Persistence, Premium, and the Illusion of Residual Reversion in SPY Options

Charlie Yan*

August 2026

Abstract

Deviations of individual option prices from a fitted implied-volatility surface ("residuals") are widely believed to mean-revert quickly, making them a natural relative-value signal. Using 2.78 million contract-snapshot observations of SPY options (hourly, 5.7 years, SVI fits per snapshot-expiry), we show this belief fails three successively stricter pre-registered tests. (1) At the one-snapshot horizon, observed convergence (the residual halving) runs 7–11 points *below* a scale-matched no-reversion null at every entry threshold, because residuals are *persistent*: AR(1) coefficient $\varphi = 0.928$ hourly, half-life ≈ 1.3 trading days. A persistent series halves less often than independent draws of the same scale; naive reversion statistics at these horizons are artifacts, with AR(1) persistence alone generating a "reversion IC" of -0.19 to -0.31 with zero tradeable content. (2) A capture-versus-cost screen nonetheless crosses positive at multi-day horizons with entry selective on the contract's own spread (+0.050 vol points net at 3× spread, 63% positive; SCREEN)—and its point estimate of the mean is *accurate*: the faithful tape realizes \$2.56 per contract against the screen's \$2.75. (3) The tape nonetheless fails: line 3 Sharpe 0.34 (95% CI $[-0.18, 0.90]$), 98% of P&L concentrated in 2022, negative under a 2× fill-haircut. Leg attribution shows the option leg *losing* money in the profitable year while the delta hedge earns it, and a side-randomized control scores always-short at 1.75 against the signal's 0.34. The apparent anomaly is three familiar objects in one costume: persistence mistaken for reversion, a mean mistaken for an edge, and the variance risk premium mistaken for alpha. We propose a six-item minimum evidence standard for residual-reversion claims and reconcile our findings with the surface-dynamics literature: the 1.3-day half-life sits inside the 1–5-day band of Cont and Da Fonseca (2002). The error was never the physics; it was the clock and the accounting.

1 Introduction

Every parametric fit of an option smile leaves residuals, and nearly every practitioner who has plotted them has had the same thought: they look like they snap back. The thought has a respectable pedigree. Higher-order factors of the implied-volatility surface mean-revert on horizons of days (Cont and Da Fonseca, 2002); smooth-surface residuals are described in the smoothing literature as "transient" imbalances (Fengler, 2009); and demand-based pricing (Gârleanu et al., 2009; Bollen

*Draft v0.3 (journal-length), August 2026. Basis discipline: SCREEN = descriptive or information-coefficient statistics, no P&L. Tape figures = Line 3 cross-spread fills (buy at ask, sell at bid, every leg both ways), ex-commission, marked-to-market daily, full business calendar, 2020-09-01 to 2026-05-12, unless labeled otherwise. All tests were pre-registered with horizons, thresholds, nulls, and decision bars fixed before execution; the registration documents, including recorded deviations, are reproduced in Appendix A. Code, pre-registration documents and a data-availability statement: `https://github.com/charlieyanhx/svi-residual-persistence`.

and Whaley, 2004) supplies the mechanism: end-user demand pushes individual contracts off fair value, market makers charge inventory premia, and prices relax as inventory recycles. From there it is a short step to a strategy: fit a surface, buy the cheap contracts, sell the rich ones, wait for convergence.

