1. (a) $\int_0^1 P(S=1;\theta)d\theta + \int_0^1 P(S=0;\theta)d\theta = 1$

$$LHS = c\left(\int_0^1 \theta^1(1-\theta)^{1-1}d\theta + \int_0^1 \theta^0(1-\theta)^{1-0}d\theta\right)$$

$$= c\left(\int_0^1 \theta\, d\theta + \int_0^1 (1-\theta)d\theta\right) = 1$$

So $c\times 1 = 1,\ c = 1$

(b) $E(S) = 1\times P(S=1;\theta) + 0\times P(S=0;\theta) = 1\times\theta + 0\times(1-\theta) = \theta$

(c) $E\left[(S-\bar{S})^2\right] = (1-\theta)^2\times P(S=1;\theta) + (0-\theta)^2\times P(S=0;\theta)$

$= \theta(1-\theta)$

2. (a) Since $\int_0^1 P_0(\theta)d\theta = \int_0^1 c'\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = c' B(\alpha,\beta) = 1$

So $c' = \dfrac{1}{B(\alpha,\beta)}$

(b)

$$P(S=s) = \int_0^1 P(S=s|\theta)\cdot P_0(\theta)d\theta = \int_0^1 \theta^s(1-\theta)^{1-s}c'\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta$$

$$= \frac{1}{B(\alpha,\beta)}\int_0^1 \theta^{\alpha+s-1}(1-\theta)^{1-s+\beta-1}d\theta$$

$$= \frac{1}{B(\alpha,\beta)}\int_0^1 \theta^{\alpha+s-1}(1-\theta)^{\beta-s}d\theta = \frac{B(\alpha+s,\beta-s+1)}{B(\alpha,\beta)}$$

(c) $E(S) = 1\times P(S=1) + 0\times P(S=0) = P(S=1) = \dfrac{B(\alpha+1,\beta-1+1)}{B(\alpha,\beta)}$

$$= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+1)\Gamma(\beta)/\Gamma(\alpha+\beta+1)}{\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)} = \frac{\alpha}{\alpha+\beta}$$

(d) $E(S^2) = 1^2\times P(S=1) + 0^2\times P(S=0) = \dfrac{\alpha}{\alpha+\beta}$ $\quad\therefore Var(S) = \dfrac{\alpha}{\alpha+\beta} - \left(\dfrac{\alpha}{\alpha+\beta}\right)\left(\dfrac{\alpha}{\alpha+\beta}\right) = \dfrac{\alpha\beta}{(\alpha+\beta)^2}$

3. (a) Since $S \sim Ber(\theta)$, so $P(S|\theta) = \theta^{S}(1-\theta)^{1-S}$

$$L(\{S_i\};\theta) = \prod_{i=1}^{N} \theta^{S_i}(1-\theta)^{1-S_i}$$ and $$\hat{\ell}(\{S_i\};\theta) = \frac{1}{N}\left(\log\left(\prod_{i=1}^{N}\theta^{S_i}(1-\theta)^{1-S_i}\right)\right)$$

$$\therefore \hat{\ell}(\{S_i\};\theta) = \frac{1}{N}\left(\sum_{i=1}^{N} S_i \log\theta + \sum_{i=1}^{N}(1-S_i)\log(1-\theta)\right)$$

$$= \frac{1}{N}\left(\hat{N}_1 \log\theta + \hat{N}_0 \log(1-\theta)\right)$$

$$\therefore \frac{\partial \hat{\ell}}{\partial\theta} = \frac{\hat{N}_1}{N}\cdot\frac{1}{\theta} + \frac{\hat{N}_0}{N}\cdot\frac{-1}{1-\theta} = 0 \Rightarrow \hat{\theta} = \frac{\hat{N}_1}{\hat{N}_1+\hat{N}_0}$$

