Break-Even Correlation Thresholds

Yannick Kälber • 19 Feb 2026

As we all know, backtesting is not a research tool, but the very end of your research pipeline. If you want to evaluate if a given signal x is predictive for returns r, you can do this more clearly and directly by regressing r on x or measuring their correlation. But "*how strong*" does that correlation need to be for the signal to be "*good enough*"? A popular heuristic by Macrocephalopod provides a practical way of thinking about this question.

This note builds on that idea, replacing all approximations with a more generalized and more detailed derivation.

Linear Model

We model the return r as a linear function of the signal value x:

$$r = \alpha + \beta x + \epsilon \tag{1}$$
$$r = \alpha + \beta x + \sigma_\varepsilon \varepsilon \tag{2}$$

where α is the intercept (the unconditional expected return), β is the slope coefficient, and the residual ϵ is a mean-zero random variable with variance σ_ε^2, which we can refactor into a scaled unit-variance residual ε with σ_ε denoting the residual standard deviation. We assume $\mathrm{Cov}(x, \varepsilon) = 0$, which we use throughout.

The correlation ρ between r and x follows directly from the linear model. To establish the exact notation used in the subsequent evaluation criteria, we briefly restate the derivation from the standard definition of the correlation coefficient:

$$\rho = \frac{\operatorname{Cov}(r,\, x)}{\sqrt{\operatorname{Var}(r)\operatorname{Var}(x)}} \tag{3}$$

$$= \frac{\operatorname{Cov}(\alpha + \beta x + \sigma_\varepsilon\,\varepsilon,\; x)}{\sqrt{\operatorname{Var}(\alpha + \beta x + \sigma_\varepsilon\,\varepsilon)\operatorname{Var}(x)}} \tag{4}$$

$$= \frac{\operatorname{Cov}(\alpha,\, x) + \operatorname{Cov}(\beta x,\, x) + \operatorname{Cov}(\sigma_\varepsilon\,\varepsilon,\, x)}{\sqrt{[\operatorname{Var}(\alpha) + \operatorname{Var}(\beta x) + \operatorname{Var}(\sigma_\varepsilon\,\varepsilon) + 2\operatorname{Cov}(\alpha,\, \beta x) + 2\operatorname{Cov}(\alpha,\, \sigma_\varepsilon\,\varepsilon) + 2\operatorname{Co}}} \tag{5}$$

$$= \frac{0 + \beta\operatorname{Var}(x) + \sigma_\varepsilon\operatorname{Cov}(\varepsilon,\, x)}{\sqrt{[0 + \beta^2\operatorname{Var}(x) + \sigma_\varepsilon^2\operatorname{Var}(\varepsilon) + 0 + 0 + 2\,\beta\,\sigma_\varepsilon\operatorname{Cov}(x,\, \varepsilon)]\operatorname{Var}(x)}} \tag{6}$$

$$= \frac{\beta\operatorname{Var}(x)}{\sqrt{\left[\beta^2\operatorname{Var}(x) + \sigma_\varepsilon^2\right]\operatorname{Var}(x)}} \tag{7}$$

$$= \frac{\beta\,\sigma_x}{\sigma_r} \tag{8}$$

In (3) we write the standard definition of the correlation coefficient. In (4) we substitute the linear model (2) for r, after which (5) applies the bilinearity of covariance in the numerator and the full variance expansion in the denominator. In (6) we evaluate each term: $\operatorname{Cov}(\alpha,\, x) = 0$ and $\operatorname{Var}(\alpha) = 0$ because α is a constant; $\operatorname{Cov}(\beta x,\, x) = \beta\operatorname{Var}(x)$ and $\operatorname{Var}(\beta x) = \beta^2\operatorname{Var}(x)$ because scalars factor out (and as their square for variance); $\operatorname{Var}(\sigma_\varepsilon\,\varepsilon) = \sigma_\varepsilon^2\operatorname{Var}(\varepsilon)$ by the same rule; all covariance terms involving the constant α vanish; in the remaining terms the scalars β and σ_ε factor out. In (7) we use $\operatorname{Cov}(\varepsilon,\, x) = 0$ by assumption and $\operatorname{Var}(\varepsilon) = 1$ by construction, which collapses both numerator and denominator. In (8) we cancel one factor of $\sqrt{\operatorname{Var}(x)} = \sigma_x$ between numerator and denominator and recognise the remaining $\sqrt{\beta^2\operatorname{Var}(x) + \sigma_\varepsilon^2}$ as σ_r, since $\operatorname{Var}(r) = \beta^2\,\sigma_x^2 + \sigma_\varepsilon^2$ follows directly from (2) and $\operatorname{Cov}(x, \varepsilon) = 0$.

For stating a signal evaluation criterion in the next step, we need β expressed in terms of ρ, which we read off directly from (8) by multiplying both sides by σ_r/σ_x:

$$\beta = \frac{\rho\,\sigma_r}{\sigma_x} \tag{9}$$

This is the standard identity linking the regression slope to the correlation coefficient.