This paper documents, on one of the most liquid option markets in the world, why that step fails—and why the standard evidence offered for it is illusory even though the underlying physics is real. Our contribution is not the discovery that transaction costs are large (that is Figlewski's (1989) point), nor that fitted surfaces disappoint out of sample (Dumas et al., 1998). It is a demonstration, on 2.78 million contract-snapshots with every test pre-registered, that the *measurement conventions* common in residual-reversion analysis manufacture the appearance of a fast, harvestable effect out of three separate, individually well-known objects:

1. **Persistence mistaken for reversion.** One-step "convergence rates" of 40–65% sound like rapid mean reversion. Against a null that preserves the residual's conditional scale while destroying its serial linkage, the observed rates are 7–11 points *below* chance: residuals are persistent (hourly AR(1) $\varphi = 0.928$), and a persistent series halves less often than independent same-scale draws. We show analytically that persistence alone manufactures a "reversion IC" of $-\sqrt{(1-\varphi)/2}$—between -0.19 and -0.31 at our horizons—indistinguishable from signal if the autoregressive null is not subtracted.
2. **A mean mistaken for an edge.** A capture-versus-cost screen on the same data is *accurate about the mean*—it projects \$2.75 per contract net and the faithful tape delivers \$2.56—yet the tape's risk-adjusted performance is indistinguishable from zero. The screen has no access to concentration (98% of P&L in one year), to the spread's share of gross (56%), or to the daily volatility that determines the Sharpe ratio.
3. **Premium mistaken for alpha.** A side-randomized negative control on identical entries shows that a constant short position outperforms the signal five-fold (line 3 Sharpe 1.75 versus 0.34). The tape—which passes coverage, reconciliation, sign, and count checks—is harvesting the variance risk premium (Carr and Wu, 2009; Bakshi and Kapadia, 2003) on half its trades and paying it on the other half; the residual signal contributes $\approx$ \$4.9 per trade of genuine discrimination against a $\approx$ \$24 per trade premium it merely dilutes.

Three positive results survive: residual persistence is precisely measured ($\varphi = 0.928$ hourly, half-life 9.3 snapshots $\approx$ 1.3 trading days, consistent with the 1–5-day factor-reversion band of Cont and Da Fonseca, 2002); the horizon structure of option-market friction is documented (the round-trip toll is flat in the holding period while the option-specific move grows with $\sqrt{t}$); and a six-item evidence standard is proposed under which residual-reversion claims can be evaluated without repeating our mistakes. The standard is cheap—each item is a few lines of code, released in the companion library `opentape`—and we believe it would have prevented at least one internal predecessor study, whose specification error we also document, from reaching its published-sounding conclusions.

2 Related literature

Surface dynamics. Cont and Da Fonseca (2002) established the factor description of the IV surface and the mean reversion of its higher-order components at 1–5 day horizons; Daglish et al. (2007) catalogue the empirical regularities of surface movement against practitioner rules of thumb. Our object is one level finer—the residual of a single contract from its own snapshot's fitted slice—and our half-life estimate (1.3 trading days) is consistent with theirs: nothing in our negative results contradicts the established physics. **Trading fitted deviations.** Dumas et al. (1998) showed that deterministic volatility functions fit the smile in sample yet failed to beat a simple benchmark on out-of-sample valuation and hedging—the earliest systematic warning that fitting a surface well confers little usable information about where it goes next; Goncalves and Guidolin (2006) and Bernales and Guidolin (2014) found forecastable surface dynamics whose economic value vanishes at realistic costs. Our contribution relative to this line is the diagnosis: *why* such strategies look attractive before they fail—the null miscalibration of convergence statistics, and the attribution failure that a premium-bearing instrument inflicts on any side-symmetric signal. Neither mechanism is articulated in that literature. **Demand-based pricing.** Gârleanu et al. (2009) and Bollen and Whaley (2004) explain why residuals exist and why they relax; Muravyev (2016) shows order flow predicts option returns at the daily horizon. Our persistence estimate quantifies the relaxation those models describe—and shows it is too slow, relative to the toll, to be harvested at the contract level. **Option returns and the premium.** Delta-hedged option positions earn negative average returns (Bakshi and Kapadia, 2003; Coval and Shumway, 2001); the variance risk premium is large and time-varying (Carr and Wu, 2009). This is the exposure our negative control isolates: any structure short options collects it, any structure long options pays it, and a 50/50 signal nets to its own thin discrimination margin. **Costs and measurement.** Figlewski (1989) anticipated that options arbitrage dies in imperfect markets; Muravyev and Pearson (2021) show effective spreads for execution-timers are far below quoted — at second-to-minute horizons. Our friction measurements use full quoted spreads and are therefore conservative; §6 shows the conclusion survives a 2× haircut in the favorable direction and reverses under a 2× stress. Methodologically we follow the pre-registration discipline argued for by Arnott et al. (2019): every test below was specified— horizons, thresholds, nulls, bars, and both possible write-ups—before execution.

