Arbeidsnotater

S T A T I S T I S K   S E N T R A L B Y R Å

Dronningensgt. 16, Dep, Oslo 1. Tlf.*(02) 41 38 20

IO 78/8 27. april 1978

SAMMENLIGNING AV MINSTE KVADRATERS METODE OG SANNSYNLIGHETSMAKSIMERINGSMETODEN I BINÆR REGRESJON

av

Henrik Dahl *)

INNHOLD

*) Dette arbeidet er utført under studiepermisjon 1977/78 og er støttet av NAVF. Forfatteren vil takke professor Erling Sverdrup, som foreslo problemstillingen, for råd og kommentarer til arbeidet. Dessuten takkes medlemmer av Gruppe for metoder for nyttige diskusjoner.

SAMMENDRAG

Når en søker å få oversikt over hvordan en binær (0-1) variabel varierer med en kontinuerlig variabel, kan det være naturlig å bruke regresjonsanalyse med den binære variable som venstresidevariabel og den kontinuerlige variable som høyresidevariabel.

I kapittel 1 diskuteres modeller for denne situasjonen. Spesielt diskuteres gyldighetsområdet for den enkleste ("p-lineære") modellen.

I kapittel 2 studeres minste kvadraters metode estimator for regresjonskoeffisienten i den "p-lineære modellen", og et asymptotisk uttrykk for variansen til denne estimatoren utledes.

I kapittel 3 studeres sannsynlighetsmaksimeringsestimatoren for regresjonskoeffisienten i den "p-lineære modellen", og et asymptotisk uttrykk for variansen til denne estimatoren utledes.

I kapittel 4 sammenlignes minste kvadraters- og sannsynlighetsmaksimeringsestimatorene for regresjonskoeffisienten i den "p-lineære modellen" ved hjelp av de asymptotiske uttrykkene som er utledet i kapitlene 2 og 3.

1. Om modeller for binær regresjon.

Vi har en situasjon der vi for ulike verdier av en "høyreside variabel" y observerer en "venstreside variabel" X som er binær. Vi ønsker å si noe om hvordan X "i gjennomsnitt" varierer med y. Det kan da være naturlig, som den enkleste modell for en slik situasjon, å bruke følgende modell (som i det følgende vil bli henvist til som "den p-lineære modellen"):

$$(1) \quad \begin{cases} X_1, X_2, \ldots, X_n \text{ er uavhengige} \\ E(X_i) = p_i = P(X_i=1) = \alpha + \beta y_i \quad ; \quad i = 1, 2, \ldots, n \end{cases}$$

I denne modellen har vi to parametre α og β. Vanligvis vil vi særlig være interesserte i parameteren β, da denne så å si måler "sammenhengen" mellom y og X.

Modellen (1) har imidlertid sine klare begrensninger. Om intet spesielt er forutsatt om variasjonsområdet for y'ene, ser vi at modellen (1) for y-verdier utenfor intervallet $\left[-\frac{\alpha}{\beta}, \frac{1-\alpha}{\beta}\right]$ vil foreslå sannsynligheter utenfor $[0,1]$. Dette kan ikke tolereres.

Den mest nærliggende modifikasjon av modellen (1) for å rette på denne skavanken, ville være:

$$\begin{aligned} &\text{Sammenhengen } p_i = \alpha + \beta y_i \text{ gjelder bare for} \\ &y_i \in \left[-\frac{\alpha}{\beta}, \frac{1-\alpha}{\beta}\right] \\ &p_i = 0 \text{ hvis } y_i < -\frac{\alpha}{\beta} \\ &p_i = 1 \text{ hvis } y_i > \frac{1-\alpha}{\beta} \end{aligned}$$

Denne modifiserte modellen vil jeg foretrekke å skrive på formen:

$$(2) \quad \begin{cases} X_1, X_2, \ldots, X_n \text{ er uavhengige} \\ E(X_i) = p_i = P(X_i=1) = F(\alpha+\beta y_i) \\ \text{der } F(x) = \begin{cases} 1 & \text{hvis } x > 1 \\ x & \text{hvis } 0 \leq x \leq 1 \\ 0 & \text{hvis } x < 0 \end{cases} \end{cases}$$

