

Sand Stat, 6.aflevering

ldg790 - Christian B. Gustafson

December 2020

Opgave 1

Vi ser at observationerne $\{y_i\}_{i=1}^n$ repræsenteres af tilfældige variable $\{Y_i\}_{i=1}^n$, hvor udfaldsrummet er $Y_i \in Y\{0, 1\}$. En relevant hvis ikke den eneste relevante fordeling vil for en binær variabel, Y_i vil være en Bernoulli-fordeling.

$$Y_i \sim \text{Bernoulli}(\theta)$$

Vi laver en antagelse om, at alle har samme ssh for at ryge, hvilket kan diskuteres om er rimeligt, eftersom at ssh kan afhænge af forskellige faktorer såsom miljø.

Desuden er der en antagelse om at observationerne er uafhængige, hvilket kan være rimeligt for et tilfældigt udfald, men kan også diskuteres om det er rimeligt, eftersom man fx er den eneste i ens omgangskreds der ikke ryger, så vil ssh for at man begynder at ryge være højere.

Opgave 2

$$L(\theta|y_1, \dots, y_n) = \prod_{i=1}^n l(\theta|y_i) = \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i}$$

Dette er vores likelihood funktion. Nu tager jeg log på begge sider for at finde den funktion.

$$\begin{aligned} \log(L(\theta|y_1, \dots, y_n)) &= \log\left(\prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i}\right) = \sum_{i=1}^n \log(\theta^{y_i} (1 - \theta)^{1-y_i}) \\ &= \sum_{i=1}^n y_i \log(\theta) + (1 - y_i) \log(1 - \theta) \end{aligned}$$

Hvilket er vores log-likelihood funktion.

Opgave 3

For at finde maximum likelihood estimatoren, vil jeg differentiere vores log funktion og sætte lig 0:

$$\begin{aligned}\frac{\partial \log(L(\theta|y_1, \dots, y_n))}{\partial \theta} &= \sum_{i=1}^n \frac{y_i}{\theta} - \frac{1 - Y_i}{1 - \theta} \\&= \frac{1}{\theta} \sum_{i=1}^n Y_i - \frac{1}{1 - \theta} (n - \sum_{i=1}^n Y_i) = 0 \Leftrightarrow \\ \frac{n}{1 - \theta} &= \frac{1}{\theta} \sum_{i=1}^n Y_i + \frac{1}{1 - \theta} \sum_{i=1}^n Y_i \Leftrightarrow \\ \frac{n}{1 - \theta} &= \sum_{i=1}^n Y_i \left(\frac{1}{\theta} + \frac{1}{1 - \theta} \right) \Leftrightarrow \\ \frac{\theta(1 - \theta)}{1 - \theta} &= \frac{1}{n} \sum_{i=1}^n \Leftrightarrow\end{aligned}$$

OPS. Der skal stå $\hat{\theta}$ alle de steder hvor vi har et theta.

$$\hat{\theta}(Y_1, \dots, Y_n) = \frac{1}{n} \sum_{i=1}^n$$

Opgave 4

$$\begin{aligned}\log(L(\theta|y_1, \dots, y_n)) &= \sum_{i=1}^n y_i \log(\theta) + (1 - y_i) \log(1 - \theta) \\ \sum_{i=1}^n y_i \log(\theta) (n - \sum_{i=1}^n y_i) \log(1 - \theta) \\ R_n \log(\theta) (n - R_n) \log(1 - \theta) &= \log(\theta | R_n) = \sum_{i=1}^n y_i\end{aligned}$$

Som vi kan se er der $n=153$ der har svaret på spørgeskemaet, hvoraf 34 personer ryger, hvilket giver os et estimat på:

$$\hat{\theta}(Y_1, \dots, Y_n) = \frac{1}{126} \cdot 34 \approx 0.27$$

Opgave 5

Benytter bernoulli igen:

$$Y_i|X_i = x_i \sim \text{Bernoulli}(\theta_i)$$

$$L(\theta_1, \theta_2|y_1, \dots, y_n, X_1, \dots, X_n) = \prod_{i=1}^n (\theta_1^{x_i} \theta_2^{1-x_i})_i^Y ((1-\theta_1)^{x_i} (1-\theta_2)^{1-x_i})^{1-Y_i}$$

