Maksimum likelihood estimation af parametrene i logitmodellen med stokastiske individparametre
Et simulationsstudie

Jørgen Kai Olsen

Institut for Afsætningsøkonomi

Handelshøjskolen i København

2003

Indholdsfortegnelse

1. Indledning

I artiklen ”En stokastisk model for total og partiel kundeloyalitet” (Olsen 2003) har vi opstillet en generel logitmodel for forbrugerens købsadfærd og defineret begrebet kundeloyalitet – herunder specielt begreberne butiksloyalitet og mærkeloyalitet – ved hjælp af de elementer, der indgår i den opstillede model.

I nærværende artikel er problemstillingen

- at vise, hvorledes parametrene i denne model kan estimeres ved hjælp af maksimum likelihood metoden,
- at vise, hvorledes ML-estimationen rent teknisk kan gennemføres ved hjælp af et specielt udviklet estimationsprogram, samt
- at vise, at det som hovedregel har betydelige konsekvenser for ML-estimaterne, om estimationen gennemføres på basis af
    - en stikprøveplan, hvor der foreligger flere uafhængige, identisk fordelte observationer for hver respondent, eller på basis af
    - en stikprøveplan, hvor der kun foreligger én observation pr. respondent.

Med henblik på at simplificere analyserne nedenfor vil vi imidlertid ikke betragte den generelle model, der er opstillet i ovennævnte artikel, men kun et forholdsvis simpelt specialtilfælde heraf.

Endvidere vil vi basere estimationen af modellens parametre på en række simulerede datasæt (48 i alt). Dette indebærer den åbenbare fordel, at ”facit” for estimationen kendes a priori, hvilket muliggør en objektiv analyse af de enkelte stikprøveplaners effektivitet.

Endvidere udgør estimation baseret på simulerede data helt generelt en yderst effektiv metode til at vurdere det statistiske estimationsværktøj, man gør brug af i et konkret tilfælde i praksis.
Fx bliver man ved simulation i stand til at besvare følgende vigtige spørgsmål:
Giver det benyttede statistiske estimationsværktøj mulighed for en tilstrækkelig nøjagtig estimation af et planlagt (i praksis af og til forholdsvis stort) antal parametre med udgangspunkt i en planlagt (i praksis af og til forholdsvis lille) stikprøve.

2. Den generelle model

Overalt i det følgende vil vi betragte en virksomhed, der sælger et mærke – kaldet mærke A – i konkurrence med ét eller flere andre mærker, som betragtes under ét og kaldes for mærke B.

Mærket udbydes på et marked med N forbrugere i alt, der hver foretager et (stokastisk) antal køb af produktkategorien pr. periode (fx pr. år).

Ved hvert køb af produktkategorien er der en sandsynlighed – kaldet θ - for, at en given forbruger vælger mærke A. Denne sandsynlighed varierer i den opstillede model fra forbruger til forbruger og afhænger for enhver given forbruger og for enhver given periode af tre størrelser – nemlig

- af forbrugerens generelle loyalitet over for mærke A
  målt ved forbrugerens individuelle loyalitetsparameter β_{0i},
- af forbrugerens generelle vurdering af prisen for mærke A
  målt ved forbrugerens individuelle prisreaktionsparameter β_{1i}, og
- af den for samtlige forbrugere gældende pris for den k-te periode
  målt ved modellens eneste forklarende variabel $\Pr is_k$.

Lad Y_{ijk} (i = 1, 2, ... , N ; j = 1, 2, ... , n_{ik} ; k = 1, 2, ... , K) være en stokastisk indikatorvariabel, der er lig med 1, hvis den i-te forbruger vælger mærke A ved det j-te køb af produktkategorien i den k-te periode, og som er lig med 0 ellers. Lad endvidere $Y_{i1k}, Y_{i2k}, ..., Y_{in_{ik}k}$ være identisk fordelt. Og lad endelig $Y_{111}, ..., Y_{Nn_{NK}K}$ være stokastisk uafhængige.

Vi betragter da modellen

$$\theta_{ik} = P[Y_{ijk} = 1] = E(Y_{ijk}) = \frac{\exp(\beta_{0i} + \beta_{1i} \Pr is_k)}{1 + \exp(\beta_{0i} + \beta_{1i} \Pr is_k)}.$$

Denne model indeholder dobbelt så mange parametre, som der er forbrugere på markedet, hvilket gør den uanvendelig i praksis.

Vi lægger derfor en simpel struktur på de enkelte forbrugeres parameterværdisæt. Mere præcist vil vi antage, at samtlige forbrugeres værdisæt af de to parametre β_{0i} *og* β_{1i} (i = 1, 2, ... , N) er uafhængige, identisk fordelte realisationer af en stokastisk variabel, der er fordelt efter den todimensionale normale fordeling med middelværdivektoren (β_0 , β_1) og med kovariansmatricen

$$\Sigma = \begin{pmatrix} \sigma_0^{\,2} & \sigma_0\sigma_1\rho \\ \sigma_0\sigma_1\rho & \sigma_1^{\,2} \end{pmatrix}.$$

Den generelle model indeholder altså de 5 parametre $\beta_0, \beta_1, \sigma_0, \sigma_1$ *og* ρ, dvs. loyalitetsparameteren β_0, prisreaktionsparameteren β_1, de to standardafvigelser σ_0 *og* σ_1 og korrelationskoefficienten ρ.

Nedenfor vil vi betragte ML-estimationen for 8 specialtilfælde af denne model, som vi vil benævne modellerne 1-8. De 8 modeller er følgende:

Model 1. Loyalitetsparameteren β_0 og standardafvigelsen σ_0 indgår alene i modellen, og der foreligger m uafhængige, identisk fordelte observationer pr. respondent.

Model 2. Loyalitetsparameteren β_0 og standardafvigelsen σ_0 indgår alene i modellen, og der foreligger kun én observation pr. respondent.

Model 3. Prisreaktionsparameteren β_1 og standardafvigelsen σ_1 indgår alene i modellen, og der foreligger m uafhængige, identisk fordelte observationer pr. respondent.

Model 4. Prisreaktionsparameteren β_1 og standardafvigelsen σ_1 indgår alene i modellen, og der foreligger kun én observation pr. respondent.

Model 5. $\beta_0, \beta_1, \sigma_0$ *og* σ_1 indgår alene i modellen, og der foreligger m uafhængige, identisk fordelte observationer pr. respondent.

Model 6. $\beta_0, \beta_1, \sigma_0$ *og* σ_1 indgår alene i modellen, og der foreligger kun én observation pr. respondent.

Model 7. Alle 5 parametre $\beta_0, \beta_1, \sigma_0, \sigma_1$ *og* ρ indgår i modellen, og der foreligger m uafhængige, identisk fordelte observationer pr. respondent.

Model 8. Alle 5 parametre $\beta_0, \beta_1, \sigma_0, \sigma_1$ *og* ρ indgår i modellen, og der foreligger kun én observation pr. respondent.

For disse 8 modeller vil vi redegøre for simulationen af data og ML-estimationen af parametrene i de følgende afsnit.

3. Simulation af modellens data

Ved simulationen af de 8 modellers data vil vi (når parameterværdierne ikke eksplicit er sat lig med nul) antage, at de sande værdier af modellens parametre er

$$\begin{aligned}\beta_0 &= \ \ 2.50\\ \beta_1 &= -0.12\\ \sigma_0 &= \ \ 1.00\\ \sigma_1 &= \ \ 0.03 \ og\\ \rho &= \ \ 0.75.\end{aligned}$$

Vi vil endvidere simulere modellens data for K = 10 perioder ved følgende priser for mærke A:

Periode 1: Pris = 15 kr.
Periode 2: Pris = 18 kr.
Periode 3: Pris = 21 kr.
Periode 4: Pris = 24 kr.
Periode 5: Pris = 27 kr.
Periode 6: Pris = 30 kr.
Periode 7: Pris = 33 kr.
Periode 8: Pris = 36 kr.
Periode 9: Pris = 39 kr.
Periode 10: Pris = 42 kr.

