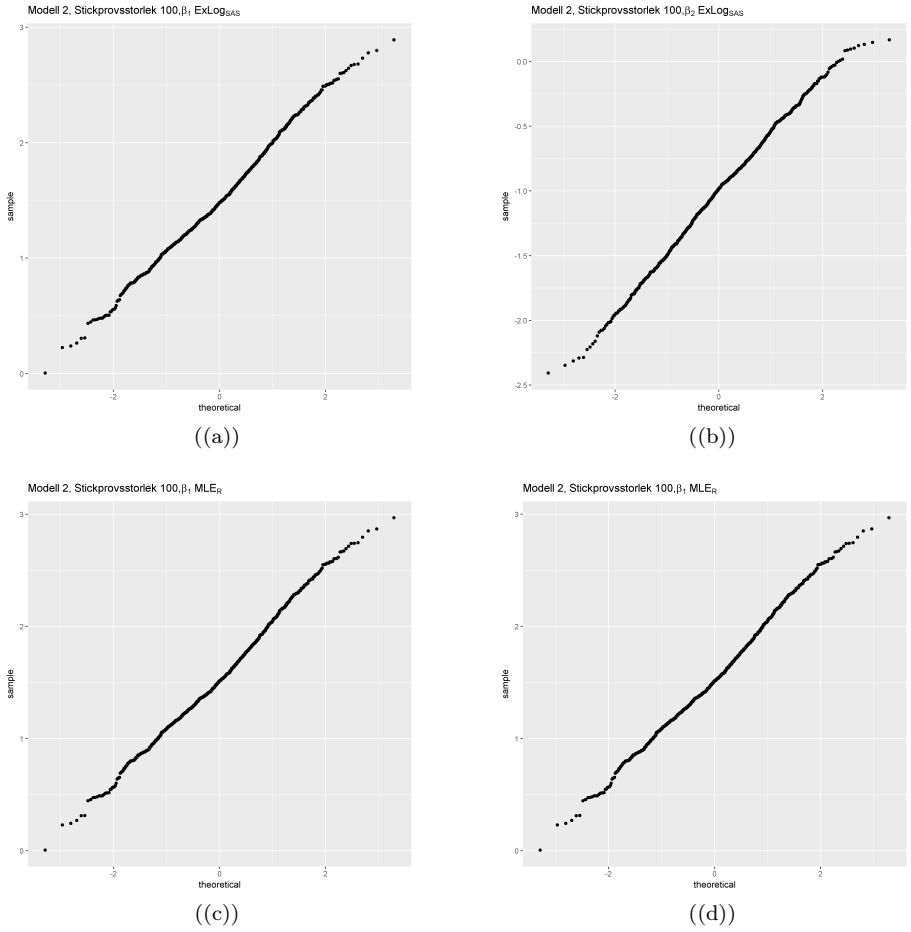
Stickprovsstorlek 50



((a)) ((b)) ((c)) ((d)) ((e)) ((f))

Figure 22: QQ-Plot för $ExLog_{SAS}$, figur ((a)) & ((b)), $B - Elrm_R$, figur ((c)) & ((d)), och MLE_R, figur ((e)) & ((f)), vid stickprovsstorlek 50, Modell 2.

Stickprovsstorlek 100



((a)) ((b)) ((c)) ((d))

Figure 23: QQ-Plot för $ExLog_{SAS}$, figur ((a)) & ((b)), MLE_R, figur ((c)) & ((d)), vid stickprovsstorlek 100, Modell 2.