Eksamen på Økonomistudiet vinteren 2019-20

Sandsynlighedsteori og Statistik

2. årsprøve

14. februar, 2020

(3-timers prøve med hjælpemidler)

Rettevejledning

Opgave 1

1. Vi har $\mathbb{E}(Y|X=0) = 1 \cdot P(Y=1|X=0) + 2 \cdot (1 - P(Y=1|X=0)) \approx 0.839 + 2 \cdot 0.161 = 1.161$ da $P(Y=1|X=0) = \frac{P(Y=1,X=0)}{P(X=0)} = \frac{0.47}{0.47+0.09} \approx 0.839$.

2. Vi har

$$\begin{aligned}
\mathbb{E}(Y) &= 1 \cdot P(Y=1) + 2 \cdot P(Y=2) \\
&= 0.65 + 2 \cdot 0.35 \\
&= 1.35
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(X) &= 0 \cdot P(X=0) + 1 \cdot P(X=1) \\
&= 0.44
\end{aligned}$$

3. Vi har $Var(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$ hvor

$$\begin{aligned}
\mathbb{E}(Y^2) &= 1^2 \cdot P(Y=1) + 2^2 \cdot P(Y=2) \\
&= 0.65 + 4 \cdot 0.35 \\
&= 2.05
\end{aligned}$$

sådan at

$$\begin{aligned}
Var(Y) &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \\
&= 2.05 - 1.35^2 \\
&\approx 0.2275
\end{aligned}$$

4. Vi har

$$\begin{aligned}
Cov(X,Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
&= \sum_{y=1}^{2} \sum_{x=0}^{1} xyp(x,y) - 0.44 \cdot 1.35 \\
&= 0.7 - 0.594 \\
&= 0.106
\end{aligned}$$

.

Opgave 2

1. Middelværdien er

$$\mu_M = \mathbb{E}[M] = \mathbb{E}[1 + 2X - 3Y] = 1 + 2\mathbb{E}[X] - 3\mathbb{E}[Y] = 1 + 2\mu_X - 3\mu_Y$$

   og variansen er

$$\begin{aligned}\sigma_M^2 = Var(M) &= Var(1 + 2X - 3Y) \\ &= Var(2X) + Var(-3Y) + 2Cov(2X, -3Y) \\ &= Var(2X) + Var(-3Y) + 2 \cdot 2 \cdot (-3)Cov(X, Y) \\ &= 4\sigma_X^2 + 9\sigma_Y^2 - 12\sigma_{X,Y}\end{aligned}$$

   og summen af to Normalfordelte stokastiske variable er Normalfordelt, så

$$M \sim N(\mu_M, \sigma_M^2)$$

2. Den betingede middelværdi af Y givet X er 0 er

$$\begin{aligned}\mathbb{E}[Y|X = 0] &= \mu_Y + \sigma_{X,Y}\sigma_X^{-2}(0 - \mu_X) \\ &= 0.5 - (0.1/0.2) \cdot 0.1 \\ &= 0.45\end{aligned}$$

3. Vi har dette resultat fordi vi betinger på en værdi af X der er lavere end μ_X samtidig med at der er en positiv samvariation mellem Y og X. Det betyder, at når vi ved, at X er lavere end sin middelværdi, så er det også sandsynligt at Y er det og den betingede middelværdi bliver lavere end den ubetingede.

4. Her bruger vi, at vi ved at den marginale fordeling af X er $N(\mu_X, \sigma_X^2)$ og finder
   (a) $p(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp(-\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2})$
   (b) grænserne er $v = 0$ og $h = \infty$
   (c) $t^{-1}(z) = \log(z)$
   (d) $\frac{\partial t^{-1}(z)}{\partial z} = \frac{1}{z}$

sådan at vi får

$$q(z) = \begin{cases} \frac{1}{z\sqrt{2\pi\sigma_X^2}} \exp(-\frac{1}{2}\frac{(\log(z)-\mu_X)^2}{\sigma_X^2}) & \text{hvis } z \in (0,\infty) \\ 0 & \text{hvis } z \notin (0,\infty) \end{cases}$$

(hvilket er tæthedsfunktionen for en log-Normalt fordelt stokastisk variabel). Der gives også fuld point, hvis konkrete værdier af μ_X og σ_X^2 indsættes.