Dermed faller den modifiserte modellen (2) inn under en mer generell klasse av modeller, som kan beskrives ved:

$$(3)\quad \begin{cases} X_1,\ X_2,\ \ldots,\ X_n \text{ er uavhengige} \\ E(X_i) = p_i = P(X_i=1) = F(\alpha_1+\beta_1 y_i) \\ \text{der } F \text{ er en kjent fordelingsfunksjon} \end{cases}$$

Modellen (2) fåes som et spesialtilfelle av (3) ved å la F være fordelingsfunksjonen til den rektangulære fordeling over $[0,1]$.

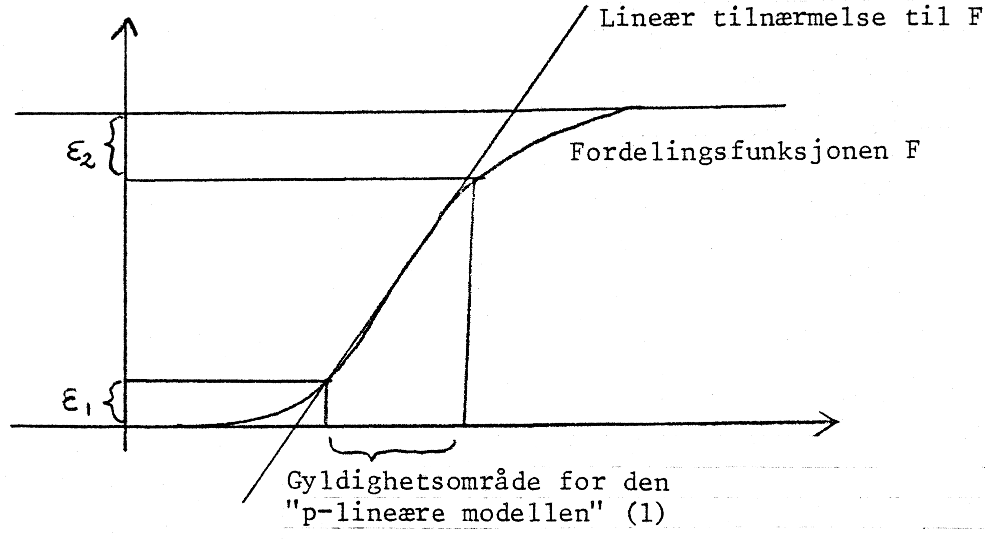
Modellen (2) fortoner seg som en temmelig urimelig modell i og med "knekkpunktene" i sammenhengen mellom p og y. Det er vanskelig å tenke seg noen konkret situasjon der det finnes slike "knekkpunkter". En annen sak er at om slike "knekkpunkter" skal trekkes inn i estimeringsmetoder, mister modellen (1) mye av sin eksistensberettigelse, nemlig enkelheten.

Moralen av dette er at modellen (1) må oppfattes som en forenkling av en modell av typen (3). Modellen (1) har et gyldighetsområde som begrenser seg til et intervall av y-verdier der fordelingsfunksjonen F med rimelig tilnærmelse er lineær. Dette innebærer at modellen (1) bare har interesse og gyldighet når y'ene er slik at:

$$\varepsilon_1 < \alpha + \beta y_i < 1 - \varepsilon_2 \quad ; \quad i = 1, 2, \ldots, n$$

der ε_1 og ε_2 avhenger av fordelingsfunksjonen F.

Rimeligvis har fordelingsfunksjonen F en S-form som vist i følgende figur:



Den "p-lineære modellen" (1) har altså bare gyldighet og interesse for y-verdier slik at $\alpha + \beta y$ holder seg unna verdiene 0 og 1. Hvor stort gyldighetsområdet for den "p-lineære modellen" (1) er, avhenger av fordelingsfunksjonen F.

Om vi ikke er villige til å begrense oss til gyldighetsområdet for den "p-lineære modellen" (1), må vi arbeide i den mer generelle modellen (3). Om F i (3) velges som standardnormalfordelingen, er vi over i probitananalyse. Om F i (3) velges som den logistiske fordeling, er vi over i logitanalyse. Både probit- og logitanalyse forutsetter at vi har et rimelig antall gjentakelser for hver y-verdi.

