3a) Regression line. Let $a, b \in \mathbb{R}^n$ two. $m_a = \text{avg}(a) = \frac{1^T a}{n}$, $m_b = \text{avg}(b) = \frac{1^T b}{n}$

$s_a = \text{std}(a) = \frac{1}{\sqrt{n}} \|a - m_a 1\|$, $s_b = \text{std}(b) = \frac{1}{\sqrt{n}} \|b - m_b 1\|$

We assume the vectors are not constant ($s_a \neq 0$ and $s_b \neq 0$) and write the correlation coefficient as $\rho = \frac{1}{n} \frac{(a - m_a 1)^T (b - m_b 1)}{s_a s_b}$

We considered the problem of fitting a straight line to the points (a_k, b_k) by minimizing $J = \frac{1}{n} \sum_{k=1}^{n} (c_1 + c_2 a_k - b_k)^2 = \frac{1}{n} \|c_1 1 + c_2 a - b\|^2$

Show that the optimal coefficients are $c_2 = \rho s_b / s_a$ and $c_1 = m_b - m_a c_2$. Show that for those values of c_1 and c_2, we have $J = (1 - \rho^2) s_b^2$.

Let $a_0 = a - m_a 1$ and $b_0 = b - m_b 1$

thus $s_a = \frac{1}{\sqrt{n}} \|a_0\|$ and $s_b = \frac{1}{\sqrt{n}} \|b_0\|$ and $\rho = \frac{1}{n} \cdot \frac{a_0^T b_0}{s_a s_b}$

$$J = \frac{1}{n} \|c_1 + c_2 a - b\|^2 = \frac{1}{n} \|c_1 + c_2 a_0 - b_0\|^2$$

$$\frac{\partial J}{\partial c_1} = \frac{2}{n} 1^T (c_1 1 + c_2 a_0 - b_0) = 2(c_1 + m_a c_2 - m_b) = 0$$

Thus $c_1 = m_b - m_a c_2$ ∎

Then $J(c_2) = \frac{1}{n} \|c_2 a_0 - b_0\|^2 = s_a^2 c_2^2 + s_b^2 - 2 c_2 \rho s_a s_b$

$J'(c_2) = 0$ so $2 s_a^2 c_2 - 2 \rho s_a s_b = 0$

$c_2 = \frac{\rho s_a s_b}{s_a^2 c} \Rightarrow$ thus $c_2 = \rho s_b / s_a$ ∎

Plug: $J(c_2) = \frac{1}{n} \|c_2 a_0 - b_0\|^2 = c_2^2 \frac{\|a_0\|^2}{n} + \frac{\|b_0\|^2}{n} - 2 c_2 \frac{a_0^T b_0}{n}$

$$J(c_2) = s_a^2 c_2^2 + s_b^2 - 2 \rho s_a s_b c_2$$

Thus $= s_a^2 (\rho \frac{s_b}{s_a})^2 + s_b^2 - 2 \rho s_a s_b (\rho \frac{s_b}{s_a})$

$= \rho^2 s_b^2 + s_b^2 - 2 \rho^2 s_b^2$

$= s_b^2 - \rho^2 s_b^2$

Thus $J = (1 - \rho^2) s_b^2$ ∎

3b) orthogonal distance regression

$\forall p \in (a_k, b_k)$, the vertical deviation from the straight line defined by $y = c_1 + c_2 x$ is given by $e_k = |c_1 + c_2 a_k - b_k|$

The orthogonal distance of (a_k, b_k) to the line: $d_k = \frac{|c_1 + c_2 a_k - b_k|}{\sqrt{1 + c_2^2}}$

We can find the straight line that minimizes the sum of the squared orthogonal distance $\sum J = \frac{1}{n} \sum_{k=1}^{n} d_k^2 = \frac{\|c_1 \mathbb{1} + c_2 a - b\|^2}{n(1 + c_2^2)}$

i) Show that the optimal value of c_1 is $c_1 = m_b - m_a c_2$ as for the least squares fit.

Let $r(c_1) = c_1 \mathbb{1} + c_2 a - b$

$$\frac{\partial}{\partial c_1} \|r(c_1)\|^2 = 2 \mathbb{1}^T r(c_1) = 2(n c_1 + c_2 \mathbb{1}^T a - \mathbb{1}^T b) = 0$$

$$c_1 = \frac{\mathbb{1}^T b}{n} - \frac{c_2 \mathbb{1}^T a}{n}$$

$$c_1 = m_b - m_a c_2 \blacksquare$$

ii) $J = \frac{s_a^2 c_2^2 + s_b^2 - 2 \rho s_a s_b c_2}{1 + c_2^2}$. Set $\frac{\partial J}{c_2} = 0$. Then $\rho c_2^2 + \left(\frac{s_a}{s_b} - \frac{s_b}{s_a}\right) c_2 - \rho = 0$

If $\rho = 0$ and $s_a = s_b$, any value of c_2 is optimal. If $\rho = 0$ and $s_a \neq s_b$ the quadratic eq. has a unique solution $c_2 = 0$. If $\rho \neq 0$, the quadratic eq. has 2 positive and a negative root. Show that the solution that minimizes J is the root c_2 with the same sign as ρ.

$$\rho c_2^2 + \left(\frac{s_a}{s_b} - \frac{s_b}{s_a}\right) c_2 - \rho = 0$$

If $\rho \neq 0$, the quadratic eq. has a positive and a negative root

Let r_1 and r_2 be roots. Thus $r_1 r_2 < 0$.

$$J'(c_2) = \frac{(2 s_a^2 c_2 - 2 \rho s_a s_b)(1 + c_2^2) - (s_a^2 c_2^2 + s_b^2 - 2 \rho s_a s_b c_2)(2 c_2)}{(1 + c_2^2)^2}$$

$$J'(0) = \frac{-2 \rho s_a s_b}{1} = -2\rho (s_a s_b)$$

If $\rho > 0$ then $J'(0) < 0$: The function is decreasing at 0. So positive c_2 is a minimum and negative c_2 is a max.

If $\rho < 0$ then $J'(0) > 0$: the function is increasing at 0. So negative c_2 is minimum and positive is a maximum.

Thus, $\rho > 0$, a and b increase so the best fit slope should be positive

$\rho < 0$, a and b decrease so the best fit slope should be negative

Thus the solution that minimizes J is the root c_2 with the same sign as ρ $\blacksquare$