3 Data and surface construction

Panel. SPY option chain, hourly snapshots (seven per trading day), 2020-09-01 to 2026-05-12: 1,429 trading days spanning the COVID-era volatility regime, the 2022 bear market, and the August-2024 and April-2025 stress episodes. Quotes, sizes, vendor Greeks and implied volatilities are as disseminated; no filtering beyond two-sided quotes (bid > 0, ask $>$ bid). **Fits.** SVI (Gatheral and Jacquier, 2014) is fitted per (snapshot, expiration) across both wings jointly, with a fixed four-point multi-start and no random number generation anywhere in the chain; the build is bitwise-reproducible (re-running it reproduces all downstream artifacts with identical checksums). The residual is $r = \sigma_{\text{obs}} - \sigma_{\text{fit}}$ in volatility units at the contract's own strike. **Universe and coverage.** Out-of-the-money contracts, $|\Delta| \in [0.05, 0.40]$, 25–45 days to expiry: 2,777,181 rows $\to$ 2,353,081 in-universe with usable quotes $\to$ 2,302,736 contract-level transitions. All joins are on exact contract identity (expiration, strike, type); forward lookups never re-bucket. Median quoted spread in the universe, expressed in volatility points through each contract's Black–Scholes vega: 0.070; median $|r|$: 0.075; 49.6% of contract-snapshots exceed their own round-trip spread—the trade is not excluded by first-pass arithmetic, which is precisely why the tests below are needed. **Whose claim is being tested.** The belief under examination is not a straw man from our own files. It is the working assumption behind residual-based relative-value trading in options, and it has explicit form in the literature: Fengler (2009) characterises residuals from smooth surface fits as transient deviations around a fitted shape, and the demand-pressure account of Gârleanu et al. (2009) implies that inventory-driven dislocations revert as inventory clears. Neither quantifies the reversion against a null, which is the gap this paper fills. We use an internal predecessor study as a concrete stalking horse for the general belief—it stated the claim in testable form—rather than as the paper's target.

A predecessor's specification error. An internal study on this data reported 58–65% of "large" deviations converging by half within one snapshot, defining large as $|r| > 0.003$ in total variance. Its own reported mean $|\varepsilon|$ was 0.037–0.049: the threshold sat *below* the typical residual, so the statistic described the entire sample, not a tail of dislocations. The scale discrepancy traces to computing total variance with maturity in days where the production chain uses years ($\sim 365\times$). We report this because any reader comparing our numbers to similar internal or published claims should first check the threshold-to-typical-residual ratio; it is the cheapest diagnostic in this paper.

4 Test 1: convergence against its own null

Design (pre-registered). Entry set: $|r_t|$ above its own 50th/75th/90th percentile (scale-free; the verdict must hold at all three). Convergence: $|r_{t+1}| < 0.5\,|r_t|$, same contract, next snapshot. Transitions are split by clock: intraday (gap $\leq$ 90 minutes) and overnight, never pooled. Null: 200 permutation draws (the null distribution's standard error is < 0.001 at this count, an order of magnitude below the effect sizes at issue) permuting r_{t+1} within (type $\times$ delta-decile $\times$ DTE-bucket) cells—preserving the conditional scale of the next residual, destroying its serial linkage to r_t. Bar: observed above the null's 97.5th percentile.