Prisen vi må betale for det større gyldighetsområdet for modellen (3) fremfor modellen (1), er altså at vi må ha et rimelig antall gjentakelser for hver y-verdi. Dette kan eventuelt oppnåes ved gruppering. Modellen (1) er altså særlig aktuell i tilfelle hvor det ikke er mulig å få til et rimelig antall gjentakelser for hver y-verdi, og der dette heller ikke kan oppnåes ved gruppering.

2. Minste kvadraters metode (MKM).

I modellen

$$(1) \quad \begin{cases} X_1, X_2, \ldots, X_n \text{ er uavhengige} \\ E(X_i) = p_i = P(X_i=1) = \alpha + \beta y_i \quad ; \quad i = 1, 2, \ldots, n \end{cases}$$

søker vi minste kvadraters metode (MKM) estimator for parameteren $\underline{p}$. Selv om α inngår som parameter i modellen og ikke a priori antas å være null, har den forholdsvis mindre interesse enn parameteren β. Jeg vil derfor konsentrere meg om estimeringen av β.

MKM-estimatoren for β : $\hat{\beta}$ er gitt ved:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i (y_i - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Under modellen (1) er $\hat{\beta}$ forventningsrett for β, og vi har følgende uttrykk for variansen:

$$\mathrm{var}(\hat{\beta}) = \frac{\sum_{i=1}^{n} (y_i-\bar{y})^2 \mathrm{var}(X_i)}{\left[\sum_{i=1}^{n} (y_i-\bar{y})^2\right]^2}$$

Siden X_i er binær, er:

$$\mathrm{var}(X_i) = p_i(1-p_i) = (\alpha+\beta y_i)(1-\alpha-\beta y_i)$$

Altså:

$$\mathrm{var}(\hat{\beta}) = \frac{\sum_{i=1}^{n} (y_i-\bar{y})^2 (\alpha+\beta y_i)(1-\alpha-\beta y_i)}{\left[\sum_{i=1}^{n} (y_i-\bar{y})^2\right]^2}$$

Som vi ser avhenger $\mathrm{var}(\hat{\beta})$ av regresjonsparametrene α og β. Vi kan følgelig ikke uttale oss om $\mathrm{var}(\hat{\beta})$ uavhengig av regresjonsparametrene, slik som i den vanlige regresjonsmodellen.

Vi vil senere sammenligne MKM-estimatoren $\hat{\beta}$ med sannsynlighetsmaksimeringsestimatoren for β. For å kunne gjennomføre en slik sammenligning vil vi anta at y'ene er generert fra en parametrisk fordelingsklasse. Som nevnt i 1., har den "p-lineære modellen" (1) bare gyldighet og interesse for y'er slik at $0 < \alpha + \beta y_i < 1$. Det er derfor naturlig å generere y'ene fra en fordeling som bare har positiv sannsynlighetsmasse for $0 < \alpha + \beta y_i < 1$. En fordelingsklasse som tilfredsstiller dette kravet får vi ved å la:

$$\alpha + \beta y_i \quad ; \quad i = 1, 2, \ldots, n$$

være uavhengige og identisk Betafordelte.

Det kan kanskje synes underlig å generere y'ene ved hjelp av en fordelingsklasse der α og β inngår. På den annen side ser vi at dette er nødvendig om en skal sikre seg mot genereringer slik at $\alpha + \beta y_i \notin [0,1]$. Vi tenker oss altså y'ene generert på følgende måte:

$$z_i = \alpha + \beta y_i \quad ; \quad i = 1, 2, \ldots, n$$

er uavhengige og Betafordelte (a,a). z'ene har altså tettheten:

$$f(z) = \begin{cases} \dfrac{\Gamma(2a)}{\Gamma(a)^2} z^{a-1}(1-z)^{a-1} & 0 < z < 1 \\ 0 & \text{ellers} \end{cases}$$

Mer generelt kunne man brukt en Betafordeling (a,b). (Se helt til slutt.) For å forenkle sammenligningen i kapittel 4, har jeg latt $a = b$. Det at $a = b$ svarer til at y'ene har en symmetrisk fordeling og at symmetripunktet svarer til $E(X) = 0.5$.

