SOLUTION

(a) $\boxed{\text{Firstly}}$ we ~~assume~~ prove $\theta^* = \max_\theta L(D,\theta)$ if $\theta^* = \max_\theta \ell(D,\theta)$.

- We assume the contrary $\theta' = \max \ell(D,\theta)$ and $\theta'' = \max_\theta L(D,\theta)$, where $\theta' \neq \theta''$. Because of the assumption, we have

$$\ell(D,\theta') > \ell(D,\theta'')$$

Since exponential function is strictly increasing, we have

$$e^{\ell(D,\theta')} > e^{\ell(D,\theta'')}.$$

which is equivalent to

$$L(D,\theta') > L(D,\theta'').$$

However, $\theta' \neq \theta''$ and $\theta'' = \max_\theta L(D,\theta)$, ~~We can~~ which implies a contradiction. We can conclude that θ^* solves $\max_\theta L(D,\theta)$ if it solves $\max_\theta \ell(D,\theta)$

$\boxed{\text{Second}}$ ~~Similarly~~. we ~~assume~~ prove $\theta^* = \max_\theta \ell(D,\theta)$ if $\theta^* = \max_\theta L(D,\theta)$.

- We assume the contrary: $\theta' = \max_\theta L(D,\theta)$, $\theta'' = \max_\theta \ell(D,\theta)$, where $\theta' \neq \theta''$, Because of the assumption, we have

$$L(D,\theta') > L(D,\theta'')$$

Since the logarithm function is strictly increasing, we have

$$\log L(D,\theta') > \log L(D,\theta'')$$

which is equivalent to

$$\ell(D,\theta') > \ell(D,\theta'').$$

Howeve, $\theta' \neq \theta''$, $\theta'' = \max_\theta \ell(D,\theta)$, which implies a contradiction. We can conclude the θ^* solves $\max_\theta \ell(D,\theta)$ if $\theta^* = \max_\theta L(D,\theta)$.

q.e.d.

(b) Maximising log-liklihood is better because

i) Summing is less expensive than multiplication, for computers and by hand

ii) Likelihood can be very small, whick can result in float number underflow. (Sometimes even float 64 is not enough)

(c) $\Pr(y=0 \mid X, \theta) = 1 - \Pr(y=1 \mid X, \theta) = 1 - \frac{1}{1+\exp(X\theta)}$

$$= \frac{1+\exp(X\theta) - 1}{1+\exp(X\theta)} = \frac{\exp(X\theta)}{1+\exp(X\theta)}$$

~~(d) $\log \Pr(y, X, \theta) = y \log \Pr($~~

(d) $\log \Pr(y, X, \theta) = \log P(X,\theta)^{y} + \log (1-P(X,\theta))^{1-y}$

$$= \log P(X,\theta)^{y} \cdot (1-P(X,\theta))^{1-y}$$

y is a binary variable. When $y=0$:

~~los~~ $\log \Pr(y, x, \theta) = \log (1-P(X,\theta))$.

When $y=1$

$\log \Pr(y, x, \theta) = \log P(X,\theta)^{y}$.

So we write

$\log \Pr(y, X, \theta) = \cancel{\theta} y \log P(X,\theta)^{y} + (1-y) \log (1-P(X,\theta))$

q.e.d.

~~natural logarithm~~.

(e) Assume the logarithm function is natural logarithm.

$$\nabla_\theta \ell(D,\theta) = \sum_{i=1}^{N} \left[\nabla_\theta\, y_i \log p(X_i,\theta) + \nabla_\theta (1-y_i) \log(1-p(X_i,\theta)) \right]$$

$$= \sum_{i=1}^{N} \left[y_i\, \nabla_\theta \log P(X_i,\theta) + (1-y_i)\, \nabla_\theta \log(1-p(X_i,\theta)) \right]$$

Since $P(X_i,\theta) = \frac{1}{1+\exp(X\theta)}$, we have $\nabla_\theta \log P(X_i,\theta) = \cancel{P(X_i,\theta)}\,(1-P(X_i,\theta)) \cdot X_i$

and $\nabla_\theta \log(1-P(X_i,\theta)) = -P(X_i,\theta) \cdot X_i$

$$\cancel{\nabla_\theta \ell(D,\theta) = \sum_{i=1}^{N} [y_i \log P(X,\theta)(1-P(X,\theta))X + (1-y_i) \log}$$

$$\nabla_\theta \ell(D,\theta) = \sum_{i=1}^{N} \left[y_i (1-P(X_i,\theta))X + (1-y_i) \cdot (-P(X_i,\theta)X) \right]$$

$$= \sum_{i=1}^{N} \left[y_i X_i \underline{- y_i P(X_i,\theta)X} - P(X_i,\theta)X_i \underline{+ y_i P(X_i,\theta)X_i} \right]$$

$$= \sum_{i=1}^{N} \left[y_i X_i - P(X_i,\theta)X_i \right] = \sum_{i=1}^{N} X(y_i - P(X_i,\theta))$$

Its Hessian matrix is

$$H_\theta(X) = \begin{bmatrix} \sum_{i=1}^{N} X(y_i - P(X_i;\theta_1)) \\ \sum_{i=1}^{N} X(y_i - P(X_i;\theta_2)) \\ \vdots \\ \sum_{i=1}^{N} X(y_i - P(X_i,\theta_N)) \end{bmatrix}$$

, where N is the dimension of the parameter vector θ.