Opgave 3

1. Vi har, at likelihood bidraget for hver kunde er $\ell(\theta|y_i) = p(y_i) = [\exp(\theta)+1]y_i^{\exp(\theta)}$ og log-likelihood bidraget er $\log(\ell(\theta|y_i)) = \log(\exp(\theta)+1) + \exp(\theta)\log(y_i)$. Log-likelihood funktionen bliver, grundet uafhænighed mellem kunderne, således

$$\begin{aligned} \log L_n(\theta) = \log L(\theta|y_1,\dots,y_{178}) &= \sum_{i=1}^n \log(\ell(\theta|y_i)) \\ &= \sum_{i=1}^n [\log(\exp(\theta)+1) + \exp(\theta)\log(y_i)] \\ &= n\cdot\log(\exp(\theta)+1) + \exp(\theta)\sum_{i=1}^n \log(y_i) \end{aligned}$$

2. Første ordens betingelsen (FOC) for den givne model er at scoren skal være nul,

$$S(\hat{\theta}) = \left.\frac{\partial \log L_n(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}} = 0$$

hvor vi her har at

$$\begin{aligned} \frac{\partial \log L_n(\theta)}{\partial\theta} &= \sum_{i=1}^n \frac{\partial \log(\ell(\theta|Y_i))}{\partial\theta} \\ &= \sum_{i=1}^n s_i(\theta) \\ &= \sum_{i=1}^n \left[\frac{\exp(\theta)}{\exp(\theta)+1} + \exp(\theta)\log(Y_i)\right] \\ &= n\frac{\exp(\theta)}{\exp(\theta)+1} + \exp(\theta)\sum_{i=1}^n \log(Y_i) \end{aligned}$$

således at maximum likelihood **estimatoren**, $\hat{\theta}$, kan findes som løsningen til ligningen

$$
\begin{aligned}
n\frac{\exp(\hat{\theta})}{\exp(\hat{\theta})+1} + \exp(\hat{\theta})\sum_{i=1}^{n}\log(Y_i) = 0 \\
\Updownarrow \\
\frac{1}{\exp(\hat{\theta})+1} &= -\frac{1}{n}\sum_{i=1}^{n}\log(Y_i) \\
\Updownarrow \\
\hat{\theta}_Y &= \log\left(-\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\log(Y_i)} - 1\right)
\end{aligned}
$$

Ved at bruge $\frac{1}{178}\sum_{i=1}^{178}\log(y_i) = -0.3634$ kan vi udlede **estimatet** for vores data som

$$
\begin{aligned}
\hat{\theta}_y = \hat{\theta}(y_1,\dots,y_{178}) &= \log\left(-\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\log(y_i)} - 1\right) \\
&= \log\left(\frac{1}{0.3634} - 1\right) \\
&\approx 0.560
\end{aligned}
$$

3. Hesse-matricen er en skalar i dette tilfælde og givet ved den anden-afledte. Vi får at bidraget for borger i er

$$
\begin{aligned}
H_i(\theta) = \frac{\partial^2 \log(\ell(\theta|Y_i))}{\partial^2\theta} = \frac{\partial s_i(\theta)}{\partial\theta} &= \frac{\exp(\theta)(1+\exp(\theta)) - \exp(\theta)\exp(\theta)}{[1+\exp(\theta)]^2} + \exp(\theta)\log(Y_i) \\
&= \frac{\exp(\theta)}{[1+\exp(\theta)]^2} + \exp(\theta)\log(Y_i)
\end{aligned}
$$

og bruger at $I(\theta_0) \approx 0.4055$ sådan at variansen bliver

$$
\begin{aligned}
Var(\hat{\theta}_Y) &= \frac{1}{n}I(\theta_0)^{-1} \\
&\approx \frac{1}{178}0.4055^{-1} \\
&\approx 0.01385
\end{aligned}
$$

Det ses at standardafvigelsen bliver $se(\hat{\theta}_Y) = \sqrt{Var(\hat{\theta}_Y)} = \sqrt{0.01385} \approx 0.1177$.

