

overfit_meanrev.py

Strategy stress test summary. The full report has every finding, test and figure.

Assets	SPY	Data period	1993-01-29 to 2026-10-02
Development period	2000-01-01 to 2012-12-31	Costs and fills	5 bps per side; filled at the next close
Generated	2026-10-09 01:05 UTC	Engine	Quantgrounds 0.2.0
Code fingerprint	067e 4b9a 8316 ff63	Replay code	1/067e4b9a8316ff63/C/SPY/2000-01-01/2012-12-31/5/10

Verdict

The edge largely disappears outside the period the strategy was developed on.

Diagnostic score

10 / 100

Serious issues

Weighted mean of six sub-scores. Capped at 10 by the edge over buy & hold (-0.87 Sharpe). Not a forecast.

Sub-score	Weight	Score	Based on
Edge over buy & hold	25%	0	out-of-sample Sharpe 0.18 vs 1.05 held
Risk-adjusted return	15%	18	out-of-sample net Sharpe 0.18
Parameter robustness	15%	52	99% of Sharpe kept (median), 4% at worst $\pm 10\%$
Cost resilience	15%	82	break-even at 5.6x your cost
Generalization	15%	70	100% of held-out assets positive, Sharpe ratio 0.41
Drawdown at equal risk	15%	21	68.0% at equal volatility vs 55.2% held

Stress tests

Stress test	Result	Status
Out-of-sample periods	Sharpe 0.95 in-sample, 0.18 out-of-sample	Fail
Walk-forward	Sharpe 0.38 over 28 one-year tests; 18 of 28 positive	Pass
Parameter sensitivity	Worst $\pm 10\%$ change keeps 4% of Sharpe (rsi_days)	Fail
Transaction costs	Break-even at 27.9 bps per side (5.6x baseline)	Pass
Execution delay	Sharpe 0.46 to 0.34 with a one-day delay	Pass
Unseen assets	Positive on 10 of 10 held-out assets; median Sharpe 0.19	Warning
Market regimes	Weakest: bear / crash, Sharpe -0.23 over 577 days	Pass
Look-ahead bias	Prevented by design: on_bar() sees past data only	Pass
Resampled histories	Sharpe 90% range 0.18 to 0.71	Pass
Random timing	Beats 14% of 1,000 random timings with the same exposure (after the development period)	Warning
Market exposure	Out-of-sample Sharpe 0.18 vs 1.05 held; full history: beta 0.07, alpha 1.4%/yr (t 1.9)	Fail
Research selection	Variants tried not given; selection risk unknown	Not run
Code review	1 note; deterministic	Warning
Drawdowns and concentration	Max drawdown -23.3%; longest underwater 10.2 yrs	Warning

Key metrics

Metric	Strategy (net of costs)	Buy & hold SPY
CAGR	2.2%	10.8%
Sharpe ratio (90% range)	0.46 (0.18 to 0.71)	0.65
Maximum drawdown	-23.3%	-55.2%
Total trades / yearly turnover	156 / 9.3x	1 / 0.0x

Most important findings

F1 **High** Performance collapses outside the development period · line 12

Evidence. In-sample (2000-01-01 to 2012-12-31): Sharpe 0.95, CAGR 4.4%. Out-of-sample: Sharpe 0.18 (before the development period: Sharpe 0.55, CAGR 2.6%; after the development period: Sharpe -0.00, CAGR -0.1%).

Likely cause. Performance is concentrated in the period you developed on, which is the signature of fitting to that period. See the tuned-looking values at line 12.

Change to try. Re-derive the rules with fewer, rounder parameters. Then judge them only on data you did not look at, such as a walk-forward split, before trusting the in-sample numbers.

F2 **High** Parameter cliff: 'rsi_days' · line 11

Evidence. Baseline rsi_days=4 gives Sharpe 0.46. A $\pm 10\%$ nudge to rsi_days=5 gives Sharpe 0.02 (4% retained). All nudges: rsi_days=3: Sharpe 0.49; rsi_days=5: Sharpe 0.02.

Likely cause. PARAMS['rsi_days'] (line 11) sits on a narrow peak. Small changes in market behavior act like a parameter nudge.

Change to try. Pick 'rsi_days' from the middle of a plateau where neighbors perform similarly, or average the signal over several values of 'rsi_days' instead of betting on one.

F3 **High** Buy & hold of the same assets did better

Evidence. Out-of-sample Sharpe 0.18 vs 1.05 held (same assets, same days, same costs). The score is capped at 10.

Likely cause. holding the assets earned more per unit of risk than trading them

Change to try. Unless the strategy has another purpose it delivers (lower drawdowns, less time invested), holding is simpler and cheaper.

F4 **Medium** Logic doesn't transfer to assets you didn't pick

Evidence. Median Sharpe on declared assets 0.46 vs 0.19 on 10 held-out assets; 100% of held-out assets were profitable.

Likely cause. The rule may be exploiting something specific to the declared asset and period rather than a general effect.

Change to try. State the mechanism you expect to work and why it should be specific to this asset. If it shouldn't be, simplify until it works across similar assets.

3 more findings in the full report.

What held up

+ Cost resilient: break-even at 27.9 bps per side (5.6x baseline).

+ Walk-forward holds up: Sharpe 0.38 when parameters are re-chosen from past data only (18 of 28 test years positive).

Educational diagnostics only; not investment advice. Results are simulated on historical data and do not predict future returns. The score summarizes this report's tests; it is not a measure of trading skill or future profitability.