Disse parameterværdier og priser er valgt, fordi vi forestiller os, at normalprisen for vare A er 30 kr. ved hvilken pris mærke A har en markedsandel på 25%. Og det gælder netop, at

$$\theta(30;2.50;-0.12)=\frac{\exp(2.50-0.12\cdot 30)}{1+\exp(2.50-0.12\cdot 30)}=\frac{\exp(-1.10)}{1+\exp(-1.10)}=0.25\,.^{1}$$

Endvidere forestiller vi os, at de variable enhedsomkostninger for mærke A er 15 kr. Dette betyder, at simulationen af modellens data foretages over et prisvariationsområde, som går fra de variable enhedsomkostninger til knap 3 gange de variable enhedsomkostninger. Dette variationsområde er valgt for at opnå stor sikkerhed på ML-estimaterne over modellens parametre. Men i praksis er det næppe muligt at arbejde med så stort et variationsområde for prisen. Ved det valgte variations-område for prisen varierer købssandsynligheden for mærke A (bestemt som nævnt i fodnoten) fra $\theta(15;2.50;-0.12)=0.67$ til $\theta(42;2.50;-0.12)=0.07$.

Standardafvigelserne $\sigma_0=1.00\ og\ \sigma_1=0.03$ er valgt forholdsvis store i forhold til regressionsparametrene $\beta_0=2.50\ og\ \beta_1=-0.12$ for at sikre, at der også reelt set er tale om heterogenitet mellem modellens forbrugere. Det er dog vigtigt at bemærke, at de nedenfor gennemførte analyser er kontrolleret meget omfattende ved andre parameterværdier – herunder ved mindre standardafvigelser – end de her nævnte. Ingen af disse analyser gav imidlertid anledning til væsentlig andre konklusioner om ML-estimaterne end de, der drages nedenfor.

Endelig er korrelationskoefficienten mellem den i-te forbrugers værdisæt af de to regressionsparametre $\beta_{0i}\ og\ \beta_{1i}$ sat til $\rho=0.75$. Når korrelationskoefficienten er forudsat positiv, skyldes det, at vi finder det realistisk at antage, at stor generel loyalitet over for mærke A hos den i-te forbruger $(\beta_{0i}>2.50)$ er knyttet sammen med lille følsomhed over for prisen for mærke A $(\beta_{1i}>-0.12)$. Og omvendt.

---

[1] Her og overalt i det følgende, hvor der beregnes en eksplicit talværdi for købssandsynligheden θ, har vi – for at simplificere beregningerne, men samtidig lidt upræcist - benyttet $\beta_0+\beta_1\,\mathrm{Pr}is_k$, dvs. forventningen af $\beta_{0i}+\beta_{1i}\,\mathrm{Pr}is_k$, som argument for den logistiske funktion for købssandsynligheden i stedet for at beregne den marginale købssandsynlighed $\overline{\theta}$ i hvert enkelt tilfælde.

For hver af de 8 specialtilfælde af den i forrige afsnit opstillede generelle model er modellens data simuleret for følgende 6 stikprøvestørrelser:

300 observationer
1000 observationer
3000 observationer
10000 observationer
30000 observationer og
100000 observationer.

Dette giver i alt 8 x 6 = 48 datasæt.

I de modeller (modellerne 1 og 2), hvor prisen ikke indgår som forklarende variabel, har vi – som en ren regneteknisk foranstaltning – sat $\beta_{1i} = \beta_1 = -0.12$ for samtlige forbrugere og simuleret alle data for de 6 stikprøvestørrelser ved normalprisen 30 kr. Dette har vi valgt at gøre for at sikre, at niveauet for købssandsynligheden θ fortsat er 0.25. Alternativt kunne vi have valgt at sætte $\beta_1 = 0$ og at ændre loyalitetsparameteren fra 2.50 til –1.10, idet $\theta(30; 2.50; -0.12) = \theta(-1.10) = 0.25$. Men uanset simulationsmetoden er det vigtigt at bemærke, at det kun er parametrene β_0 *og* σ_0, der estimeres i modellerne 1 og 2. ($\beta_1 = -0.12$ indgår altså som en konstant i estimationsprogrammet).

I de modeller (modellerne 3 og 4), hvor loyalitetsparameteren ikke indgår i problemstillingen, har vi – igen som en ren regneteknisk foranstaltning – sat $\beta_{0i} = \beta_0 = 2.50$ for samtlige forbrugere og simuleret alle data med denne konstante hjælpevariabel. Dette har vi valgt at gøre for at sikre, at niveauet for købssandsynligheden θ fortsat er 0.25 ved normalprisen 30 kr. Alternativt kunne vi have valgt at sætte $\beta_0 = 0$ og at ændre prisreaktionsparameteren fra –0.12 til –0.0367 (–1.10/30). Denne løsning er dog utilfredsstillende, fordi den ikke giver den ønskede spredning på θ-værdierne ved de øvrige priser. Men uanset simulationsmetoden er det vigtigt at bemærke, at det kun er parametrene β_1 *og* σ_1, der estimeres i modellerne 3 og 4. ($\beta_0 = 2.50$ indgår altså som en konstant i estimationsprogrammet).

I de modeller (modellerne 3 - 8), hvor prisen for mærke A indgår som egentlig forklarende variabel for købssandsynligheden, er der altid simuleret lige mange observationer fra hver af de 10 perioder.

I de modeller (modellerne 1, 3, 5 og 7), hvor der foreligger flere uafhængige, identisk fordelte observationer pr. respondent, er der altid simuleret lige mange observationer pr. respondent – nemlig m = 10 i alt. For en given respondent er disse 10 observationer altid simuleret fra den samme periode. (Men periodens nummer varierer naturligvis fra respondent til respondent).
Det er muligt, at der ville kunne opnås større sikkerhed på ML-estimaterne over modellens parametre, hvis de 10 observationer pr. respondent havde været simuleret med én observation pr. periode. Men en analyse af dette forhold ligger (ligesom en analyse af gentagelsernes antal) uden for artiklens hovedproblemstilling, der som nævnt i indledningen er at analysere forskellen på nøjagtigheden af parameterestimaterne henholdsvis med og uden gentagelser pr. respondent.

For hver af de 10 perioder, er der simuleret købsdata for n forskellige respondenter i alt.
I de stikprøveplaner, hvor der foreligger 10 observationer pr. respondent, er n derfor lig med den totale stikprøvestørrelse (fx 1000) divideret med 100 (10 x 10). I de stikprøveplaner, hvor der kun foreligger én observation pr. respondent, er n derimod lig med den totale stikprøvestørrelse (fx 1000) divideret med 10.

