1.

因為x_n 和x_m 是獨立的，所以有：

$$\mathrm{E}[x_n x_m] = \mathrm{E}[x_n]\mathrm{E}[x_m]$$

由於x_n 和x_m 都來自於平均值為μ 、方差為σ^2 的高斯分布，因此有：

$$\mathrm{E}[x_n] = \mathrm{E}[x_m] = \mu$$

綜合上面兩個公式，得到：

$$\mathrm{E}[x_n x_m] = \mathrm{E}[x_n]\mathrm{E}[x_m] = \mu\mu = \mu^2$$

如果$n = m$，那麼：

$$\mathrm{E}[x_n x_m] = \mathrm{E}[x_n^2] = \mathrm{Var}[x_n] + \mathrm{E}[x_n]^2 = \sigma^2 + \mu^2$$

否則，$n \neq m$，x_n 和x_m 獨立，所以它們的協方差為 0，即：

$$\mathrm{E}[x_n x_m] = 0$$

所以有：

$$\mathrm{E}[x_n x_m] = \mu^2 + I_{nm}\sigma^2 - (3)$$

當$n = m$時，$I_{nm} = 1$，否則$I_{nm} = 0$，得證。

為了證明$\mathrm{E}[\mu_{ML}] = \mu$，我們需要顯示最大近似估計平均值參數的期望值等於數據生成過程的真實平均值。

對於具有未知平均值μ和已知方差σ^2的高斯分布，其概似函數可以寫為：

$$\mathrm{L}(\mu|x_1, x_2, \ldots, x_n) = (2\pi\sigma^2)^{-\left(\frac{n}{2}\right)} \; e^{-(1/2\sigma^2)\Sigma(x_i-\mu)^2}$$

對數近似函數為：

$$\ln \; \mathrm{L}(\mu|\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n) = (2\pi\sigma^2)^{-\left(\frac{n}{2}\right)} \; e^{-\left(\frac{1}{2\sigma^2}\right)\Sigma(\mathrm{x_i}-\mu)^2}$$

$$= -\left(\frac{n}{2}\right)\ln(2\pi) - \left(\frac{n}{2}\right)\ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)\Sigma(\mathrm{x_i} - \mu)^2$$

為了找到 μ 的最大概似估計，我們對對數概似函數對 μ 求導數並將其設置為零：

$$\frac{d}{du}\ln \; \mathrm{L}(\mu|\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n) = \left(\frac{1}{\sigma^2}\right)\Sigma(\mathrm{x_i} - \mu) = 0$$

解出 μ，我們得到：

$$\mu_{ML} = \frac{1}{n}\Sigma\mathrm{x_i}$$

μ 的最大概似估計值期望可以通過對μ_{ML}取期望值得到：

$$\mathrm{E}[\mu_{ML}] = \mathrm{E}\left[\frac{1}{n}\Sigma\mathrm{x_i}\right] = \frac{1}{n}\Sigma E[\mathrm{x_i}] = \mu$$

在這裡，我們使用了每個數據點$\mathrm{x_i}$的期望值等於真實平均值 μ 的事實。

為了證明$\mathrm{E}\left[\sigma_{ML}^2\right] = \frac{N-1}{N}\sigma^2$，可以通過將 μ 的最大概似估計值代入對數概似函數

並對σ^2進行最大化來找到變量參數的最大概似估計：

$$\ln\ L(\sigma^2|x_1, x_2, \dots, x_n) = -\left(\frac{n}{2}\right)\ln(2\pi) - \left(\frac{n}{2}\right)\ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)\Sigma(x_i - \mu)^2$$

$$\frac{d}{d\sigma^2}\ln\ L(\sigma^2|x_1, x_2, \dots, x_n) = -\left(\frac{n}{2}\right)\sigma^2 + \frac{1}{2\sigma^4}\Sigma(x_i - \mu)^2 = 0$$

解出σ^2，我們得到：

$$\sigma^2_{ML} = \frac{1}{n}\Sigma(x_i - \mu)^2$$

σ^2的最大概似估計期望可以通過對σ^2_{ML}取期望值得到：

$$E[\sigma^2_{ML}] = E[\frac{1}{n}\Sigma(x_i - \mu)^2]$$

使用$\mathrm{Var}(x_i)$，因此，我們得出 MLE 的期望值是偏小估計量，即

$$E[\sigma^2_{ML}] = \frac{N-1}{N}\sigma^2$$

得證。

2.

我們可以使用期望的線性來證明兩個獨立隨機向量a和b的和$y = a + b$ 的平均值由每個變量的平均值分別組成，

即：

$$E[y] = E[a + b] = E[a] + E[b]$$

其中E[y]是隨機變量y的期望值，E[a]和E[b]分別是隨機變量a和b的期望值，這是由於隨機變量的和之期望值等於它們各自期望值的總和。

為了顯示y的協方差矩陣，由a和b的協方差矩陣之和給出，我們可以使用協方差的定義，即：

$$\mathrm{cov}(y) = E[(y - E[y])(y - E[y])^T]$$

其中cov(y)是隨機向量y的協方差矩陣，E[y]是y的平均值。

展開上式，並使用期望的線性性質，我們得到：

$$\mathrm{cov}(y) = E[(a - E[a] + b - E[b])(a - E[a] + b - E[b])^T] =$$
$$E[(a - E[a])(a - E[a])^T] + E[(a - E[a])(b - E[b])^T]$$
$$+E[(b - E[b])(a - E[a])^T] + E[(b - E[b])(b - E[b])^T]$$
$$= \mathrm{cov}(a) + \mathrm{cov}(b)$$

其中cov(a)和cov(b)分別是a和b的協方差矩陣。這是由於兩個獨立隨機變量的和的協方差等於它們各自的協方差之和。因此y的協方差矩陣，由a和b的協方差矩陣之和給出。

3.