(b)
$$\begin{cases} P(S=1 \mid \hat{N}_0, \hat{N}_1) = \hat{\theta} = \dfrac{\hat{N}_1}{\hat{N}_0+\hat{N}_1} \\ P(S=0 \mid \hat{N}_0, \hat{N}_1) = 1-\hat{\theta} = \dfrac{\hat{N}_0}{\hat{N}_0+\hat{N}_1} \end{cases}$$

(c) From the results of question (a) and (b), for next experiment,

$$\begin{cases} P(S=1 \mid \hat{N}_0=0, \hat{N}_1=1) = \dfrac{1}{1+0} = 1 \\ P(S=0 \mid \hat{N}_0=0, \hat{N}_1=1) = \dfrac{0}{1+0} = 0 \end{cases}$$

It seems ridiculous, but also make sense because we have too few samples and observations. If we have done more experiments, the estimator $\hat{\theta}$ would converge to the true value.

4. Since $S \sim Ber(\theta)$, $\theta \sim Beta(\alpha, \beta)$

ⓐ

So $P(\theta | \{S_i\}) = \dfrac{\prod_{i=1}^{N} P(S_i|\theta) P_0(\theta)}{P(\{S_i\})} = \dfrac{\prod_{i=1}^{N}\left[\theta^{S_i}(1-\theta)^{1-S_i}\right] Beta(\alpha,\beta)}{P(\{S_i\})} = \dfrac{\theta^{\hat{N}_1}(1-\theta)^{\hat{N}_0} Beta(\alpha,\beta)}{P(\{S_i\})} = \cancel{C \cdot \theta^{\hat{N}_1}(1-\theta)^{\hat{N}_0}}$

$$= \frac{\theta^{\hat{N}_1}(1-\theta)^{\hat{N}_0} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}}{P(\{S_i\})\, B(\alpha,\beta)}$$

and $\int_0^1 d\theta P(\theta|\{S_i\}) = 1$

$$= C \cdot \theta^{(\hat{N}_1+\alpha-1)}(1-\theta)^{(\hat{N}_0+\beta-1)}$$

So $\int_0^1 d\theta \cdot C \cdot \theta^{\hat{N}_1+\alpha-1}(1-\theta)^{\hat{N}_0+\beta-1} = 1$

$\therefore C \cdot B(\hat{N}_1+\alpha, \hat{N}_0+\beta) = 1$ $\quad\therefore C = \dfrac{1}{B(\hat{N}_1+\alpha, \hat{N}_0+\beta)}$

$\therefore P(\theta|\{S_i\}) = \dfrac{1}{B(\hat{N}_1+\alpha, \hat{N}_0+\beta)} \theta^{\hat{N}_1+\alpha-1}(1-\theta)^{\hat{N}_0+\beta-1}$

ⓑ

$\therefore P(S=1|\{S_i\}) = \int_0^1 \theta \cdot P(\theta|\{S_i\}) d\theta = \int_0^1 C \cdot \theta^{(\hat{N}_1+\alpha-1+1)}(1-\theta)^{(\hat{N}_0+\beta-1)} d\theta = \dfrac{B(\hat{N}_1+\alpha+1, \hat{N}_0+\beta)}{B(\hat{N}_1+\alpha, \hat{N}_0+\beta)}$

$$= \frac{\Gamma(\hat{N}_1+\alpha+1)\Gamma(\hat{N}_0+\beta)}{\Gamma(\hat{N}_1+\alpha+1+\hat{N}_0+\beta)} \Big/ \frac{\Gamma(\hat{N}_1+\alpha)\Gamma(\hat{N}_0+\beta)}{\Gamma(\hat{N}_1+\alpha+\hat{N}_0+\beta)} = \frac{\hat{N}_1+\alpha}{\hat{N}_1+\hat{N}_0+\alpha+\beta}$$

$$P(S=0|\{S_i\}) = 1 - \frac{\hat{N}_1+\alpha}{\hat{N}_1+\hat{N}_0+\alpha+\beta} = \frac{\hat{N}_0+\beta}{\hat{N}_1+\hat{N}_0+\alpha+\beta}$$