For et givet af de 48 datasæt er selve simulationen af modellens data gennemført på følgende måde:

Lad U_{ik} (i = 1, 2, ..., n ; k = 1, 2, ..., K) være en endimensional stokastisk variabel, der er defineret således:

$$U_{ik} = \beta_{0i} + \beta_{1i} \Pr is_k .$$

Da følger det af den ovenfor specificerede forudsætning om fordelingen af regressionsparametrene β_{0i} *og* β_{1i}, at U_{ik} er fordelt efter den endimensionale normale fordeling med middelværdien

$$\mu_k = \beta_0 + \beta_1 \Pr is_k$$

og variansen

$$\tau_k^{\,2} = \sigma_0^{\,2} + \sigma_1^{\,2}\,\mathrm{Pr}\,is_k^{\,2} + 2\sigma_0\sigma_1\rho\,\mathrm{Pr}\,is_k .$$

Endvidere er den i-te forbrugers købssandsynlighed for mærke A i periode nummer k en stokastisk variabel, θ_{ik} , der er defineret således:

$$\theta_{ik} = \frac{\exp(\beta_{0i} + \beta_{1i}\,\mathrm{Pr}\,is_k)}{1+\exp(\beta_{0i} + \beta_{1i}\,\mathrm{Pr}\,is_k)} = \frac{\exp(U_{ik})}{1+\exp(U_{ik})}.$$

Simulationen er derfor foretaget således, at vi for hver af de n respondenter, der er til rådighed i en given periode, og for hver af de 10 perioder har genereret enten m = 10 uafhængige, identisk fordelte observationer af U_{ik} (for de stikprøveplaner, hvor der foreligger gentagelser) eller en enkelt observation af U_{ik} (for de stikprøveplaner, hvor der kun foreligger én observation pr. respondent). Herefter er θ_{ik} beregnet. Endelig er selve indikatorvariablen Y_{ijk} for valg af mærke A simuleret således, at Y_{ijk} er sat lig med 1, hvis et genereret tilfældigt tal R_{ijk}, der er rektangulært fordelt mellem 0 og 1, er mindre end θ_{ik} , og er sat lig med 0 ellers.

Det er vigtigt at bemærke, at simulationen er foretaget vha. en såkaldt tilfældighedsgenerator (programmeret i Pascal), der som bekendt kun frembringer pseudotilfældige tal.

Men på den anden side er det (vha. U-tests og χ^2 -tests) for hver stikprøve undersøgt, om de simulerede data virker rimeligt tilfældige. Konklusionen på disse analyser er, at der ikke er noget der tyder på, at simulationen af modellens data er uacceptabel.

De forholdsvis få talmaterialer med lav signifikanssandsynlighed er da også bibeholdt i analysen, fordi de ved ML-estimationen senere hen ikke giver anledning til "besynderlige" parameter-estimater.

4. Likelihoodfunktionen

Vi vil opstille likelihoodfunktionen og maksimere denne for den generelle model, hvor samtlige 5 parametre $\beta_0, \beta_1, \sigma_0, \sigma_1$ *og* ρ indgår, og hvor der foreligger m = 10 observationer pr. respondent.

I dette og i næste afsnit er det nødvendigt at sondre mellem det generelle og det sande parameterpunkt i parameterrummet. Disse to parameterpunkter vil vi henholdsvis kalde for $(\beta_0, \beta_1, \sigma_0, \sigma_1, \rho)$ og $(\beta_{00}, \beta_{01}, \sigma_{00}, \sigma_{01}, \rho_0)$. Når denne sondring ikke længere er nødvendig, går vi tilbage til vor oprindelige betegnelse for de sande parametre. (Uden det ekstra fodtegn nul).

Lad Y_{ijk} (i = 1, 2, ..., n ; j = 1, 2, ..., m ; k = 1, 2, ..., K) være den ovenfor definerede indikatorvariabel, der er lig med 1, hvis den i-te respondent vælger mærke A ved det j-te køb af produktkategorien i periode nummer k, og som er lig med 0 ellers.
(For en given respondent antager k som nævnt ovenfor kun én af de 10 mulige værdier).

Da er den betingede fordeling af Y_{ijk} givet værdisættet (β_{0i}, β_{1i}) af den i-te respondents regressionsparametre bestemt ved, at

$$\theta_{ik} = P[Y_{ijk} = 1] = \frac{\exp(\beta_{0i} + \beta_{1i} \operatorname{Pr} is_k)}{1 + \exp(\beta_{0i} + \beta_{1i} \operatorname{Pr} is_k)}.$$

Derfor er den betingede sandsynlighed for observationssættet ($y_{i1k}, y_{i2k}, ..., y_{imk}$)

$$f_i(y_{i1k}, y_{i2k}, ..., y_{imk}) = \prod_{j=1}^{m} \theta_{ik}^{\,y_{ijk}} (1-\theta_{ik})^{1-y_{ijk}} = \theta_{ik}^{\,y_{i.k}} (1-\theta_{ik})^{m-y_{i.k}},$$

hvor $y_{i.k} = \sum_{j=1}^{m} y_{ijk}$, og hvor f_i afhænger af β_{0i} *og* β_{1i}.

Idet θ_{ik} er en funktion af de to stokastiske regressionsparametre β_{0i} *og* β_{1i}, følger det, at den marginale fordeling af observationssættet $(y_{i1k}, y_{i2k}, ..., y_{imk})$ bliver

$$f(y_{i1k}, y_{i2k}, ..., y_{imk}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} {\theta_{ik}}^{y_{i.k}} (1-\theta_{ik})^{m-y_{i.k}} \, g_0(\beta_{0i}, \beta_{1i}) \, d\beta_{0i} \, d\beta_{1i}$$

$$= E_0[{\theta_{ik}}^{y_{i.k}} (1-\theta_{ik})^{m-y_{i.k}}],$$

hvor g_0 er sandsynlighedstætheden for den todimensionale normale fordeling af β_{0i} *og* β_{1i} taget i det sande parameterpunkt, og hvor E_0 betegner forventningsoperatoren, igen taget i det sande parameterpunkt.

Likelihoodfunktionen $L = L(\beta_0, \beta_1, \sigma_0, \sigma_1, \rho)$ bliver derfor

$$L = \prod_{k=1}^{K} \prod_{i=1}^{n} E[{\theta_{ik}}^{y_{i.k}} (1-\theta_{ik})^{m-y_{i.k}}],$$

hvor forventningsoperatoren E er defineret som forventningsoperatoren E_0, bortset fra, at det sande parameterpunkt $(\beta_{00}, \beta_{01}, \sigma_{00}, \sigma_{01}, \rho_0)$ i sandsynlighedstætheden g_0 nu er erstattet af det generelle parameterpunkt $(\beta_0, \beta_1, \sigma_0, \sigma_1, \rho)$ i den tilsvarende generelle sandsynlighedstæthed g.

Heraf følger, at logaritmen af likelihoodfunktionen bliver

$$Ln(L) = \sum_{k=1}^{K} \sum_{i=1}^{n} Ln(E[{\theta_{ik}}^{y_{i.k}} (1-\theta_{ik})^{m-y_{i.k}}]),$$

som vi vil maksimere approksimativt i næste afsnit.

5. Maksimering af likelihoodfunktionen

Udtrykket

$$E\left[\theta_{ik}^{\ y_{i.k}}(1-\theta_{ik})^{m-y_{i.k}}\right],$$

hvor

$$\theta_{ik} = \frac{\exp(\beta_{0i} + \beta_{1i}\,\mathrm{Pr}\,is_k)}{1+\exp(\beta_{0i} + \beta_{1i}\,\mathrm{Pr}\,is_k)},$$

der indgår i logaritmen af den i forrige afsnit opstillede likelihoodfunktion, kan ikke bestemmes eksplicit ad analytisk vej. Derimod kan det bestemmes approksimativt ved Monte Carlo simulation, idet der netop er tale om en forventning, der kan approksimeres ved et gennemsnit.
(Se fx Ross [2000], kap. 11).

I det følgende vil vi imidlertid benytte en anden approksimation, der bygger på numerisk integration og indebærer, at likelihoodfunktionen og logaritmen af likelihoodfunktionen kan opskrives eksplicit og dermed forholdsvis let kan maksimeres. Ideen til denne approksimation skyldes Tue Tjur.