Result: inverted, everywhere (Figure 1, Table 1). Observed convergence is not merely short of its null; it sits 7–11 points below it at every threshold and both clocks.

Arm	n	Observed	Null mean [95%]	Obs − null
intraday, p75	486,164	0.048	0.156 [0.155, 0.157]	**−0.108**
intraday, p90	192,412	0.052	0.124 [0.123, 0.124]	−0.071
overnight, p75	89,520	0.125	0.225 [0.224, 0.226]	−0.100
overnight, p90	37,862	0.142	0.216 [0.214, 0.218]	−0.074

Table 1: One-step convergence versus the scale-matched null. p50 arms (not shown) agree. Differences are computed on unrounded values and may therefore differ by 0.001 from the difference of the rounded columns shown.

Mechanism. The pooled AR(1) coefficient of the residual is $\varphi = +0.928$ intraday ($n = 1.97$M transitions; half-life $\ln 0.5 / \ln \varphi = 9.3$ snapshots ≈ 1.3 trading days) and $+0.808$ overnight (half-life 3.3 steps). Under an AR(1) with innovation scale matched to the data, the probability that $|r_{t+1}| < 0.5|r_t|$ for an entry drawn from the upper tail of $|r_t|$ is small precisely because $r_{t+1} \approx \varphi r_t$; the shuffled null, which draws r_{t+1} from the unconditional same-cell distribution, halves far more often. The "convergence" convention thus reverses its meaning at high persistence.

The reversion-IC trap. The same persistence contaminates correlation-based evidence. For an AR(1), the population correlation between the current level r_t and the forward change $r_{t+h} - r_t$ is $-\sqrt{(1-\varphi^h)/2}$. At our measured φ this is -0.19 (one intraday step) to -0.31 (one overnight step): a large, significant, and completely non-tradeable "reversion IC" produced by persistence alone. Any level-on-forward-change correlation must be reported as an *excess* over this null; reporting the raw number, as is common, is reporting the autocorrelation function in costume.

Reconciliation. A 1.3-day half-life sits inside the 1–5-day band that Cont and Da Fonseca (2002) report for higher-order surface factors. The failure documented here is not of the physics but of the clock: the reversion is real and slow, while the claims under test asserted it was real and fast.

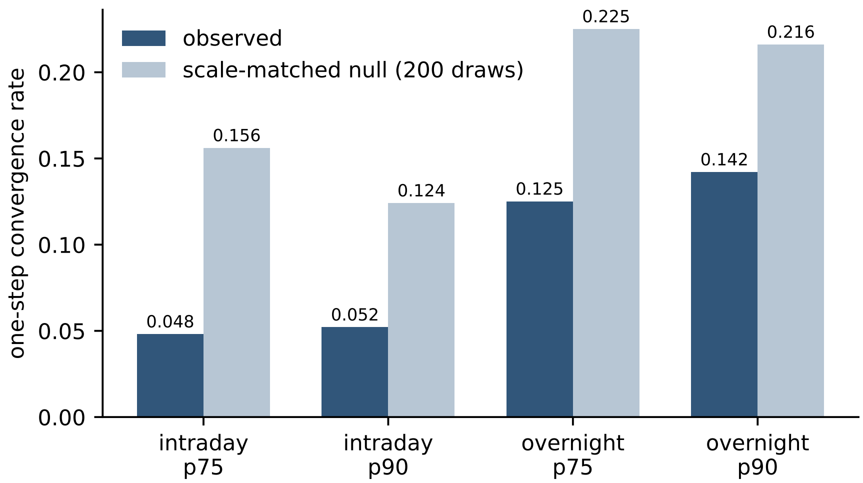

Figure 1: Observed one-step convergence versus the scale-matched null, four pre-registered arms. A persistent series halves less often than independent draws of the same scale.