ⓒ

$$E(S|\hat{N}_1,\hat{N}_0;\alpha,\beta) = 1 \times P(S=1|\hat{N}_0,\hat{N}_1;\alpha,\beta) + 0 \times P(S=0|\hat{N}_0,\hat{N}_1;\alpha,\beta)$$

$$= \frac{\hat{N}_1+\alpha}{\hat{N}_0+\hat{N}_1+\alpha+\beta}$$

ⓓ

$$\begin{cases} P(S=1|\{S_i\}) = \dfrac{1+\alpha}{1+\alpha+\beta} \\ P(S=0|\{S_i\}) = \dfrac{\beta}{1+\alpha+\beta} \end{cases}$$

At this senario, the Prior distribution parameters (α, β) counts, and would fix some problem by giving a prior to the event' belief that at which stage the event would happen.

And as the experiments go on, we gather more and more data, we would get more precise predictions.

5. $$y = \sum_{d=1}^{D} x_d \theta_d + \epsilon$$

(a) For 1 observation, $E(y_i) = E\left(\sum_{d=1}^{D} x_{id}\theta_d\right) + E(\epsilon)$

$$= \sum_{d=1}^{D} x_{id}\theta_d$$

Since $\epsilon \sim N(0,\sigma^2)$ so $y_i \sim N(E(y),\sigma^2)$, $f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \sum_{d=1}^{D} x_{id}\theta_d)^2}{2\sigma^2}\right)$

$$P(y_i \le y \mid x_{i,d};\theta) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y' - \sum_{d=1}^{D} x_{id}\theta_d)^2}{2\sigma^2}\right) dy'$$

(b) Since for each i,j, ϵ_i is independent of ϵ_j, so y_i is also independent of y_j

so $$E(\vec{y}) = \begin{pmatrix} \sum_{d=1}^{D} x_{1d}\theta_d \\ \vdots \\ \sum_{d=1}^{D} x_{Nd}\theta_d \end{pmatrix} \triangleq \vec{\mu}$$

and $$\mathrm{Cov}(y_i,y_j) = \begin{cases} 0 & \text{if } i \ne j \\ \sigma^2 & \text{if } i = j \end{cases}$$ so $\Sigma = \sigma^2 I_N$.

So the join density $P(\{y_i\} \mid \{x_{i,d}\};\theta)$ of all the N observations $\{y_i, x_{id}\}_{i=1}^{N}$ should be:

$$f(\vec{y}) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\vec{y}-\vec{\mu})^T \Sigma^{-1} (\vec{y}-\vec{\mu})\right)$$

where $\vec{\mu}$ and Σ are defined before.

(c) $\hat{l}(\theta;\{y_i,x_{id}\}) = \frac{1}{N}\sum_{i=1}^{N} \log f(y_i)$ Since $\log f(y_i) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(\exp\left(-\frac{(y_i - \sum_{d=1}^{D} x_{id}\theta_d)^2}{2\sigma^2}\right)\right)$

So $$\hat{l}(\theta;\{y_i,x_{id}\}_{i=1}^{N}) = -\frac{1}{2}\log(2\pi\sigma^2) + \left(-\frac{1}{2\sigma^2}\right) \times \frac{1}{N} \times \left(\sum_{i=1}^{N} (y_i - \sum_{d=1}^{D} x_{id}\theta_d)^2\right)$$

(d) $$E(\theta;\{y_i,x_{id}\}_{i=1}^{N}) = -\hat{l}(\theta;\{y_i,x_{id}\}_{i=1}^{N})$$
$$= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2 N}\left(\sum_{i=1}^{N}(y_i - \sum_{d=1}^{D} x_{id}\theta_d)^2\right)$$

(e) $$\frac{\partial E}{\partial \theta_d} = \frac{1}{\sigma^2 N}\sum_i \left(\sum_{d'=1}^{D} x_{id'}\theta_{d'} - y_i\right) x_{id}.$$

(f) $(X^T X)\theta = X^T Y$

(g) The OLS method is to some extent equivalent to the ~~method~~ MLE method given that the error terms are normally distributed.