Lad U_{ik} (i = 1, 2, ..., n ; k = 1, 2, ..., K) være den i afsnit 3 indførte endimensionale stokastiske variabel, der er defineret således:

$$U_{ik} = \beta_{0i} + \beta_{1i}\,\mathrm{Pr}\,is_k\,.$$

Da følger det af forudsætningen om fordelingen af regressionsparametrene $\beta_{0i}\ og\ \beta_{1i}$, at U_{ik} er fordelt efter den endimensionale normale fordeling med middelværdien

$$\mu_k = \beta_0 + \beta_1\,\mathrm{Pr}\,is_k$$

og variansen

$$\tau_k^{\,2} = \sigma_0^{\,2} + \sigma_1^{\,2}\,\mathrm{Pr}is_k^{\,2} + 2\sigma_0\sigma_1\rho\,\mathrm{Pr}is_k .$$

Endvidere er den i-te forbrugers købssandsynlighed for mærke A i periode nummer k en stokastisk variabel, θ_{ik}, der er defineret således:

$$\theta_{ik} = \frac{\exp(\beta_{0i} + \beta_{1i}\,\mathrm{Pr}is_k)}{1+\exp(\beta_{0i} + \beta_{1i}\,\mathrm{Pr}is_k)} = \frac{\exp(U_{ik})}{1+\exp(U_{ik})},$$

og som dermed alene afhænger af U_{ik}. Dette betyder, at udtrykket

$$E\,[\theta_{ik}^{\,y_{i.k}}(1-\theta_{ik})^{m-y_{i.k}}]$$

kan bestemmes vha. den endimensionale normale fordeling.

Lad endvidere V_{ik} være en endimensional stokastisk variabel, der er defineret således:

$$V_{ik} = \frac{U_{ik} - \mu_k}{\tau_k}.$$

Da er V_{ik} fordelt efter den endimensionale standardiserede normale fordeling, og

$$U_{ik} = \mu_k + \tau_k V_{ik}.$$

Dette betyder, at udtrykket

$$E\,[\theta_{ik}^{\,y_{i.k}}(1-\theta_{ik})^{m-y_{i.k}}] = E[(\frac{\exp(\mu_k + \tau_k V_{ik})}{1+\exp(\mu_k + \tau_k V_{ik})})^{y_{i.k}}\ (1-\frac{\exp(\mu_k + \tau_k V_{ik})}{1+\exp(\mu_k + \tau_k V_{ik})})^{m-y_{i.k}}]$$

kan bestemmes vha. den endimensionale standardiserede normale fordeling.

Men udtrykket kan stadig ikke bestemmes eksplicit.

Vi approksimerer derfor den standardiserede normale fordeling ved en diskret stokastisk variabel Z med s lige sandsynlige udfald $z_1, z_2, ..., z_s$, altså med

$$P[Z = z_r] = \frac{1}{s},$$

hvor z_r er defineret således:

$$z_r = \Phi^{-1}(\frac{1}{2s} + \frac{r-1}{s}) \quad ; \quad r = 1, 2, ..., s,$$

hvor Φ^{-1} er den inverse af fordelingsfunktionen for den standardiserede normale fordeling, og hvor s er et stort (ulige) positivt helt tal. (I estimationsprogrammet, der omtales nedenfor, har det vist sig tilstrækkeligt at sætte s = 101, men i andre sammenhænge har vi også benyttet s = 6001).

Variationsområdet for Φ^{-1} er altså (her) de s punkter fra 1/2s til 1-1/2s. Og når s er ulige, bliver medianen i fordelingen af Z lig med $z_{(s+1)/2} = \Phi^{-1}(½) = 0$.

Med denne approksimation bliver logaritmen af likelihoodfunktionen

$$Ln(L) = \sum_{k=1}^{K}\sum_{i=1}^{n} Ln(E\,[{\theta_{ik}}^{y_{i.k}}(1-\theta_{ik})^{m-y_{i.k}}])$$

$$\approx \sum_{k=1}^{K}\sum_{i=1}^{n} Ln(\frac{1}{s}\sum_{r=1}^{s}\{(\frac{\exp(\mu_k + \tau_k z_r)}{1+\exp(\mu_k + \tau_k z_r)})^{y_{i.k}}\ (1-\frac{\exp(\mu_k + \tau_k z_r)}{1+\exp(\mu_k + \tau_k z_r)})^{m-y_{i.k}}\}).$$

Da dette udtryk kan beregnes eksplicit, har vi udviklet et specielt estimationsprogram, der (med rimelig nøjagtighed) maksimerer udtrykket mht. det generelle parameterpunkt $(\beta_0, \beta_1, \sigma_0, \sigma_1, \rho)$, og dermed giver os de approksimative maksimum likelihood estimater $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_0, \hat{\sigma}_1, \hat{\rho})$.

Det skal bemærkes, at den benyttede approksimative metode til bestemmelse af ML-estimaterne, der som nævnt bygger på numerisk integration, i virkeligheden er en "styret" diskret Monte Carlo simulationsmetode.

De s uafhængige og identisk fordelte observationer af den stokastiske variabel Z simuleres nemlig ikke - som ved kontinuert Monte Carlo simulation af s standardiserede normalt fordelte observationer - fra den rektangulære fordeling på intervallet mellem 0 og 1, men fra den diskrete ligefordeling i punkterne 1/2s, 1/2s + 1/s, ..., 1-1/2s. (I begge tilfælde ved benyttelse af funktionen Φ^{-1}). Men til gengæld "styres" simulationen, således at hver af de s mulige værdier af Z, $z_1, z_2, ..., z_s$, indgår i simulationen én og kun én gang.

I næste afsnit vil vi vise, hvorledes det specielt udviklede estimationsprogram kan benyttes til at estimere parametrene i de i afsnit 2 opstillede 8 modeller.

I denne forbindelse er det vigtigt at bemærke, at estimationsprogrammet er kontrolleret (mange gange) ved at simulere en række datamaterialer, hvor de to standardafvigelser σ_0 *og* σ_1 og (derfor) korrelationskoefficienten ρ alle er lig med 0. I så fald bliver modellen en sædvanlig logistisk regressionsanalysemodel med eller uden gentagelser pr. respondent.

For disse talmaterialer har vi herefter estimeret loyalitetsparameteren β_0 og prisreaktionsparameteren β_1, dels vha. det specielt udviklede estimationsprogram, dels vha. den statistiske programpakke ISUW (der er udviklet af Tue Tjur).

I alle de betragtede tilfælde (dvs. for de sædvanlige 6 stikprøvestørrelser hhv. med og uden gentagelser pr. respondent) fik vi præcis de samme ML-estimater ved de to forskellige estimationsmetoder.

I øvrigt er det en ekstra kontrol af det udviklede estimationsprogram, at det nedenfor viser sig at give de generelle resultater, man ville forvente a priori.

6. ML-estimation af parametrene i de 8 modeller

6.1 Model 1

I denne model er interesseparametrene loyalitetsparameteren β_0 og standardafvigelsen σ_0. Dette betyder, at modellen for købssandsynligheden for mærke A for alle respondenter (i = 1, 2, …, n), for alle uafhængige, identisk fordelte gentagelser (j = 1, 2, …, m) og for alle perioder (k = 1, 2, …, K), er

$$\theta_i = \frac{\exp(\beta_{0i} - 0.12 \cdot 30)}{1 + \exp(\beta_{0i} - 0.12 \cdot 30)},$$

hvor vi, som nævnt i afsnit 3, alene har medtaget konstanten $-0.12 \cdot 30$ for at sikre, at købssandsynligheden er på niveauet 0.25, og hvor standardafvigelsen σ_0 indgår i normalfordelingen for β_{0i}.

For denne model har så simuleret de 6 stikprøvestørrelser, der er omtalt i afsnit 3 (dvs. i alt hhv. 300, 1000, 3000, 10000, 30000 og 100000 observationer). Endelig har vi benyttet det i forrige afsnit omtalte specielt udviklede estimationsprogram til at estimere parametrene β_0 *og* σ_0.

Resultatet af den approksimative ML-estimation fremgår af tabel 1.