Under forutsetningen at y'ene er generert som beskrevet, søker vi:

$$\underset{n\to\infty}{\text{p lim}}\; n \text{ var } \hat{\beta}$$

der "$\underset{n\to\infty}{\text{p lim}}$" står for "grense i sannsynlighet".

$$n \text{ var } \hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2(\alpha+\beta y_i)(1-\alpha-\beta y_i)}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^2}$$

Innfører $z_i = \alpha + \beta y_i$ og utnytter at $y_i - \bar{y} = \frac{1}{\beta}(z_i-\bar{z})$:

$$n \text{ var } \hat{\beta} = \beta^2 \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})^2 z_i(1-z_i)}{\left[\frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})^2\right]^2}$$

Vi bemerker at $n \text{ var } \hat{\beta}$ ikke avhenger av verdien på α. Dette er intuitivt rimelig.

"Telleren" omformes:

$$\frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})^2 z_i(1-z_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(z_i^2-2\bar{z}z_i+\bar{z}^2)z_i(1-z_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n}z_i^3(1-z_i) - 2\bar{z}\cdot\frac{1}{n}\sum_{i=1}^{n}z_i^2(1-z_i) + \bar{z}^2\cdot\frac{1}{n}\sum_{i=1}^{n}z_i(1-z_i)$$

Anta $z_i, z_2, \ldots, z_n$ er uavhengige med samme fordeling som Z. Hvis de oppsatte forventningene eksisterer, gjelder ifølge Khinchins setning:

$$\text{(4)} \qquad \underset{n\to\infty}{\text{p lim}}\ \frac{1}{n}\sum_{i=1}^{n}z_i^m(1-z_i)^n = E(Z^m(1-Z)^n).$$

For Z Betafordelt (a,a) har vi at hvis $m > 0$, $n > 0$, gjelder:

$$\text{(5)} \qquad E(Z^m(1-Z)^n) = \frac{\Gamma(2a)\cdot\Gamma(a+m)\Gamma(a+n)}{\Gamma(a)^2\Gamma(2a+m+n)}$$

Av dette følger:

$$E(Z) = \frac{1}{2}$$

$$E(Z^3(1-Z)) = \frac{a(a+2)}{4(2a+1)(2a+3)}$$

$$E(Z^2(1-Z)) = \frac{a}{4(2a+1)}$$

$$E(Z(1-Z)) = \frac{a}{2(2a+1)}$$

$$\text{var}(Z) = E(Z^2) - (E(Z))^2 = \frac{1}{4(2a+1)}$$

Ved å bruke disse resultatene, samt Slutskys teorem, får vi:

$$\underset{n\to\infty}{\text{p lim}}\ n \text{ var } \hat{\beta} = 2\beta^2\,\frac{a(2a+1)}{2a+3}$$

Mer generelt har vi at hvis z'ene er Betafordelte (a,b), gjelder:

$$\underset{n\to\infty}{\text{p lim}}\ n \text{ var } \hat{\beta} = \beta^2 \frac{(a+b)(a+b+1)(a^2b+ab^2+2a^2+2b^2-2ab)}{ab(a+b+3)(a+b+2)}$$

3. Sannsynlighetsmaksimeringsmetoden (SMM).

I modellen:

$$(1) \quad \begin{cases} X_1, X_2, \ldots, X_n \text{ er uavhengige} \\ E(X_i) = p_i = P(X_i=1) = \alpha + \beta y_i \ ; \ i = 1, 2, \ldots, n. \end{cases}$$

søker vi sannsynlighetsmarkeringsmetode (SMM) estimator for parameteren β.

SMM-estimatoren for β : β^* fremkommer ved å løse ligningssystemet:

$$\sum_{i=1}^{n} \frac{x_i}{\alpha^*+\beta^* y_i} = \sum_{i=1}^{n} \frac{1-x_i}{1-\alpha^*-\beta^* y_i}$$

$$\sum_{i=1}^{n} \frac{x_i y_i}{\alpha^*+\beta^* y_i} = \sum_{i=1}^{n} \frac{(1-x_i)y_i}{1-\alpha^*-\beta^* y_i}$$

med hensyn på β^*.

