Eksamen på Økonomistudiet vinter 2015-16
Sandsynlighedsteori og Statistik
2. Årsprøve
15. januar, 2016

(3-timers prøve med hjælpemidler)

Rettevejledning

Opgaven består af tre delopgaver, som alle skal besvares. De tre opgaver kan regnes uafhængigt af hinanden. Opgaver 1 og 2 indgår med samme vægt som opgave 3.

Opgave 1

1. Sandsynligheden $P(X_I = 0)$ og $P(X_{II} = 0)$ skal bestemmes. X_I (X_{II}) er poissonfordelt med parameter $\lambda_I = 0.5(\lambda_{II} = 1)$. Punktsandsynligheden i 0 udregnes som punktsandsynligheden i en Poissonfordeling (se Sørensen, side 121 formel (4.1.6)):

$$\begin{aligned} P(X_I &= 0) = \exp(-0.5) = 0.6065 \\ P(X_{II} &= 0) = \exp(-1) = 0.3679 \end{aligned}$$

2. Det forventede antal skader for henholdsvis type I og II skal bestemmes. Det forventede antal bestemmes som middelværdierne af X_I og X_{II}. Middelværdien i en poissonfordeling er givet ved λ (se Sørensen eksempel 4.1. side 131):

$$\begin{aligned} E(X_I) &= 0.5 \\ E(X_{II}) &= 1 \end{aligned}$$

3. Det antages, at forsikringsselskabet har 300 forsikringstagere og at de fordeler sig således: 200 personer af type I og 100 af type II. Lad Z være antallet af skader for alle de forsikrede. Det forventede antal skader $E(Z)$, kan udregnes ved at anvende sætningen 4.4.6 (se Sørensen side 135):

$$\begin{aligned} E(Z) &= E(\sum_{i=1}^{200} X_{I,i} + \sum_{j=1}^{100} X_{I,j}) = \sum_{i=1}^{200} E(X_{I,i}) + \sum_{j=1}^{100} E(X_{I,j}) \\ &= \sum_{i=1}^{200} 0.5 + \sum_{j=1}^{100} 1 = 200 \cdot 0.5 + 100 \cdot 1 = 200 \end{aligned}$$

4. Fordelingen af det samlede antal skader (Z) skal bestemmes. Her kan man anvende sætning 4.3.4 i Sørensen side 129 og benytter at enkelte forsikringstagere kan antages at være uafhængige. Det følger så at Z er Poissonfordelt med $\lambda = 200$.

Opgave 2

I denne opgave undersøges, hvordan forsikringsselskabet skal fastlægge sine præmier, hvis forsikringsselskabet ikke ved om personen er type I eller type II.

Det antages, at forsikringsselskabet har kunder i to forskellige regioner og at man kender fordelingen af type I og type II kunder i de to regioner A og B. Det antages, at Y angiver om kunden er type I eller type II og R angiver regionen. Lad Y og R være stokastisk variable, hvor deres simultane fordeling er angivet i Tabel 1. $Y = 1$ betyder, at personen er type I.

Tabel 1: *Den simultane fordeling af type Y og region R*

		Y	
		$Y = 1$ (type I)	$Y = 0$ (type II)
	Region A	0.10	0.20
R	Region B	0.35	0.35

1. Den marginale fordeling for regionerne (R) kan bestemmes ved at anvende sæting 4.2.1 (Sørensen side 124)

$$\begin{aligned} P(R = A) &= p(1, A) + p(0, A) = 0.1 + 0.2 = 0.3 \\ P(R = B) &= p(1, B) + p(0, B) = 0.35 + 0.35 = 0.7 \end{aligned}$$

2. Sandsynligheden $P(Y = 1|R = A)$ angiver den betingede sandsynlighed for $Y = 1$ givet at region A. Det betyder, at vi har sandsynligheden for, at den tilfældig udvalgt person, som bor i region A, er af type I. Den betingede sandsynlighed kan beregnes ved at anvende definitionen 1.4.1 (se Sørensen side 24)

$$P(Y = 1|R = A) = \frac{P(Y = 1 \cap R = A)}{P(R = A)} = \frac{0.1}{0.3} = 1/3$$

3. Der er flere måde at vise, at Y og R ikke er uafhængige. En måde er at benytte definitionen af uafhængighed og vise at

$$P(Y = 1, R = A) \neq P(Y = 1) \cdot P(R = A)$$

man kan vise at $P(Y = 1) = 0.45$ og da $0.45 \cdot 0.3 = 0.135 \neq 0.10$, kan Y og R ikke være uafhængige. Alternative måder at vise afhængighed er ved at vise, at $Cov(R, Y) \neq 0$ eller $P(Y = 1|R = A) \neq P(Y = 1)$.