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \in \Theta = \{\theta \in R^2 | 0 < \theta_1, \theta_2 < 1\}$$

$$\log(L(\theta_1, \theta_2|y_1, \dots, y_n, X_1, \dots, X_n)) = \log\left(\prod_{i=1}^n (\theta_1^{x_i} \theta_2^{1-x_i})_i^Y ((1-\theta_1)^{x_i} (1-\theta_2)^{1-x_i})^{1-Y_i}\right)$$

$$= \sum_{i=1}^n \log(\theta_1^{x_i} \theta_2^{1-x_i})_i^Y ((1-\theta_1)^{x_i} (1-\theta_2)^{1-x_i})^{1-Y_i}$$

$$= \sum_{i=1}^n \log(Y_i(X_i \log \theta_1 + (1-X_i) \log \theta_2) + (1-Y_i)(X_i \log(1-\theta_1) + (1-X_i) \log(1-\theta_2)))$$

$$\begin{pmatrix} s(\hat{\theta}_1) \\ s(\hat{\theta}_2) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \left(\frac{Y_i X_i}{\theta_1} - \frac{X_i(1-Y_i)}{1-\theta_1} \right) \\ \sum_{i=1}^n \left(\frac{Y_i(1-X_i)}{\theta_2} - \frac{(1-Y_i)(1-X_i)}{1-\theta_2} \right) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Leftrightarrow$$

$$\begin{pmatrix} \frac{1}{\theta_1} \sum_{i=1}^n Y_i X_i \\ \frac{1}{\theta_2} \sum_{i=1}^n Y_i (1-X_i) \end{pmatrix} = \begin{pmatrix} \frac{1}{1-\theta_1} \sum_{i=1}^n X_i (1-Y_i) \\ \frac{1}{1-\theta_2} \sum_{i=1}^n (1-Y_i) (1-X_i) \end{pmatrix} \Leftrightarrow$$

$$\begin{pmatrix} \frac{1-\theta_1}{\theta_1} \\ \frac{1-\theta_2}{\theta_2} \end{pmatrix} = \begin{pmatrix} \frac{\sum_{i=1}^n X_i (1-Y_i)}{\sum_{i=1}^n (1-X_i - Y_i + X_i Y_i)} \\ \frac{\sum_{i=1}^n (1-Y_i) (1-X_i)}{\sum_{i=1}^n (Y_i - X_i Y_i)} \end{pmatrix}$$

$$\begin{pmatrix} \frac{1}{\theta_1} \\ \frac{1}{\theta_2} \end{pmatrix} = \begin{pmatrix} \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n (1-X_i)} \\ \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n (Y_i - X_i Y_i)} \end{pmatrix}$$

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i (1-X_i)} \\ \frac{\sum_{i=1}^n X_i (1-Y_i)}{\sum_{i=1}^n (1-X_i)} \end{pmatrix}$$

Herfra er estimatoren fundet for $\hat{\theta}(Y_1, \dots, Y_n)$ og jeg vil nu finde estimatet, $\hat{\theta}(y_1, \dots, y_n)$:

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \frac{9}{44} \\ \frac{22}{82} \end{pmatrix}$$

Estimatet $\hat{\theta}_1 = 0,2045$ som angiver forholdet mellem hvor mange ryger og som bruger hjelm og hvor mange der bruger hjelm i alt.

Estimatet $\hat{\theta}_2 = 0,2683$ er forholdet mellem hvor mange rygere der ikke bruger hjelm og hvor mange der ikke bruger hjelm i alt.

Opgave 6

$$\theta = \begin{pmatrix} \gamma \\ \delta \end{pmatrix} \in \Theta = \{\theta \in R^2 | 0 < \gamma, \delta < 1\}$$

Hvor delta ville beskrive den ekstra ssh for, at person i der mindst ryger til fester ikke bruger hjelm.