5 Test 2: the capture screen, and exactly what it gets right

If reversion completes in ∼1.3 days, the natural repair is to hold longer and select harder. A capture-versus-cost screen (SCREEN basis; signed capture, so an overshoot through zero counts as capture, not as failure) delivers exactly the encouragement such repairs usually deliver:

Entry	Hold (snaps)	n	Capture (vp)	Net of spread (vp)	% net > 0
$> 1\times$ spread	21	693,875	0.024	−0.034	33.3
$> 1\times$ spread	63	161,337	0.049	−0.005	48.4
$> 3\times$ spread	63	36,707	0.093	+0.050	63.1
$> 5\times$ spread	63	11,440	0.108	+0.074	67.5
$> 8\times$ spread	63	2,752	0.134	+0.104	73.6

Table 2: Signed capture vs the contract's own round-trip spread. Monotone in selectivity and horizon; the 3×/63-snapshot cell projects ≈ \$2.75/contract net. Entries overlap heavily: at $h = 63$, effective $n \approx n/63$.

A screen of this shape is why residual strategies get funded. We emphasize what it gets *right*, because the standard critique—"screens overstate"—is not what happens here: the faithful tape below realizes \$2.56 per contract at the cross against the screen's \$2.75, and \$5.89 at mid against the screen's ≈ \$5 gross. The screen's point estimate of the mean is accurate to within 7%. What it cannot see is everything else that determines whether the mean is investable: concentration across years, the friction share of gross, and the daily variance of a stacked, overlapping book. Screens are accurate about means and silent about risk; the dangerous ones are those validated on average P&L per trade.

6 Test 3: the faithful tape

Machinery (pre-registered). Single-leg fade of the residual: sell rich contracts, buy cheap ones, at $K\times$ the contract's own spread ($K = 3$ primary; 2 and 5 declared sensitivities); entry at the *next* snapshot after the signal; hold nine trading days located on the calendar; force-close at the last real quote if the contract leaves the panel or crosses DTE < 7 (candidates are never dropped); no re-entry while a contract is open; daily SPY delta hedge at \$0.005 per share. Marking is MTM-daily via a shared marking library with a full-coverage assertion (one mark row per trade per held trading day); quote-less held days are re-marked by carry plus intrinsic shift. Accounting is three-line; the decision line is full cross (line 3), buy at ask and sell at bid, every leg, both directions; commissions (\$0.65/contract/side) are a separate line. Bar: line 3 full-calendar Sharpe ≥ 1.5 with a block-bootstrap 95% CI (Politis and Romano, 1994) excluding zero (2,000 draws, 21-trading-day blocks; the block length is set to the maximum position holding period, so that a block spans a complete trade cycle and the resampling preserves the overlap structure that drives the daily series). Trade counts tie the entry log; line 3 $\leq$ line 1 is asserted; both write-ups were committed before the run.

Entry ($\times$ spread)	Trades	Line 1 (mid)	**line 3 (cross)**	95% CI	Net of comm.
$2\times$	41,928	1.07	0.55	$[-0.03, 1.23]$	0.36
3$\times$	28,152	0.77	**0.34**	$[\mathbf{-0.18}, \mathbf{0.90}]$	0.17
$5\times$	13,072	0.47	0.11	$[-0.45, 0.61]$	-0.06
$3\times$, $2\times$ haircut	28,152	—	-0.10	$[-0.71, 0.41]$	-0.27

Table 3: The tape. Annualized Sharpe, full calendar. The pre-registered bar (1.5, CI excluding zero) fails at every K; the friction stress reverses the sign.

Where the P&L lives. Per-year line 3 P&L at $K = 3$: 2020 $-$\$110; 2021 $-$\$470; **2022** +\$70,394; 2023 +\$16,816; 2024 $-$\$25,486; 2025 +\$422; 2026 (partial) +\$10,496. One year contributes 98% of the total; excluding it, the strategy loses money—$-$\$8,828—over the five *full* years that remain (2026 is a partial year; including its stub the remainder is a statistically meaningless +\$1,668). The spread consumes 56% of gross (line 1 \$165,707 $\rightarrow$ line 3 \$72,062); commissions consume half of what survives (\$72,062 $\rightarrow$ \$35,464). No pooled per-trade statistic—and Table 2 is built from 36,707 of them—can exhibit this structure.