Tabel 1: **ML-estimation af β_0 og σ_0**

10 observationer pr. respondent

	Beta.0	Beta.1	Sigma.0	Sigma.1	Rho	-2*Ln(L)
Sand Model	**2.50**	**(-0.12)**	**1.00**	**0.00**	**0.00**	**-----**
300	**2.38**	**-----**	**0.79**	**-----**	**-----**	**330.02**
300	**2.38**	**-----**	**0.79**	**-----**	**-----**	**330.02**
1000	**2.41**	**-----**	**0.82**	**-----**	**-----**	**1111.51**
1000	**2.41**	**-----**	**0.82**	**-----**	**----**	**1111.51**
3000	**2.40**	**-----**	**0.93**	**-----**	**-----**	**3323.85**
3000	**2.40**	**-----**	**0.93**	**-----**	**-----**	**3323.85**
10000	**2.53**	**-----**	**0.93**	**-----**	**-----**	**11466.01**
10000	**2.53**	**-----**	**0.93**	**-----**	**-----**	**11466.01**
30000	**2.47**	**-----**	**1.01**	**-----**	**-----**	**33859.02**
30000	**2.47**	**-----**	**1.01**	**-----**	**-----**	**33859.02**
100000	**2.50**	**-----**	**1.01**	**-----**	**-----**	**113728.10**
100000	**2.50**	**-----**	**1.01**	**-----**	**-----**	**113728.10**

I tabel 1 (og i alle tabeller nedenfor) viser første række den sande model, dvs. den model, der ligger til grund for simulationen af data.

Endvidere er der for hver stikprøvestørrelse angivet to ML-estimater og to værdier af minus 2 gange logaritmen af likelihoodfunktionen. Det første sæt af disse tal er altid det resultat, der opnås, når iterationen i det specielt udviklede estimationsprogram startes i de sande parameter-værdier, medens det andet sæt tal altid er det resultat, der opnås, når iterationen startes i et andet (tilfældigt valgt) parameterpunkt.

I de modeller nedenfor, hvor der opstår problemer med ML-estimaterne, er der af kontrolhensyn altid afprøvet flere forskellige tilfældigt valgte startpunkter for iterationen. De forskellige kørsler af estimationsprogrammet fører dog altid til de samme generelle konklusioner.

Af tabel 1 fremgår det,

- at ML-estimaterne er de samme, uanset om iterationen startes i det sande parameterpunkt eller i et andet parameterpunkt,
- at –2*Ln(L) antager den samme værdi, uanset om iterationen startes i det sande parameterpunkt eller i et andet parameterpunkt, og
- at ML-estimaterne ligger tættere og tættere ved de sande parameterværdier, efterhånden som stikprøven vokser.

Konklusionen er derfor, at model 1 – dels fordi den er simpel, men specielt fordi der foreligger flere uafhængige, identisk fordelte observationer pr. respondent - ikke giver anledning til estimations-problemer.

6.2 Model 2

I denne model er problemstillingen, modelstrukturen, parameterværdierne og stikprøvestørrelserne ganske som i model 1. Den eneste forskel er, at der nu ikke foreligger m = 10 observationer pr. respondent, men kun én observation pr. respondent.
(Antallet af forskellige respondenter er derfor 10 gange så stort som i model 1).

For denne model fremgår ML-estimaterne for β_0 *og* σ_0 af tabel 2.

Tabel 2: **ML-estimation af β_0 og σ_0**

1 observation pr. respondent

	Beta.0	Beta.1	Sigma.0	Sigma.1	Rho	-2*Ln(L)
Sand Model	**2.50**	**(-0.12)**	**1.00**	**0.00**	**0.00**	**-----**
300	**2.47**	**-----**	**0.99**	**-----**	**-----**	**355.77**
300	**1.01**	**-----**	**4.07**	**-----**	**-----**	**355.77**
1000	**2.56**	**-----**	**1.02**	**-----**	**-----**	**1218.32**
1000	**0.94**	**-----**	**4.69**	**-----**	**----**	**1218.32**
3000	**2.50**	**-----**	**1.00**	**-----**	**-----**	**3593.02**
3000	**2.62**	**-----**	**0.58**	**-----**	**-----**	**3593.02**
10000	**2.48**	**-----**	**0.99**	**-----**	**-----**	**11898.52**
10000	**-1.42**	**-----**	**8.52**	**-----**	**-----**	**11898.52**
30000	**2.50**	**-----**	**1.00**	**-----**	**-----**	**35858.67**
30000	**-0.50**	**-----**	**6.99**	**-----**	**-----**	**35858.67**
100000	**2.50**	**-----**	**1.00**	**-----**	**-----**	**119576.08**
100000	**-2.58**	**-----**	**10.75**	**-----**	**-----**	**119576.08**

Denne tabel viser tydeligt, at der – ved anvendelse af klassisk maksimum likelihood estimation – opstår betydelige problemer, når der kun foreligger én observation pr. respondent.

- Når iterationen i det specielt udviklede estimationsprogram startes i de sande parameterværdier, finder estimationsprogrammet – selv for forholdsvis små stikprøvestørrelser – frem til nogenlunde acceptable estimater.
- Derimod opnås der – trods ganske samme værdi af –2*Ln(L) – helt misvisende resultater, når iterationen startes i et andet parameterpunkt.
  Og resultaterne forbedres ikke, når stikprøvestørrelsen vokser.

Forklaringen på dette forhold er, at likelihoodfunktionen er overparametriseret.
Datamaterialet indeholder simpelt hen ikke information nok til en entydig estimation af de to parametre β_0 *og* σ_0.

Vi vil illustrere dette for stikprøven på 1000 observationer. For denne stikprøve er den eneste information, man har til rådighed, når man skal estimere de to parametre β_0 *og* σ_0, at 298 af de 1000 respondenter har valgt mærke A, medens de resterende 702 respondenter har valgt et andet mærke (mærke B).[2] Det er oplagt, at det er realistisk at antage, at de 298 respondenter, der har valgt mærke A, ikke alle har samme præference (loyalitet) over for mærke A. Der kan endog være tale om betydelig variation i de 298 respondenters præference. Men for to givne respondenter P og Q råder vi alene over den information, at de begge har købt produktkategorien én gang, og at de begge har valgt mærke A. Hvor forskellig P og Q's præference for mærke A er, kan vi naturligvis ikke sige noget om på det foreliggende datagrundlag.

På tilsvarende måde er det oplagt, at det er realistisk at antage, at de 702 respondenter, der har valgt et andet mærke end mærke A, ikke alle har samme præference over for mærke A. Men for to andre givne respondenter R og S råder vi alene over den information, at de begge har købt produktkategorien én gang, og at de begge har valgt et andet mærke end mærke A. Hvor forskellig R og S's præference for mærke A er, kan vi naturligvis heller ikke sige noget.

Disse kendsgerninger resulterer da også i, at den betragtede likelihoodfunktion antager sin maksimale værdi (som er konstant) for uendelig mange kombinationer af de to parametre β_0 *og* σ_0. For nogle få eksempler fremgår dette af følgende tabel:

Beta.0	**Sigma.0**	**-2*Ln(L)**
0.00	**6.55**	**1218.32**
0.50	**5.57**	**1218.32**
1.00	**4.57**	**1218.32**
1.50	**3.54**	**1218.32**
2.00	**2.47**	**1218.32**
2.50	**1.22**	**1218.32**

Problemet svarer i virkeligheden til at løse ligningen x * y = 60 mht. x og y.

[2] Når den relative hyppighed for valg af mærke A er 0.298, dvs. større end 0.25, skyldes det, at den marginale købssandsynlighed - med de valgte parameterværdier - er større end 0.25 – nemlig lig med 0.286.

Konklusionen på analysen af model 2 er derfor, at der – selv i denne ganske simple model med kun to parametre β_0 *og* σ_0 – opstår estimationsproblemer, fordi der kun foreligger én observation pr. respondent.