4. Det forventede antal skader for en person som hhv. bor i region A og B udregnes. Vi skal her bestemme den betingede middelværdi $E(Y \cdot X_I + (1 - Y) \cdot X_{II}|R = A)$. Man kan nu regne på ovenstående udtryk

$$\begin{aligned} E(Y \cdot X_I + (1 - Y) \cdot X_{II}|R = A) &= E(Y \cdot X_I|R = A) + E((1 - Y) \cdot X_{II}|R = A) \\ &= E(X_I)E(Y|R = A) + E(X_{II})E((1 - Y)|R = A) \\ &= E(X_I)E(Y|R = A) + E(X_{II})(1 - E(Y|R = A)) \\ &= \frac{1}{2}\frac{1}{3} + 1 \cdot (1 - \frac{1}{3}) = \frac{5}{6} \end{aligned}$$

Vi benytter her sætning 4.4.6. Første lighedstegn følger af middelværdien af en sum er summen af middelværdierne. Andet lighedstegn følger af at X_I og X_{II} er uafhængige af Y og R. Sidste lighedstegn følger at at $E(Y|R = A) = P(Y = 1|R = A)$.Tilsvarende kan $E(Y \cdot X_I + (1 - Y) \cdot X_{II}|R = B)$ udregnes

$$\begin{aligned} E(Y \cdot X_I + (1 - Y) \cdot X_{II}|R = B) &= E(X_I)E(Y|R = B) + E(X_{II})(1 - E(Y|R = B)) \\ &= \frac{1}{2} \cdot \frac{1}{2} + 1 \cdot (1 - \frac{1}{2}) \\ &= \frac{3}{4} \end{aligned}$$

Forikringsselskabet ønsker at fastlægge prisen på præmien, således at præmien dækker den forventede udgift til skader. Forsikringsselskabet kan ikke observere om personen er type I eller type II, kun hvilken region personen bor i. Det vil være optimalt for forsikringsselskabet at have en prispolitik, som diskriminerer mellem regionerne. Personer, som bor i region A, bør betale en højere præmie, da man forventer, at de vil have flere skader end personer, som bor i region B.

Opgave 3

1. Figuren viser en klart tendens til, at variansen på afkastet afhænger af låneandelen. Det er præcis hvad der opnås med modellen. Det kaldes ofte for *heteroskedasticitet*.

2. Hvis $(Z \mid X_i = x_i) \overset{d}{=} N(0,1)$, så gælder der, at

$$Y_i = \sigma_i \cdot Z \overset{d}{=} N(0 \cdot \sigma_i, 1 \cdot \sigma_i^2).$$

Det betyder at

$$(Y_i \mid X_i = x_i) \overset{d}{=} N(0, \sigma_i^2), \quad \text{med} \quad \sigma_i^2 = \beta x_i^2,$$

som er den ønskede model.

3. Tæthedsfunktionen for normalfordelingen,

$$(Y_i \mid X_i = x_i) \overset{d}{=} N(\mu_i, \sigma_i^2),$$

er givet som

$$f_{Y_i|X_i}\left(y_i \mid \mu_i, \sigma_i^2\right) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left\{-\frac{(y_i - \mu_i)^2}{2\sigma_i^2}\right\}, \quad y_i \in \mathbb{R}.$$

Med $\mu_i = 0$ og $\sigma_i^2 = \beta \cdot x_i^2$ indsat, fås

$$f_{Y_i|X_i}\left(y_i \mid x_i, \beta\right) = \frac{1}{\sqrt{2\pi\beta x_i^2}} \cdot \exp\left\{-\frac{y_i^2}{2\beta x_i^2}\right\}.$$

Udgangspunktet med de betingede fordelinger bygger på en antagelse om *eksogenitet.*

Da der antages *identiske betingede fordelinger* for alle i, er likelihood bidraget fra observation i givet ved

$$\ell(\beta \mid y_i, x_i) = f_{Y_i|X_i}\left(y_i \mid x_i, \beta\right) = \frac{1}{\sqrt{2\pi\beta x_i^2}} \cdot \exp\left\{-\frac{y_i^2}{2\beta x_i^2}\right\},$$

mens log-likelihood bidraget er givet ved

$$\begin{aligned}\log \ell(\beta \mid y_i, x_i) &= -\frac{1}{2}\log(2\pi\beta x_i^2) - \frac{y_i^2}{2\beta x_i^2} \\ &= -\frac{1}{2}\log(2\pi x_i^2) - \frac{1}{2}\log(\beta) - \frac{1}{\beta}\frac{1}{2}\frac{y_i^2}{x_i^2}.\end{aligned}$$