要證明 $\sigma^2_{N+1}(x) \le \sigma^2_N(x)$，需要證明

$$\boldsymbol{\phi}(x)^T \boldsymbol{S}_{N+1} \boldsymbol{\phi}(x) \le \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \boldsymbol{\phi}(x)$$

我們使用 Woodbury matrix identity from Appendix C in [Bishop, 2006]

$$\boldsymbol{S}_{N+1} = [\boldsymbol{S}_N^{-1} + \beta \boldsymbol{\phi}(x) \boldsymbol{\phi}(x)^T]^{-1} = \boldsymbol{S}_N - \frac{\boldsymbol{S}_N \boldsymbol{\phi}(x) \boldsymbol{\phi}(x)^T \boldsymbol{S}_N}{1 + \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \boldsymbol{\phi}(x)}$$

將這方程式帶入上面不等式，可以得到

$$\boldsymbol{\phi}(x)^T \frac{\boldsymbol{S}_N \boldsymbol{\phi}(x) \boldsymbol{\phi}(x)^T \boldsymbol{S}_N}{1 + \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \boldsymbol{\phi}(x)} \boldsymbol{\phi}(x) \ge 0$$

只要協方差矩陣$\boldsymbol{S}_N$是半正定，$\gamma = \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \boldsymbol{\phi}(x) \ge 0$，上述不等是會變成

$$\frac{\gamma^2}{1 + \gamma} \ge 0$$

隨著數據集大小的增加，線性回歸函數相關的不確定性會減少。

4.

這個問題提到獨立的噪音被添加到每個維度的輸入變量x_n，因此我們的新模型變成，

$$\mathrm{y}'(x_n, \boldsymbol{\omega}) = \omega_0 + \sum_{d=1}^{D} \omega_d (x_{nd} + \epsilon_{nd})$$

$$= \omega_0 + \sum_{d=1}^{D} \omega_d x_{nd} + \sum_{d=1}^{D} \omega_d \epsilon_{nd} = \mathrm{y}(x_n, \boldsymbol{\omega}) + \sum_{d=1}^{D} \omega_d \epsilon_{nd}$$

其中 the noise ϵ_{nd} is independent across both the n and d indices，所以新的 error function is

$$\mathrm{E}'_D(\boldsymbol{\omega}) = \frac{1}{2} \sum_{n=1}^{N} \{\mathrm{y}'(x_n, \boldsymbol{\omega}) - t_n\}^2$$

$$= \frac{1}{2} \sum_{n=1}^{N} \{\mathrm{y}(x_n, \boldsymbol{\omega}) + \sum_{d=1}^{D} \omega_d \epsilon_{nd} - t_n\}^2$$

$$\frac{1}{2} \sum_{n=1}^{N} \left\{ (\mathrm{y}(x_n, \boldsymbol{\omega}) - t_n)^2 + 2(\mathrm{y}(x_n, \boldsymbol{\omega}) - t_n) \left(\sum_{d=1}^{D} \omega_d \epsilon_{nd} \right) + \left(\sum_{d=1}^{D} \omega_d \epsilon_{nd} \right)^2 \right\}$$

取期望值，並使用期望的線性性質，得到

$$\mathbf{E}[\mathbf{E'}_D(\boldsymbol{\omega})] = \frac{1}{2}\sum_{n=1}^{N}\left\{(\mathbf{y}(x_n,\boldsymbol{\omega}) - t_n)^2 + 2(\mathbf{y}(x_n,\boldsymbol{\omega}) - t_n)\left(\sum_{d=1}^{D}\omega_d E[\epsilon_{nd}]\right)\right.$$

$$\left. + E\left[\left(\sum_{d=1}^{D}\omega_d\epsilon_{nd}\right)^2\right]\right\}$$

$E[\epsilon_{nd}]$ is 0, so the second term disappears. Now we look at the third term

$$\mathbf{E}\left[\left(\sum_{n=1}^{N}\omega_d\epsilon_{nd}\right)^2\right] = E\left[\sum_{d=1}^{D}\sum_{d'=1}^{D}\omega_d\omega_{d'}\epsilon_{nd}\epsilon_{nd'}\right]$$

$$= \sum_{d=1}^{D}\sum_{d'=1}^{D}\omega_d\omega_{d'}\,E[\epsilon_{nd}\epsilon_{nd'}]$$

$$= \sum_{d=1}^{D}\sum_{d'=1}^{D}\omega_d\omega_{d'}\,\delta_{dd'} = \sum_{d=1}^{D}{\omega_d}^2$$

Using these results, we get

$$\mathbf{E}[\mathbf{E'}_D(\boldsymbol{\omega})] = \frac{1}{2}\sum_{n=1}^{N}\left\{(\mathbf{y}(x_n,\boldsymbol{\omega}) - t_n)^2 + \sum_{d=1}^{D}{\omega_d}^2\right\} = \mathbf{E}_D(\boldsymbol{\omega}) + \frac{N}{2}\sum_{d=1}^{D}{\omega_d}^2$$

and we see that we get a L_2 regularization term without the bias parameter ω_0, as desired.