Vi har altså ikke noe sluttet uttrykk for β^* som vi kan bruke til å studere estimatorens egenskaper.

Imidlertid kan vi bruke asymptotisk teori for SMM-estimatorer.

Under visse regularitetskrav må en kunne gå ut fra at det kan vises at β^* er asymptotisk normalfordelt med riktig forventning og varians som finnes ved å invertere informasjonsmatrisen. Dette bør undersøkes nærmere.

Informasjonsmatrisen finnes ved å derivere loglikelihood-funksjonen:

$$L(\alpha,\beta) = \sum_{i=1}^{n} [X_i \ln(\alpha+\beta y_i) + (1-X_i)\ln(1-\alpha-\beta y_i)]$$

Informasjonsmatrisen I er:

$$I = \begin{bmatrix} -E\frac{\partial^2}{\partial\alpha^2}L(\alpha,\beta), & -E\frac{\partial^2}{\partial\alpha\partial\beta}L(\alpha,\beta) \\ -E\frac{\partial^2}{\partial\alpha\partial\beta}L(\alpha,\beta), & -E\frac{\partial^2}{\partial\beta^2}L(\alpha,\beta) \end{bmatrix}$$

I vårt tilfelle får vi:

$$I = \begin{bmatrix} \sum_{i=1}^{n}\left(\frac{1}{\alpha+\beta y_i} + \frac{1}{1-\alpha-\beta y_i}\right), & \sum_{i=1}^{n}\left(\frac{y_i}{\alpha+\beta y_i} + \frac{y_i}{1-\alpha-\beta y_i}\right) \\ \sum_{i=1}^{n}\left(\frac{y_i}{\alpha+\beta y_i} + \frac{y_i}{1-\alpha-\beta y_i}\right), & \sum_{i=1}^{n}\left(\frac{y_i^2}{\alpha+\beta y_i} + \frac{y_i^2}{1-\alpha-\beta y_i}\right) \end{bmatrix}$$

Determinanten til I er:

$$D = \sum_{i=1}^{n}\left(\frac{1}{\alpha+\beta y_i} + \frac{1}{1-\alpha-\beta y_i}\right) \sum_{i=1}^{n}\left(\frac{y_i^2}{\alpha+\beta y_i} + \frac{y_i^2}{1-\alpha-\beta y_i}\right)$$
$$- \left[\sum_{i=1}^{n}\left(\frac{y_i}{\alpha+\beta y_i} + \frac{y_i}{1-\alpha-\beta y_i}\right)\right]^2$$

Altså har vi:

$$I^{-1} = \frac{1}{D}\begin{bmatrix} \sum_{i=1}^{n}\left(\frac{y_i^2}{\alpha+\beta y_i} + \frac{y_i^2}{1-\alpha-\beta y_i}\right), & -\sum_{i=1}^{n}\left(\frac{y_i}{\alpha+\beta y_i} + \frac{y_i}{1-\alpha-\beta y_i}\right) \\ \sum_{i=1}^{n}\left(\frac{y_i}{\alpha+\beta y_i} + \frac{y_i}{1-\alpha-\beta y_i}\right), & \sum_{i=1}^{n}\left(\frac{1}{\alpha+\beta y_i} + \frac{1}{1-\alpha-\beta y_i}\right) \end{bmatrix}$$

Av dette kan den asymptotiske variansen til β^* leses ut som:

$$\frac{1}{D}\sum_{i=1}^{n}\left(\frac{1}{\alpha+\beta y_i} + \frac{1}{1-\alpha-\beta y_i}\right)$$