由 扫描全能王 扫描创建

6. Similar to questions 5, we have:

(a) $P(y_i|x_i;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\sum_{d=1}^{D} h_d(x)\theta_d)^2}{2\sigma^2}\right)$

(b) $\hat{\ell}(\theta; \{y_i, x_{i,d}\}_{i=1}^{N}) = \frac{1}{-2N\sigma^2}\left(\sum_{i=1}^{N}(y_i - \sum_{d=1}^{D} h_d(x)\theta_d)^2\right) - \frac{1}{2}\log(2\pi\sigma^2)$

$$E(\theta; \{y_i, x_{i,d}\}_{i=1}^{N}) = -\hat{\ell}(\theta; \{y_i, x_{i,d}\}_{i=1}^{N})$$

$$= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2 N}\left(\sum_{i=1}^{N}(y_i - \sum_{d=1}^{D} h_d(x)\theta_d)^2\right)$$

$$= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2 N}(Y-H\theta)^T(Y-H\theta)$$

(c) Similarly, since for linear X, $(X^TX)\theta = X^TY$

now we can consider H as new X,

so $(H^TH)\theta = H^TY$

(d) It is the same.

Once we get the matrix H computed, the dimensionality of x does not matter.

(e) The function can be expressed, so the matrix H is not full-ranked.

Some rows/columns of the matrix is redundant.

We need to do Q-R decomposition of H.

Say $H = QR$, where $Q^TQ = I$, and R is upper-triangle matrix

7. (a) $P(\Theta) = \mathcal{N}(\theta; 0, \frac{\sigma^2}{\lambda} I)$

$$E = 2\log 2\pi\sigma^2 + \frac{1}{2\sigma^2 N}(Y - X\theta)^T(Y - X\theta) + \frac{\lambda}{2\sigma^2 N}\theta^T\theta$$

$$\frac{\partial E}{\partial \theta_d} = \frac{1}{\sigma^2 N}\left\{\sum_{i=1}^{N} x_{id}\left(\sum_d (x_{id}\theta_d - y_i)\right) + \lambda\theta_d\right\}$$

or $$\frac{\partial E}{\partial \theta} = \frac{1}{\sigma^2 N}\left\{X^T(X\theta - Y) + \lambda\theta\right\}$$

(b) $P_0(\theta) = Ce^{-\lambda\sum_d|\theta_d|}$

$$E = 2\log 2\pi\sigma^2 + \frac{1}{2\sigma^2 N}\sum_i\left(y_i - \sum_d x_{id}\theta_d\right)^2 + \frac{\lambda}{N}\sum_d|\theta_d|$$

$$\frac{\partial E}{\partial \theta_d} = \frac{1}{\sigma^2 N}\left\{\sum_i x_{id}\left(\sum_{d'} x_{id'}\theta_{d'} - y_i\right)\right\} + \frac{\lambda}{N}\operatorname{sgn}\theta_d$$

(c) $P_0(\theta) = Ce^{-\lambda_1\sum_d|\theta_d|}\mathcal{N}(\theta; 0, \frac{\sigma^2}{\lambda_2} I)$

$$E = 2\log 2\pi\sigma^2 + \frac{1}{2\sigma^2 N}\sum_i\left(y_i - \sum_d x_{id}\theta_d\right)^2 + \frac{\lambda_1}{N}\sum_d|\theta_d| + \frac{\lambda_2}{\sigma^2 N}\sum_d\theta_d^2.$$

$$\frac{\partial E}{\partial \theta_d} = \frac{1}{\sigma^2 N}\left\{\sum_i x_{id}\left(\sum_{d'} x_{id'}\theta_{d'} - y_i\right)\right\} + \frac{\lambda_1}{N}\operatorname{sgn}\theta_d + \frac{\lambda_2}{\sigma^2 N}\theta_d$$