Det følger af *uafhængighed*, at likelihood funktionen er

$$
\begin{aligned}
L(\beta \mid y_1, ..., y_n, x_1, ..., x_n) &= \prod_{i=1}^{n} \ell(\beta \mid y_i, x_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\beta x_i^2}} \cdot \exp\left\{-\frac{y_i^2}{2\beta x_i^2}\right\},
\end{aligned}
$$

mens log-likelihood funktionen er

$$
\log L(\beta \mid y_1, ..., y_n, x_1, ..., x_n) = \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi x_i^2) - \frac{1}{2}\log(\beta) - \frac{1}{\beta}\frac{1}{2}\frac{y_i^2}{x_i^2}\right].
$$

4. For at udlede estimatoren som funktion ad de stokastiske variable, bruges

$$
\begin{aligned}
\log L(\beta \mid Y_1, ..., Y_n, X_1, ..., X_n) &= \sum_{i=1}^{n} \log \ell\left(\beta \mid Y_i, X_i\right) \\
&= \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi X_i^2) - \frac{1}{2}\log(\beta) - \frac{1}{\beta}\frac{1}{2}\frac{Y_i^2}{X_i^2}\right].
\end{aligned}
$$

Først findes score-bidraget som den første-afledte,

$$
\begin{aligned}
s_i(\beta) &= \frac{\partial \log \ell\left(\beta \mid Y_i, X_i\right)}{\partial \beta} \\
&= \frac{\partial}{\partial \beta}\left(-\frac{1}{2}\log(2\pi X_i^2) - \frac{1}{2}\log(\beta) - \frac{1}{\beta}\frac{1}{2}\frac{Y_i^2}{X_i^2}\right) \\
&= -\frac{1}{2}\left(\frac{1}{\beta} - \frac{1}{\beta^2}\frac{Y_i^2}{X_i^2}\right).
\end{aligned}
$$

Scoren er derfor

$$
\begin{aligned}
S(\beta) &= \sum_{i=1}^{n} s_i(\beta) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\left(\frac{1}{\beta} - \frac{1}{\beta^2}\frac{Y_i^2}{X_i^2}\right) \\
&= -\frac{1}{2}\left(\frac{n}{\beta} - \frac{1}{\beta^2}\sum_{i=1}^{n}\frac{Y_i^2}{X_i^2}\right)
\end{aligned}
$$

Dermed første-ordens betingelsen givet ved

$$
\begin{aligned}
S(\hat{\beta}_n) &= 0 \\
\frac{n}{\hat{\beta}_n} &= \frac{1}{\hat{\beta}_n^2}\sum_{i=1}^{n}\left(\frac{Y_i^2}{X_i^2}\right),
\end{aligned}
$$

og estimatoren er givet ved

$$
\hat{\beta}_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i^2}{X_i^2}\right).
$$

5. Dette spørgsmål er relativt teknisk og man kan næppe vente helt samme præcision som i besvarelsen her. Den anden-afledte af log-likelihood bidraget er givet ved

$$\begin{aligned}
\frac{\partial^2 \log \ell\left(\beta \mid Y_i, X_i\right)}{\partial\beta\partial\beta} &= \frac{\partial}{\partial\beta} s_i(\beta) \\
&= \frac{\partial}{\partial\beta}\left(-\frac{1}{2}\frac{1}{\beta} + \frac{1}{\beta^2}\frac{1}{2}\frac{Y_i^2}{X_i^2}\right) \\
&= \frac{1}{2}\frac{1}{\beta^2} - \frac{1}{\beta^3}\frac{Y_i^2}{X_i^2}.
\end{aligned}$$

Dermed er Hessematricen, evalueret i β_0, givet ved

$$\begin{aligned}
H_i(\beta_0) &= \left.\frac{\partial^2 \log \ell\left(\beta \mid Y_i, X_i\right)}{\partial\beta\partial\beta}\right|_{\beta=\beta_0} \\
&= \frac{1}{2}\frac{1}{\beta_0^2} - \frac{1}{\beta_0^3}\frac{Y_i^2}{X_i^2}.
\end{aligned}$$