Som i studiet av var $\hat{\beta}$ antar vi også her at

$$z_i = \alpha + \beta y_i \quad ; \quad i = 1, 2, \ldots, n$$

er uavhengige og identisk Betafordelte (a,a).
Den asymptotiske variansen til β^* er altså:

$$\frac{1}{D} \sum_{i=1}^{n} \left(\frac{1}{z_i} + \frac{1}{1-z_i} \right)$$

La $K_i = \frac{1}{z_i} + \frac{1}{1-z_i}$

Vi har:

$$D = \frac{1}{\beta^2}\left[\left(\sum_{i=1}^{n} K_i \right)\left(\sum_{i=1}^{n} (z_i-\alpha)^2 K_i \right) - \left(\sum_{i=1}^{n} (z_i-\alpha) K_i \right)^2 \right]$$

$$= \frac{1}{\beta^2}\left[\left(\sum_{i=1}^{n} K_i \right)\left(\sum_{i=1}^{n} z_i^2 K_i \right) - 2\alpha\left(\sum_{i=1}^{n} z_i K_i \right) + \alpha^2\left(\sum_{i=1}^{n} K_i \right)^2 \right.$$

$$\left. - (\Sigma z_i K_i)^2 + 2\alpha\left(\sum_{i=1}^{n} K_i \right)\left(\sum_{i=1}^{n} z_i K_i \right) - \alpha^2\left(\sum_{i=1}^{n} K_i \right)^2 \right]$$

$$= \frac{1}{\beta^2}\left[\left(\sum_{i=1}^{n} K_i \right)\left(\sum_{i=1}^{n} z_i^2 K_i \right) - \left(\sum_{i=1}^{n} z_i K_i \right)^2 \right]$$

Dette viser at den asymptotiske variansen for β^* ikke avhenger av verdien på α. Dette er intuitivt rimelig.

Under forutsetningen at y'ene er generert som beskrevet, søker vi:

$$\underset{n\to\infty}{\text{p lim}}\ n \text{ var } \beta^* = \underset{n\to\infty}{\text{p lim}} \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{z_i} + \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1-z_i}}{\frac{1}{n^2}D}$$

"Nevneren" omformes:

$$\frac{1}{n^2}D = \frac{1}{\beta^2}\left[\left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{z_i} + \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1-z_i} \right)\left(\frac{1}{n}\sum_{i=1}^{n} z_i + \frac{1}{n}\sum_{i=1}^{n} \frac{z_i^2}{1-z_i} \right) \right.$$

$$\left. - \left(1 + \frac{1}{n}\sum_{i=1}^{n} \frac{z_i^2}{1-z_i} \right)^2 \right]$$

Ved å bruke formlene (4) og (5) i 2. får vi:

$$\underset{n\to\infty}{p\lim} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{z_i} = E(Z^{-1}) = \frac{2a-1}{a-1} \qquad \text{for } a > 1$$

$$\underset{n\to\infty}{p\lim} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1-z_i} = E((1-Z)^{-1}) = \frac{2a-1}{a-1} \qquad \text{for } a > 1$$

$$\underset{n\to\infty}{p\lim} \frac{1}{n} \sum_{i=1}^{n} z_i = E(Z) = \frac{1}{2}$$

$$\underset{n\to\infty}{p\lim} \frac{1}{n} \sum_{i=1}^{n} \frac{z_i^2}{1-z_i} = E(Z^2(1-Z)^{-1}) = \frac{a+1}{2(a-1)} \qquad \text{for } a > 1$$

$$\underset{n\to\infty}{p\lim} \frac{1}{n} \sum_{i=1}^{n} \frac{z_i}{1-z_i} = E(Z(1-Z)^{-1}) = \frac{a}{a-1} \qquad \text{for } a > 1$$

Ved å bruke disse resultatene samt Slutskys teorem, får vi:

$$\underset{n\to\infty}{p\lim}\, n \operatorname{var} \beta^* = 2\beta^2(a-1)$$

Denne utledningen gjelder bare for $a > 1$ (ellers eksisterer ikke alle de oppsatte forventningene).

Mer generelt har vi at hvis z'ene er Betafordelte (a,b), gjelder:

$$\underset{n\to\infty}{p\lim}\, n \operatorname{var} \beta^* = \beta^2(a+b-2)$$

Her må $a > 1$ og $b > 1$ for at utledningen skal gjelde.