6.3 Model 3

I model 3 – hvor der igen foreligger m = 10 uafhængige, identisk fordelte observationer pr. respondent - er problemstillingen ændret, idet prisen for mærke A nu inddrages som egentlig forklarende variabel for købssandsynligheden θ.

I denne model er interesseparametrene prisreaktionsparameteren β_1 og standardafvigelsen σ_1. Når loyalitetsparameteren $\beta_0 = 2.50$ på trods heraf alligevel inddrages i modellen (som en konstant, der er den samme for alle respondenter), er det derfor alene for at sikre, at købssandsynligheden ved normalprisen 30 kr. er på niveauet 0.25.

Modellen for købssandsynligheden for mærke A er altså

$$\theta_{ik} = \frac{\exp(2.50 + \beta_{1i} \Pr is_k)}{1 + \exp(2.50 + \beta_{1i} \Pr is_k)},$$

hvor σ_1 indgår i normalfordelingen for β_{1i}.

For denne model fremgår ML-estimaterne for β_1 *og* σ_1 af tabel 3.

Tabel 3: **ML-estimation af β_1 og σ_1**

10 observationer pr. respondent

	Beta.0	Beta.1	Sigma.0	Sigma.1	Rho	-2*Ln(L)
Sand Model	**(2.50)**	**-0.12**	**0.00**	**0.03**	**0.00**	**-----**
300	**-----**	**-0.11**	**-----**	**0.03**	**-----**	**361.29**
300	**-----**	**-0.11**	**-----**	**0.03**	**-----**	**361.29**
1000	**-----**	**-0.13**	**-----**	**0.03**	**-----**	**1016.10**
1000	**-----**	**-0.13**	**-----**	**0.03**	**----**	**1016.10**
3000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**3222.14**
3000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**3222.14**
10000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**11199.62**
10000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**11199.62**
30000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**33135.54**
30000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**33135.54**
100000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**110223.48**
100000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**110223.48**

Af tabellen fremgår det,

- at ML-estimaterne er de samme, uanset om iterationen startes i det sande parameterpunkt eller i et andet parameterpunkt,
- at –2*Ln(L) antager den samme værdi, uanset om iterationen startes i det sande parameterpunkt eller i et andet parameterpunkt, og
- at ML-estimaterne ligger endog meget tæt ved de sande parameterværdier, selv for de forholdsvis små stikprøver.

Konklusionen er derfor, at model 3 – dels fordi den er simpel, dels fordi der foreligger flere uafhængige, identisk fordelte observationer pr. respondent - ikke giver anledning til estimationsproblemer.

6.4 Model 4

I denne model er problemstillingen, modelstrukturen, parameterværdierne og stikprøvestørrelserne ganske som i model 3. Den eneste forskel er, at der nu ikke foreligger m = 10 observationer pr. respondent, men kun én observation pr. respondent.

For denne model fremgår ML-estimaterne for β_1 *og* σ_1 af tabel 4.

Tabel 4: ML-estimation af β_1 *og* σ_1

1 observation pr. respondent

	Beta.0	Beta.1	Sigma.0	Sigma.1	Rho	-2*Ln(L)
Sand Model	**(2.50)**	**-0.12**	**0.00**	**0.03**	**0.00**	**-----**
300	**-----**	**-0.12**	**-----**	**0.00+**	**-----**	**319.97**
300	**-----**	**-0.12**	**-----**	**0.00+**	**-----**	**319.97**
1000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**1135.58**
1000	**-----**	**-0.12**	**-----**	**0.03**	**----**	**1135.58**
3000	**-----**	**-0.12**	**-----**	**0.02**	**-----**	**3352.70**
3000	**-----**	**-0.12**	**-----**	**0.02**	**-----**	**3352.70**
10000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**11572.94**
10000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**11572.94**
30000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**34408.77**
30000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**34408.77**
100000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**114125.42**
100000	**-----**	**-0.12**	**-----**	**0.03**	**-----**	**114125.42**

Tabel 4 viser,

- at ML-estimaterne for alle 4 parametre er de samme, uanset om iterationen startes i det sande parameterpunkt eller i et andet parameterpunkt,
- at –2*Ln(L) antager den samme værdi, uanset om iterationen startes i det sande parameterpunkt eller i et andet parameterpunkt, og

- at ML-estimaterne – bortset fra estimatet for σ_1, der er udartet i nul i den mindste stikprøve - ligger endog meget tæt ved de sande parameterværdier for enhver af de betragtede stikprøvestørrelser.

Konklusionen på analysen af model 4 er derfor, at der ikke opstår problemer med ML-estimationen, på trods af at der kun foreligger én observation pr. respondent. Dette skyldes dels, at modellen er simpel, men specielt at der nu – til forskel fra model 2 – indgår en egentlig forklarende variabel på 10 niveauer i modellen. Dette medfører en væsentlig forøgelse af informationen i de foreliggende data i forhold til model 2.

6.5 Model 5

I model 5 – hvor der foreligger m uafhængige, identisk fordelte observationer pr. respondent – indgår loyalitetsparameteren β_0 og prisreaktionsparameteren β_1 sammen med standardafvigelserne σ_0 og σ_1 på helt normal vis i modellen. Den eneste forskel mellem denne model og den generelle model, der behandles nedenfor, er, at korrelationskoefficienten ρ mellem de individuelle regressionsparametre β_{0i} *og* β_{1i} er forudsat lig med 0. Modellen for den betingede købssandsynlighed for mærke A er altså

$$\theta_{ik} = \frac{\exp(\beta_{0i} + \beta_{1i} \Pr is_k)}{1 + \exp(\beta_{0i} + \beta_{1i} \Pr is_k)},$$

og resultatet af ML-estimationen af denne models 4 parametre fremgår for de 6 simulerede stikprøver af tabel 5.

Tabel 5: **ML-estimation af $\beta_0, \beta_1, \sigma_0 \, og \, \sigma_1$**

10 observationer pr. respondent

	Beta.0	Beta.1	Sigma.0	Sigma.1	Rho	-2*Ln(L)
Sand Model	**2.50**	**-0.12**	**1.00**	**0.03**	**0.00**	**-----**
300	**2.61**	**-0.12**	**0.00+**	**0.05**	**-----**	**340.34**
300	**2.61**	**-0.12**	**0.00+**	**0.05**	**-----**	**340.34**
1000	**1.72**	**-0.10**	**0.90**	**0.05**	**-----**	**1062.95**
1000	**1.72**	**-0.10**	**0.90**	**0.05**	**-----**	**1062.95**
3000	**2.73**	**-0.13**	**1.38**	**0.02**	**-----**	**3099.43**
3000	**2.73**	**-0.13**	**1.38**	**0.02**	**-----**	**3099.43**
10000	**2.64**	**-0.13**	**0.91**	**0.03**	**-----**	**10709.27**
10000	**2.64**	**-0.13**	**0.91**	**0.03**	**-----**	**10709.27**
30000	**2.45**	**-0.12**	**0.96**	**0.03**	**-----**	**32287.77**
30000	**2.45**	**-0.12**	**0.96**	**0.03**	**-----**	**32287.77**
100000	**2.47**	**-0.12**	**1.00**	**0.03**	**-----**	**108303.34**
100000	**2.47**	**-0.12**	**1.00**	**0.03**	**-----**	**108303.34**

Af tabellen fremgår det,

- at ML-estimaterne for alle 4 parametre er de samme, uanset om iterationen startes i det sande parameterpunkt eller i et andet parameterpunkt,
- at –2*Ln(L) antager den samme værdi, uanset om iterationen startes i det sande parameterpunkt eller i et andet parameterpunkt, og
- at ML-estimaterne ligger tættere og tættere ved de sande parameterværdier, efterhånden som stikprøven vokser.