Vi bruger nu, at informationen er givet ved

$$\begin{aligned}
I(\beta_0) &= -E\left[H_i(\beta_0) \mid X_i = x_i\right] \\
&= -E\left[\frac{1}{2}\frac{1}{\beta_0^2} - \frac{1}{\beta_0^3}\frac{Y_i^2}{X_i^2} \mid X_i = x_i\right] \\
&= -\frac{1}{2\beta_0^2} + \frac{1}{\beta_0^3}\frac{E\left[Y_i^2 \mid X_i = x_i\right]}{x_i^2}
\end{aligned}$$

Vi husker at $E(Y_i^2 \mid X_i = x_i) = \sigma_i^2 = \beta_0 x_i^2$, sådan at

$$\begin{aligned}
I(\beta_0) &= -\frac{1}{2\beta_0^2} + \frac{1}{\beta_0^3}\frac{\beta_0 x_i^2}{x_i^2} \\
&= -\frac{1}{2\beta_0^2} + \frac{1}{\beta_0^2} \\
&= \frac{1}{2\beta_0^2}.
\end{aligned}$$

Variansen på estimatoren, $V(\hat{\beta}_n)$, er derfor givet ved

$$V(\hat{\beta}_n) = \frac{I(\beta_0)^{-1}}{n} = \frac{2\beta_0^2}{n}.$$

6. Besvarelsen skal redegøre for, at konsistens betyder, at når n bliver stor nok vil estimatoren være vilkårligt tæt på den sande værdi, dvs. formelt

$$P\left(\left|\hat{\beta}_n - \beta_0\right| > \epsilon\right) \to 0 \quad \text{når} \quad n \to \infty,$$

for alle $\epsilon > 0$. Hvis der er valgt et ϵ, kan man derefter altid vælge et n sådan et afstanden mellem $\hat{\beta}_n$ og β_0 er mindre end ϵ, med sandsynlighed en.

Den asymptotiske fordeling vil være karakteriseret ved at for $n \to \infty$, gælder

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, \Omega_\beta), \quad \Omega_\beta = I(\beta_0)^{-1}.$$

Nogle vil foretrække at skrive resultatet som

$$\hat{\beta}_n \overset{a}{\sim} N(0, V(\hat{\beta}_n)), \quad V(\hat{\beta}_n) = n^{-1} I(\beta_0)^{-1}.$$

Besvarelsen skal forklare at den sande værdi indgår i grænseresultaterne fordi egenskaberne for estimatoren er udledt under antagelse af korrekt specifikation, dvs. i dette tilfælde

$$(Y_i \mid X_i = x_i) \overset{d}{=} N(0, \beta_0 x_i^2).$$

Da β_0 ikke er kendt i praksis indsættes estimatoren $\hat{\beta}_n$,og man anvender approksimationen

$$V(\hat{\beta}_n) = \frac{2\beta_0^2}{n} \approx \frac{2\hat{\beta}_n^2}{n}.$$

7. I det konkrete tilfælde med $\hat{\beta}_n = 2.037$ og $V(\hat{\beta}_n) = 0.0333$ er

$$\mathsf{se}(\hat{\beta}_n) = \sqrt{0.0333} = 0.182.$$

Så er 95% konfidens-intervallet givet ved

$$\begin{aligned}\left\{\underline{\beta} \leq \beta_0 \leq \overline{\beta}\right\} &= \left\{\hat{\beta}_n - 1.96 \cdot \mathsf{se}(\hat{\beta}_n) \leq \beta_0 \leq \hat{\beta}_n + 1.96 \cdot \mathsf{se}(\hat{\beta}_n)\right\} \\ &= \{2.037 - 1.96 \cdot 0.182 \leq \beta_0 \leq 2.037 + 1.96 \cdot 0.182\} \\ &= \{1.680 \leq \beta_0 \leq 2.394\}.\end{aligned}$$

Hypotesen, $\beta_0 = 2.5$, kan testes med hypoteserne

$$H_0 : \beta_0 = 2.5 \quad \text{mod} \quad H_A : \beta_0 \neq 2.5.$$

Teststørrelsen er givet ved

$$z_n(\beta_0 = 2.5) = \frac{\hat{\beta}_n - 2.5}{\mathsf{se}(\hat{\theta}_\mathsf{n})} = \frac{2.037 - 2.5}{0.182} = -2.544.$$

Den absolutte teststørrelse skal sammenligns med en kritisk værdi fra en standard normalfordeling. På et 5% signifikans-niveau er den kritiske værdi $\Phi^{-1}(0.975) = 1.96$ og hypotesen om $\beta_0 = 2.5$ afvises.