4. Sammenligning av MKM og SMM.

Vi har i 2. og 3. funnet følgende asymptotiske uttrykk:

$$n \operatorname{var}(\hat{\beta}) \approx 2\beta^2 \frac{a(2a+1)}{2a+3}$$

$$n \operatorname{var}(\beta^*) \approx 2\beta^2(a-1)$$

Som en kunne vente, er SMK-estimatoren β^* bedre enn MKM-estimatoren $\hat{\beta}$. Dessuten ser en at forskjellen i varians er størst for små verdier av a, det vil si når y-verdiene har stor spredning.

Effisienstapet ved å bruke $\hat{\beta}$ fremfor β^* er altså asymptotisk:

$$1 - e = 1 - \frac{\mathrm{var}(\beta^*)}{\mathrm{var}(\hat{\beta})} = \frac{3}{a(2a+1)}$$

Det kan være av interesse å bruke:

$$\sigma^2 = \mathrm{var}(y_i) = \frac{1}{\beta^2}\,\mathrm{var}(z_i)$$

som parameter i stedet for a. Siden z_i er Betafordelt (a,a), er:

$$\mathrm{var}(z_i) = \frac{1}{4(2a+1)}$$

Altså har vi følgende sammenheng mellom σ^2 og a:

$$\sigma^2 = \frac{1}{4\beta^2(2a+1)}$$

eller:

$$a = \frac{1}{8\beta^2\sigma^2} - \frac{1}{2}$$

Uttrykt ved σ^2 har vi:

$$n\,\mathrm{var}(\hat{\beta}) \approx \frac{\frac{1}{4\sigma^2} - 1}{1+8\beta^2\sigma^2}$$

$$n\,\mathrm{var}(\beta^*) \approx \frac{1}{4\sigma^2} - 3\beta^2$$

I utledningen av det asymptotiske uttrykket for $n\,\mathrm{var}(\beta^*)$ forutsatte vi $a > 1$. Dette svarer til at:

$$\sigma^2 < \frac{1}{12\beta^2}$$

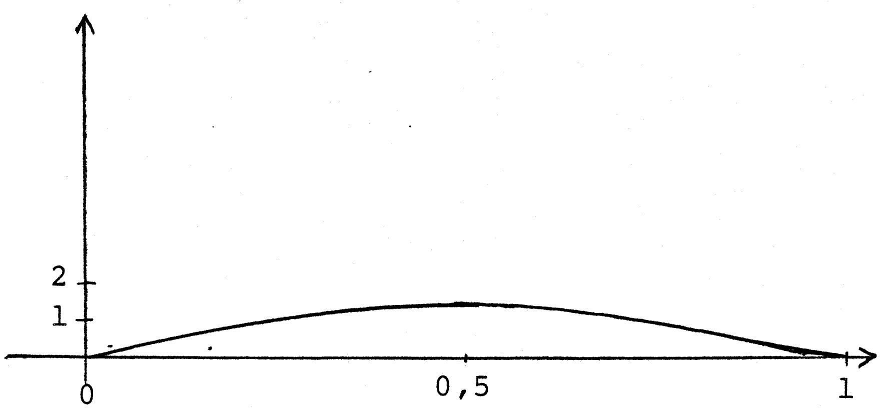
For å få en oversikt over følgene av å velge $\hat{\beta}$ fremfor β^*, kan vi tabellere effisienstapet som funksjon av a (eller σ^2):

TABELL 1

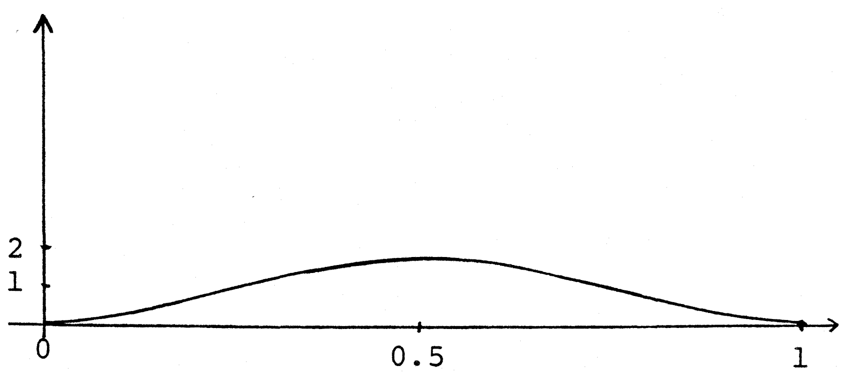
a	σ^2	Effisienstapet i % : $100 \cdot \frac{3}{a(2a+1)}$
2	$\frac{1}{20\beta^2}$	30
3	$\frac{1}{28\beta^2}$	14.3
5	$\frac{1}{44\beta^2}$	5.5
12	$\frac{1}{100\beta^2}$	1
49.5	$\frac{1}{400\beta^2}$	0.06