Konklusionen er derfor, at model 5 ikke giver anledning til estimationsproblemer. Dette skyldes hovedsagelig, at der foreligger flere uafhængige, identisk fordelte observationer pr. respondent, og ikke - som ovenfor - at modellen er meget simpel. Thi nu indgår der 4 mod hidtil kun 2 parametre i modellen.

6.6 Model 6

I denne model er problemstillingen, modelstrukturen, parameterværdierne og stikprøvestørrelserne ganske som i model 5. Den eneste forskel er, at der nu ikke foreligger m = 10 observationer pr. respondent, men kun én observation pr. respondent.

For denne model fremgår ML-estimaterne for $\beta_0, \beta_1, \sigma_0 \, og \, \sigma_1$ af tabel 6.

Tabel 6: ML-estimation af $\beta_0, \beta_1, \sigma_0 \, og \, \sigma_1$

1 observation pr. respondent

	Beta.0	Beta.1	Sigma.0	Sigma.1	Rho	-2*Ln(L)
Sand Model	**2.50**	**-0.12**	**1.00**	**0.03**	**0.00**	**-----**
300	**21.54**	**-0.99**	**15.46**	**0.00+**	**-----**	**341.69**
300	**37.88**	**-1.75**	**27.31**	**0.00+**	**-----**	**341.69**
1000	**1.44**	**-0.07**	**0.00+**	**0.00+**	**-----**	**1233.78**
1000	**1.44**	**-0.07**	**0.00+**	**0.00+**	**-----**	**1233.78**
3000	**1.83**	**-0.08**	**0.00+**	**0.01**	**-----**	**3647.04**
3000	**1.83**	**-0.08**	**0.00+**	**0.01**	**-----**	**3647.04**
10000	**1.81**	**-0.09**	**0.00+**	**0.00+**	**-----**	**11957.88**
10000	**1.81**	**-0.09**	**0.00+**	**0.00+**	**-----**	**11957.88**
30000	**2.17**	**-0.11**	**0.00+**	**0.03**	**-----**	**35566.43**
30000	**2.17**	**-0.11**	**0.00+**	**0.03**	**-----**	**35566.43**
100000	**2.10**	**-0.10**	**0.30**	**0.02**	**-----**	**118688.22**
100000	**2.05**	**-0.10**	**0.00+**	**0.02**	**-----**	**118688.22**

Af tabellen fremgår det tydeligt, at der opstår en række problemer i forbindelse med estimationen

- For det første er ML-estimaterne for regressionsparametrene β_0 *og* β_1 meget upålidelige i de små og middelstore stikprøver.
- For det andet er ML-estimaterne for modellens to standardafvigelser σ_0 *og* σ_1 som hovedregel helt upålidelige (og ofte udartede i nul) for enhver af de betragtede stikprøvestørrelser. Denne konklusion gælder selv i de tilfælde, hvor iterationen startes i de sande parameterværdier.

Konklusionen på analysen af model 6 er derfor, at der opstår betydelige problemer i forbindelse med klassisk ML-estimation af modellens 4 parametre. Dette skyldes i al væsentlighed, at der (til forskel fra model 5) kun foreligger én observation pr. respondent, men også at modellen (til forskel fra model 4) er mere kompliceret med 4 (og ikke kun 2) parametre.

6.7 Model 7

I dette afsnit betragter vi den generelle model, der er opstillet i afsnittene 2 - 5 ovenfor, og som er den model, vi finder det mest realistisk at estimere i praksis.

Den eneste forskel på denne model og på de i afsnittene 6.5 og 6.6 analyserede modeller er imidlertid, at det nu forudsættes, at den individuelle loyalitetsparameter β_{0i} og den individuelle prisreaktionsparameter β_{1i} er positivt korrelerede. Vi vil her (og i næste afsnit) antage, at korrelationskoefficienten ρ i den todimendionale normale fordeling af β_{0i} *og* β_{1i} er lig med 0.75. Alle øvrige parameterværdier er som i modellerne 5 og 6.

I model 7 vil vi endvidere antage, at der foreligger m = 10 uafhængige, identisk fordelte observationer pr. respondent.

For model 7 fremgår ML-estimaterne for de 5 parametre $\beta_0, \beta_1, \sigma_0, \sigma_1$ *og* ρ af tabel 7.

Tabel 7: **ML-estimation af $\beta_0, \beta_1, \sigma_0, \sigma_1 \, og \, \rho$**

10 observationer pr. respondent

	Beta.0	Beta.1	Sigma.0	Sigma.1	Rho	-2*Ln(L)
Sand Model	**2.50**	**-0.12**	**1.00**	**0.03**	**0.75**	**-----**
300	**3.36**	**-0.12**	**1.70**	**0.06**	**-0.87**	**350.79**
300	**3.36**	**-0.12**	**1.70**	**0.06**	**-0.87**	**350.79**
1000	**2.41**	**-0.12**	**0.81**	**0.08**	**-0.84**	**1053.80**
1000	**2.41**	**-0.12**	**0.81**	**0.08**	**-0.84**	**1053.80**
3000	**2.98**	**-0.14**	**0.76**	**0.04**	**1.00**	**3021.91**
3000	**2.98**	**-0.14**	**0.76**	**0.04**	**1.00**	**3021.91**
10000	**2.57**	**-0.12**	**1.06**	**0.05**	**-0.10**	**10396.77**
10000	**2.57**	**-0.12**	**1.06**	**0.05**	**-0.10**	**10396.77**
30000	**2.37**	**-0.12**	**1.40**	**0.05**	**-0.32**	**31636.19**
30000	**2.37**	**-0.12**	**1.40**	**0.05**	**-0.32**	**31636.19**
100000	**2.52**	**-0.12**	**1.01**	**0.03**	**0.66**	**104359.28**
100000	**2.52**	**-0.12**	**1.01**	**0.03**	**0.66**	**104359.28**

Af tabel 7 fremgår det,

- at ML-estimaterne for alle 5 parametre er de samme, uanset om iterationen startes i det sande parameterpunkt eller i et andet parameterpunkt,
- at –2*Ln(L) antager den samme værdi, uanset om iterationen startes i det sande parameterpunkt eller i et andet parameterpunkt,
- at ML-estimaterne for de 4 parametre $\beta_0, \beta_1, \sigma_0 \, og \, \sigma_1$ ligger tættere og tættere ved de sande parameterværdier, efterhånden som stikprøven vokser, og
- at der tydeligvis opstår problemer med at estimere korrelationskoefficienten ρ. Estimatet for denne parameter bliver først acceptabelt, når den totale stikprøve er på 100000 observationer. (Hvilket som hovedregel er en urealistisk stor stikprøvestørrelse i praksis).

Når det er vanskeligt at estimere ρ, selv om der foreligger flere observationer pr. respondent, skyldes det, at den i afsnit 3 indførte stokastiske variabel

$$U_{ik} = \beta_{0i} + \beta_{1i} \Pr is_k$$

under den opstillede model er normalfordelt med middelværdien

$$\mu_k = \beta_0 + \beta_1 \Pr is_k$$

og variansen

$$\tau_k^{\,2} = \sigma_0^{\,2} + \sigma_1^{\,2} \Pr is_k^{\,2} + 2\sigma_0\sigma_1\rho \Pr is_k .$$

Som det fremgår af udtrykket for variansen, indgår ρ i et produkt sammen med standardafvigelserne σ_0 *og* σ_1. Dette medfører, at det - selv i meget store stikprøver - er yderst vanskeligt at adskille effekten af korrelationskoefficienten fra effekten af de to standardafvigelser.

Konklusionen er derfor, at model 7 – på trods af, at der foreligger flere uafhængige, identisk fordelte observationer pr. respondent - giver anledning til visse estimationsproblemer.
Men kun i forbindelse med estimation af korrelationskoefficienten ρ, idet denne indgår på en kompliceret måde i modellens varians.

