Suppose we have a set of lines with points (x_i, y_i). Then the best-line fit is

$$\begin{aligned}
r &= \frac{N\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}} \\
b &= r\frac{\sigma_y}{\sigma_x} \\
a &= \langle y\rangle - b\langle x\rangle
\end{aligned}$$

If we write $SX = \sum_i x_i$, $SY = \sum_i y_i$, $SXY = \sum_i x_i y_i$, $SXX = \sum_i x_i^2$, and $SYY = \sum_i y_i^2$, then

$$r = \frac{N \times SXY - SX \times SY}{\sqrt{(N \times SXX - SX^2)(N \times SYY - SY^2)}}$$

Also $\sigma_x = \frac{1}{N}\sqrt{N \times SXX - SX^2}$, ditto for σ_y, and $\langle x\rangle = SX/N$. Then we can rewrite

$$r = \frac{NS_{XY} - S_X S_Y}{N^2\sigma_x\sigma_y}$$

$$b = \frac{NS_{XY} - S_X S_Y}{N^2\sigma_x^2} = \frac{NS_{XY} - S_X S_Y}{NS_{XX} - S_X^2}$$

$$\begin{aligned}
a &= \frac{S_Y}{N} - \frac{NS_{XY} - S_X S_Y}{NS_{XX} - S_X^2}\frac{S_X}{N} \\
&= \frac{1}{N}\left[\frac{(NS_Y S_{XX} - S_Y S_X^2) - (NS_{XY}S_X - S_X^2 S_Y)}{NS_{XX} - S_X^2}\right] \\
&= \frac{S_Y S_{XX} - S_{XY}S_X}{NS_{XX} - S_X^2}
\end{aligned}$$

However, there is a problem if all the x's or y's are equal, because then the denominator of r is zero. Bleh.

Let's parametrize the line instead using $p(x_0, y_0) + (1-p)(x_n, y_n)$. Then the variance from this line is

$$V = \sum_i (x_i - (px_0 + (1-p)x_n))^2 + \sum_i (y_i - (px_0 + (1-p)x_n))^2$$

Let's write $f_{x,i}(p) = (x_i - x_n) + p(x_n - x_0)$. Then

$$V = \sum_i f_{x,i}(p)^2 + f_{y,i}(p)^2$$

We want

$$\begin{aligned}
0 &= \frac{\partial V}{\partial p} = \sum_i \frac{\partial f_{x,i}^2}{\partial p} + \frac{\partial f_{y,i}^2}{\partial p} \\
&= 2\sum_i (f_{x,i}f'_{x,i} + f_{y,i}f'_{y,i})
\end{aligned}$$

Now $f'_x(p) = (x_n - x_0)$. Thus $f_x(p)f'_x(p) = (x_n - x_0)(x_i - x_n + p(x_n - x_0)) = x_i(x_n - x_0) - x_n(x_n - x_0) + p(x_n - x_0)^2$.

$$0 = (x_n - x_0)\sum_i x_i - N(-x_n(x_n - x_0) + p(x_n - x_0)^2) + (y_n - y_0)\sum_i y_i - N(-y_n(y_n - y_0) + p(y_n - y_0)^2)$$

$$
\begin{aligned}
&= p[-(x_n - x_0)^2 - (y_n - y_0)^2] + (x_n - x_0)(\langle x \rangle - x_n) + (y_n - y_0)(\langle y \rangle - y_n) \\
\implies p &= \frac{(x_n - x_0)(\langle x \rangle - x_n) + (y_n - y_0)(\langle y \rangle - y_n)}{(x_n - x_0)^2 + (y_n - y_0)^2}
\end{aligned}
$$

No, let's try something different. Suppose we parametrize $x_i = a_x t_i + b_x$ and $y_i = a_y t_i + b_y$. Then if we write $T_1 = \sum_i t_i$ and $T_2 = \sum_i t_i^2$,

$$
\begin{aligned}
S_X &= a_x T_1 + b_x N \\
S_Y &= a_y T_1 + b_y N \\
S_{XX} &= a_x^2 T_2 + 2a_x b_x T_1 + b_x^2 N \\
S_{YY} &= a_y^2 T_2 + 2a_y b_y T_1 + b_y^2 N \\
S_{XY} &= a_x a_y T_2 + (a_x b_y + a_y b_x) T_1 + b_x b_y N \\
N S_{XX} - S_X^2 &= N(a_x^2 T_2 + 2a_x b_x T_1 + b_x^2 N) - (a_x^2 T_1^2 + 2a_x b_x N T_1 + b_x^2 N^2) \\
&= N a_x^2 (T_2 - T_1^2)
\end{aligned}
$$

Then the Pearson coefficient becomes

$$
\begin{aligned}
r &= \frac{N \times SXY - SX \times SY}{\sqrt{(N \times SXX - SX^2)(N \times SYY - SY^2)}} \\
&= \frac{N(a_x a_y T_2 + (a_x b_y + a_y b_x) T_1 + b_x b_y N) - (a_x T_1 + b_x N)(a_y T_1 + b_y N)}{\sqrt{N^2 a_x^2 a_y^2 (T_2 - T_1^2)}} \\
&= \frac{(N a_x a_y T_2 + N(a_x b_y + a_y b_x) T_1 + b_x b_y N^2) - (a_x a_y T_1^2 + (a_x b_y + a_y b_x) T_1 N + b_x b_y N^2)}{N a_x a_y \sqrt{T_2 - T_1^2}} \\
&= \frac{a_x a_y (N T_2 - T_1^2)}{N a_x a_y \sqrt{T_2 - T_1^2}} \\
&= \frac{N T_2 - T_1^2}{N \sqrt{T_2 - T_1^2}}
\end{aligned}
$$

and this is absolute codswollop because the a's disappeared entirely, and it depends entirely on the parametrization. Bah.