4. Vi kan bruge den estimerede model til at beregne sandsynligheden for at en kundes

elforbrug udgør højest halvdelen af kundens samlede forbrug som

$$P(Y_i \leq 0.5) = \int_0^{0.5} p(y)dy = \int_0^{0.5} [\exp(\theta) + 1] y^{\exp(\theta)} dy = [y^{\exp(\theta)+1}]_0^{0.5} = 0.5^{\exp(\theta)+1}$$

sådan at hvis vi indsætter estimatet får vi $0.5^{\exp(0.560)+1} \approx 0.1486$. Sandsynligheden er altså ca. 15%.

5. Log-likelihood funktionen for den betingede model er

$$\begin{aligned}\log L_n(\theta, \delta) &= \log L(\theta, \delta | y_1, \ldots, y_{178}, d_1, \ldots, d_{178}) \\ &= \log \left(\prod_{i=1}^{178} [\exp(\theta + \delta d_i) + 1] y_i^{\exp(\theta + \delta d_i)} \right) \\ &= \sum_{i=1}^{178} \{ \log [\exp(\theta + \delta d_i) + 1] + \exp(\theta + \delta d_i) \log(y_i) \}\end{aligned}$$

6. Hvis $\delta \neq 0$ så er der en tendens til at der er forskel på el-forbrugets andel af det samlede forbrug i weekenderne i forhold til hverdagene. Hvis $\delta > 0$ så er der en tendens til at elforbruget tager en relativt større andel af forbruget i weekenden.

7. Vi kunne estimere den betingede model i STATA ved at skrive
```
mlexp( log(exp({theta}+{delta}*d) + 1) + exp({theta}+{delta}*d)*log(y))
```

8. Vi skal teste om der er en signifikant forskel på weekend og hverdage. Det svarer til hypotesen

$$\mathcal{H}_0 : \delta_0 = 0$$

med alternativ-hypotesen

$$\mathcal{H}_A : \delta_0 \neq 0.$$

Vi beregner vores z-statistik som

$$z_n(\delta_0 = 0) = \frac{\hat{\delta} - 0}{se(\hat{\delta})} = \frac{0.46155}{0.2419} \approx 1.9080.$$

Vi ved at $z_n(\delta_0 = 0) \overset{a}{\sim} \mathcal{N}(0, 1)$ under $\mathcal{H}_0$. Så vi kan beregne den kritiske værdi på et 5% signifikans-niveau, $\alpha = 0.05$, som $c = \Phi^{-1}(0.975) \approx 1.96$ (to-sidet test). Da $z_n < |c|$ kan vi **ikke** afvise på et 5% signifikansniveau, at der IKKE er en forskel på weekend og hverdage.
(p-værdien er $2 \cdot (1 - \Phi(1.9080)) \approx 0.0564$, hvilket er marginalt højere end de 5%)

Alternativt kan LR-test benyttes da vi har fået log-likelihood funktionen under den restrikterede og urestrikterede model. Her får vi LR-statistikken

$$LR(\delta_0 = 0) = 2(68.763549 - 66.865291) = 3.796516$$

og vi ved at under $\mathcal{H}_0$, så er $LR \overset{a}{\sim} \chi^2_1$ med 1 frihedsgrad. Den kritiske værdi er således $F^{-1}_{\chi^2_1}(0.95) = 3.84$. Da $LR(\delta = 0) < 3.84$ kan vi her heller **ikke** afvise at der IKKE er en forskel. p-værdien bliver næsten den samme $p = 1 - F_{\chi^2_1}(3.796516) \approx .0514$.
Vi kan altså ikke på et 5% signifikans-niveau afvise at der ikke er forskel på weekend og hverdage. Det er dog marginalt og på et 6% signifikans-niveau eller højere, vil vi kunne afvise at der ikke er en forskel.