6.8 Model 8

Til sidst vil vi betragte det tilfælde, hvor modellen og parameterværdierne er helt de samme som i den netop behandlede model 7, men hvor der nu kun foreligger én observation pr. respondent.

For denne model fremgår ML-estimaterne for de 5 parametre og de sædvanlige 6 stikprøvestørrelser af følgende tabel 8.

Tabel 8: **ML-estimation af $\beta_0, \beta_1, \sigma_0, \sigma_1 \, og \, \rho$**

1 observation pr. respondent

	Beta.0	Beta.1	Sigma.0	Sigma.1	Rho	-2*Ln(L)
Sand Model	**2.50**	**-0.12**	**1.00**	**0.03**	**0.75**	**-----**
300	**1.52**	**-0.10**	**4.22**	**0.00+**	**1.00**	**391.56**
300	**15.19**	**-1.15**	**78.35**	**0.64**	**-1.00**	**391.56**
1000	**490.64**	**-25.51**	**210.21**	**9.39**	**1.00**	**#NUM!**
1000	**55.97**	**-2.92**	**23.45**	**1.09**	**1.00**	**1194.46**
3000	**29.20**	**-1.35**	**22.15**	**0.13**	**1.00**	**3655.80**
3000	**36.35**	**-1.68**	**27.65**	**0.16**	**1.00**	**3655.80**
10000	**41.13**	**-1.99**	**20.15**	**0.65**	**1.00**	**12271.49**
10000	**2.45**	**-0.12**	**0.00+**	**0.05**	**-1.00**	**12271.60**
30000	**1.84**	**-0.09**	**0.89**	**0.05**	**-1.00**	**37021.81**
30000	**7.25**	**-0.35**	**7.08**	**0.21**	**-0.52**	**37021.76**
100000	**9.31**	**-0.44**	**8.63**	**0.23**	**-0.39**	**123176.78**
100000	**8.88**	**-0.42**	**8.21**	**0.22**	**-0.39**	**123176.78**

Af denne tabel fremgår det, at der opstår mange problemer i forbindelse med klassisk ML-estimation af modellens 5 parametre.

- For det første afhænger ML-estimaterne klart af, om iterationen startes i det sande parameterpunkt eller i et andet parameterpunkt.
  Men –2*Ln(L) antager den samme værdi i de to tilfælde.
  Dette viser, at maksimum for likelihoodfunktionen ikke er entydigt bestemt i denne model.
- For det andet er estimaterne såvel for de to regressionsparametre $\beta_0 \, og \, \beta_1$ som for de to standardafvigelser $\sigma_0 \, og \, \sigma_1$ meget upålidelige – selv for de store stikprøvestørrelser.
- For det tredje er det (som i model 7) ikke muligt, at opnå et blot nogenlunde pålideligt estimat for korrelationskoefficienten ρ.

Konklusionen på analysen af model 8 - der (sammen med model 7) er den generelle og for praksis mest relevante model – er derfor, at der opstår betydelige problemer i forbindelse med klassisk ML-estimation af modellens 5 parametre. Dette skyldes først og fremmest, at der (til forskel fra model 7) kun forligger én observation pr. respondent. Men det skyldes også, at modellens variansstruktur (ligesom i model 7) er kompliceret, fordi korrelationskoefficienten indgår som faktor i et produkt sammen med de to standardafvigelser.

7. Konklusion

I denne artikel har vi indledningsvis opstillet en række simple logitmodeller med stokastiske individparametre og med højst én egentlig forklarende variabel for købssandsynligheden for et givet mærke – nemlig prisen for det pågældende mærke.

Med henblik på at estimere parametrene i denne type modeller har vi herefter udviklet et specielt estimationsprogram, der approksimativt bestemmer maksimum likelihood estimaterne. Programmet, der bygger på numerisk integration, minder om Monte Carlo simulation, men det har den åbenbare fordel, at likelihoodfunktionen kan opskrives eksplicit og dermed forholdsvis let kan maksimeres.

Endelig har vi ved hjælp af det udviklede estimationsprogram estimeret parametrene i 8 modeller i alt – 4 modeller med og de samme 4 modeller uden gentagne observationer pr. respondent.

For disse 8 modeller har vi i artiklen opnået følgende hovedkonklusioner:

Nøjagtigheden på ML-estimaterne over parametrene i en given model afhænger generelt af følgende 5 forhold:

1. Hvor kompliceret modellen er, dvs. om der indgår få eller mange parametre i modellen.
2. Hvor kompliceret modellens parameterstruktur – specielt dens variansstruktur - er, dvs. om effekten af de enkelte parametre kan adskilles.
3. Om der indgår forklarende variable i modellen eller ej.
4. Om der foreligger flere uafhængige, identisk fordelte observationer pr. respondent eller ej, og (naturligvis)
5. Om den totale stikprøve er stor eller lille.

Artiklens hovedproblemstilling har været at analysere det under punkt 4 nævnte forhold.
Om dette forhold kan vi drage følgende konklusion:

- Når modellens parametre varierer fra forbruger til forbruger (efter en nærmere specificeret model), og der foreligger flere uafhængige, identisk fordelte observationer pr. respondent, opstår der som hovedregel ingen problemer med ML-estimationen af modellens parametre. Korrelationskoefficienten mellem de individuelle regressionskoefficienter er dog vanskelig at estimere – i hvert tilfælde i normale stikprøvestørrelser. Dette skyldes modellens komplicerede variansstruktur.

- Når modellens parametre varierer fra forbruger til forbruger, og der kun foreligger én observation pr. respondent, opstår der derimod som hovedregel betydelige problemer med ML-estimationen af samtlige parametre i modellen.
  Disse problemer er størst, hvis der ikke indgår forklarende variable i modellen og/eller, hvis der indgår (relativt) mange parametre i modellen og/eller, hvis modellens parameterstruktur – specielt dens variansstruktur - er kompliceret.
  Denne konklusion gælder generelt, dvs. også i de tilfælde, hvor der foreligger endog meget store stikprøvestørrelser.

De sidst omtalte estimationsproblemer forstærkes formodentlig, hvis man (som det af og til er tilfældet i praksis) betragter modeller med mange forklarende variable – og dermed med mange regressionsparametre, mange standardafvigelser og - især - mange korrelationskoefficienter. Også fordi estimationen i praksis (af og til) baseres på forholdsvis få observationer.

Tak til Tue Tjur for mange værdifulde diskussioner om artiklens problemstilling.

Litteraturfortegnelse

Daniel McFadden (1974)
Conditional Logit Analysis of Qualitative Choice Behavior
Frontiers in Econometrics, Chapter Four, Ed. by Paul Zarembka
New York

Daniel McFadden and K. Train (1999)
Mixed MNL Models for Discrete Response
Applied Econometrics

S. P. B. Murthi and K. Srinivasan (1998)
Performance of the integrated random coefficients covariance probit model
Implications for brand choice
International Journal of Research in Marketing, 15, 137-156

Jørgen Kai Olsen (2001)
En operationel model til måling af kundeloyalitet
Research Paper
Institut for Afsætningsøkonomi, Handelshøjskolen i København

Jørgen Kai Olsen (2003)
En stokastisk model for total og partiel kundeloyalitet
Research Paper
Institut for Afsætningsøkonomi, Handelshøjskolen i København

Sheldon M. Ross (2000)
Introduction to Probability Models
Harcourt Academic Press

Hans Stubbe Solgaard and Torben Hansen (2002)
A Hierarchical Bayes Model of Choice Between Supermarket Formats
Journal of Retailing and Consumer Services

Tue Tjur (2002)
Logistic regression models for single-source data – a simulation study
Department of Management Science and Statistics
Copenhagen Business School

Tue Tjur (2003)
A warning concerning random effects and random coefficients
in logistic regression models for binary data
Department of Management Science and Statistics
Copenhagen Business School