Som vi ser kan effisienstapet bli stort hvis y'ene har stor spredning, mens effisienstapet er neglisjerbart for a-verdier fra ca. 10 og oppover.

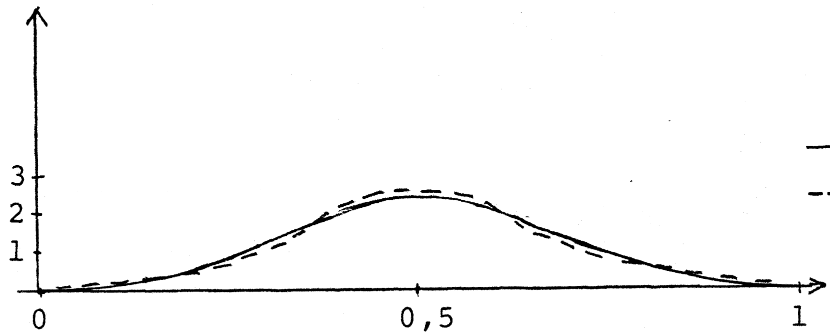
For å illustrere hva de ulike verdiene av a (eller σ^2) betyr, har jeg skissert tetthetene til Betafordelingene (a,a) for $a = 2, 3, 5, 12$ og 49.5.



$a = 2$
Tettheten $6z(1-2)$



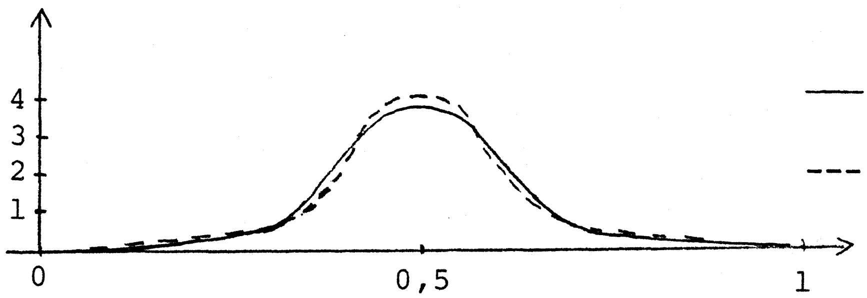
$a = 3$
Tettheten $30z^2(1-z)^2$



a = 5

—— Tettheten 630 $z^4(1-z)^4$

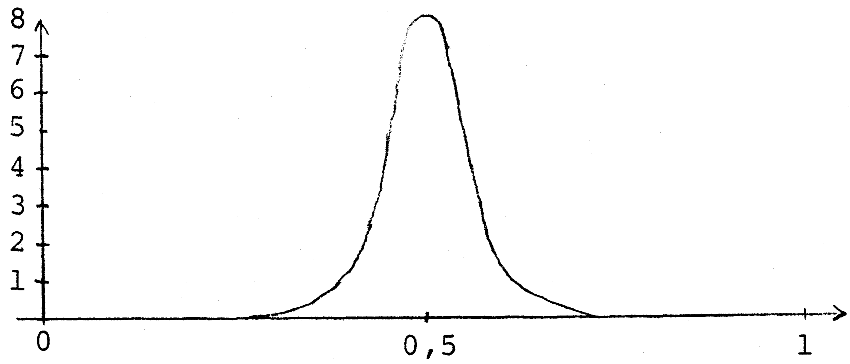
--- Dessuten N(0.5, 0.15)



a = 12

—— Tettheten 16224936 z"(1-z)"

--- Dessuten N(0.5, 0.1)



a = 49.5 som er lik N(0.5, 0.05)