سوال ۳:

در این مسئله، داده‌ها به صورت $y_i = wx_i + \epsilon_i$ مدل شده‌اند که در آن $\epsilon_i$ نویز گوسی با توزیع $N(0,1)$ است. بنابراین، احتمال شرطی $y_i$ با توجه به $x_i$ به صورت زیر خواهد بود:

$$P(y_i|x_i, w) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i - wx_i)^2}{2}\right)$$

چون داده‌ها مستقل و هم‌توزیع (i.i.d) هستند، احتمال کل مجموعه داده $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ به صورت ضرب احتمالات تک‌تک داده‌ها خواهد بود:

$$P(D|w) = \prod_{i=1}^{n} P(y_i|x_i, w) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i - wx_i)^2}{2}\right)$$

برای محاسبه log-likelihood، لگاریتم احتمال کل را می‌گیریم:

$$\log P(D|w) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i - wx_i)^2}{2}\right)\right)$$

با ساده‌سازی، داریم:

$$\log P(D|w) = \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi) - \frac{(y_i - wx_i)^2}{2}\right)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - wx_i)^2$$

چون $\log(2\pi)$ و ثابت‌ها مستقل از $w$ هستند، بیشینه کردن log-likelihood با بیشینه کردن عبارت زیر معادل است:

$$\arg\max_w \log P(D|w) = \arg\max_w -\frac{1}{2}\sum_{i=1}^{n}(y_i - wx_i)^2$$

که این معادل است با کمینه کردن مجموع مجذور خطاها:

$$\arg\max_w \log P(D|w) = \arg\min_w \sum_{i=1}^{n}(y_i - wx_i)^2$$

بنابراین نشان دادیم که بیشینه کردن log-likelihood معادل است با کمینه کردن مجموع مجذور خطا.