Signal Evaluation Criterion

Finally, we state what it means for a signal to be "*good enough*". We require that, at a signal level k standard deviations from its mean, i.e. at $x = \mu_x + k\sigma_x$, the corresponding absolute expected return exceeds a trading cost threshold $c > 0$:

$$\left|\mathbb{E}[r \mid x = \mu_x + k\sigma_x]\right| > c \tag{10}$$

$$\left|\alpha + \beta\big(\mu_x + k\sigma_x\big)\right| > c \tag{11}$$

$$\left|\alpha + \frac{\rho\,\sigma_r}{\sigma_x}\big(\mu_x + k\sigma_x\big)\right| > c \tag{12}$$

In (10) we state the criterion in general terms: the conditional expected return, evaluated at a signal realization k standard deviations from its mean $\mu_x = \mathbb{E}[x]$, must exceed the threshold c in absolute value. In (11) we substitute $\mathbb{E}[r \mid x] = \alpha + \beta x$ from (2), and in (12) we replace β using (9), which expresses the criterion entirely in terms of ρ. The absolute value reflects that the signal can be profitable in either direction (long or short).

The parameter k controls how strict the criterion is and has a direct probabilistic interpretation. Since $\mathbb{E}[r \mid x] = \alpha + \beta x$ is linear in x, all realizations of x closer to the mean μ_x, i.e. where $|x - \mu_x| \leq k\sigma_x$, fail as well, if $|\mathbb{E}[r \mid x = \mu_x + k\sigma_x]|$ fails. By Chebyshev's inequality,

$$P\big(|x - \mu_x| \leq k\sigma_x\big) \geq 1 - \frac{1}{k^2} \tag{13}$$

So at least a fraction $1 - 1/k^2$ of all signal realizations fall within this range. If ρ fails to clear c at the $k\sigma_x$ boundary, the signal may be economically non-viable for the majority of realizations and should be discarded. A smaller k raises the bar on ρ because a lower fraction of unprofitable realizations is accepted, whereas a larger k lowers the bar because a higher fraction of unprofitable realizations is accepted.

The absolute value in (12) splits into two cases, depending on whether the expression inside is strictly positive or strictly negative:

$$\text{(A)} \quad \alpha + \frac{\rho\,\sigma_r}{\sigma_x}\big(\mu_x + k\sigma_x\big) > c \tag{14}$$

$$\text{(B)} \quad \alpha + \frac{\rho\,\sigma_r}{\sigma_x}\big(\mu_x + k\sigma_x\big) < -c \tag{15}$$

Case (A) corresponds to the signal pushing expected returns above the positive threshold $+c$ (profitable for a long position), while Case (B) corresponds to pushing expected returns below $-c$ (profitable for a short position).

Case (A): Long Profitability

We rearrange (14) by moving α to the right-hand side and dividing by $\sigma_r(\mu_x + k\sigma_x)/\sigma_x$:

$$\frac{\rho\,\sigma_r}{\sigma_x}\big(\mu_x + k\sigma_x\big) > c - \alpha \tag{16}$$

$$\rho\,\sigma_r\big(\mu_x + k\sigma_x\big) > (c - \alpha)\,\sigma_x \tag{17}$$

Dividing both sides of (17) by $\sigma_r(\mu_x + k\sigma_x)$, which is nonzero by assumption, yields two subcases depending on its sign:

$$\mu_x + k\sigma_x > 0\colon \quad \rho > \frac{(c - \alpha)\,\sigma_x}{\sigma_r\big(\mu_x + k\sigma_x\big)} \tag{18}$$

$$\mu_x + k\sigma_x < 0\colon \quad \rho < \frac{(c - \alpha)\,\sigma_x}{\sigma_r\big(\mu_x + k\sigma_x\big)} \tag{19}$$

In (18) the evaluation point is positive, so dividing preserves the inequality direction, and ρ must exceed the threshold on the right. In (19) the evaluation point is negative, so dividing reverses the direction, and ρ must fall below the threshold. Notably, if $\alpha > c$, profitability does not require a strictly positive correlation in (18) or a strictly negative correlation in (19) since the unconditional return already exceeds the cost threshold c.

Case (B): Short Profitability

We rearrange (15) analogously:

$$\rho\,\sigma_r\big(\mu_x + k\sigma_x\big) < -(c + \alpha)\,\sigma_x \tag{20}$$

Dividing both sides of (20) by $\sigma_r(\mu_x + k\sigma_x)$:

$$\mu_x + k\sigma_x > 0\colon \quad \rho < \frac{-(c + \alpha)\,\sigma_x}{\sigma_r\big(\mu_x + k\sigma_x\big)} \tag{21}$$

$$\mu_x + k\sigma_x < 0\colon \quad \rho > \frac{-(c + \alpha)\,\sigma_x}{\sigma_r\big(\mu_x + k\sigma_x\big)} \tag{22}$$

In (21) the evaluation point is positive, so dividing preserves the direction and ρ must fall below the threshold. In (22) the evaluation point is negative, so dividing reverses it and ρ must exceed the threshold. Analogously, if $\alpha < -c$, profitability does not require a strictly positive correlation in (21) or a strictly negative correlation in (22) since the unconditional return already lies below the cost threshold.

Application

Which bound of $(18)/(19)$ and $(21)/(22)$ applies is fully determined by the input parameters. Given a concrete signal with intercept α, correlation ρ, return volatility σ_r, signal mean μ_x, signal volatility σ_x, a cost threshold $c > 0$, and an evaluation level k, the procedure is as follows:

First, determine whether you are checking for long profitability (Case (A)) or short profitability (Case (B)), keeping in mind that both can be checked independently and a signal may satisfy one, both, or neither. Second, check the sign of the evaluation point $\mu_x + k\sigma_x$. If it is positive, use (18) or (21); if negative, use (19) or (22). Third, choose $|k|$ according to how selective you wish to be.