7 Attribution: what the tape was actually measuring

Leg decomposition. Splitting line 3 P&L into option leg and delta hedge, per year: in 2022 the option leg *lost* \$129,483 while the hedge earned \$199,772. The strategy's best year was not a residual-convergence harvest; it was a hedge-side gain in a trending selloff. **Side split.** Short trades earn in every year of the sample (+\$6 to +\$45 per trade); long trades lose in every year ($-$\$15 to $-$\$39). That is the delta-hedged option return premium (Bakshi and Kapadia, 2003; Coval and Shumway, 2001; Carr and Wu, 2009), present on both sides of a side-symmetric signal. **The negative control** (Figure 2). Re-running the identical tape—same entries, same contracts, same holds, same accounting—with the side determined by (i) the signal, (ii) a coin flip, (iii) always short, (iv) always long: signal 0.34; random -1.62 $[-2.69, -1.09]$; **always-short 1.75** $[0.69, 3.32]$; always-long -2.28 $[-4.02, -1.09]$. A constant side beats the signal five-fold. The signal's genuine content—its margin over its own side-mix—is $\approx$ \$4.9 per trade against the $\approx$ \$24 per trade premium it dilutes. **Regime, not skill.** Within the top realized-volatility quintile of entry days, 2022 entries earned +\$7.7 per trade while other high-volatility periods earned −\$27.0: the profitable year was a slow trending repricing (which pays the accidental long-gamma half of a hedged book), not a generic high-volatility harvest. Sharp, mean-reverting stress episodes (August 2024, April 2025) lose money on the same construction.

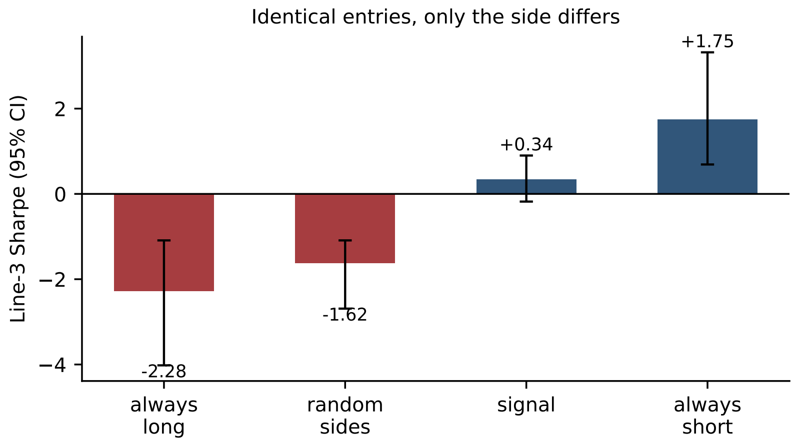

Figure 2: The side-randomized control. Dollars, timing, coverage and fills are identical across bars; only the side assignment differs. If a constant side beats the signal, the tape is measuring the premium, not the signal.

8 What survives, and a proposed evidence standard

Three facts survive every test in this paper. Smile residuals on a liquid index ETF are *persistent* ($\varphi \approx 0.93$ hourly). They revert on a ≈1.3-trading-day half-life, consistent with the factor-level literature. And they typically move less than the cost of a round trip at the horizons where a residual trade needs them to move more—while at horizons long enough for the move to grow, the position is dominated by the premium and gamma exposures of the instrument that carries it.

For residual-reversion claims we propose the following minimum standard, each item a few lines of code (companion library `opentape`): (i) a **scale-matched convergence null**—any convergence rate quoted without one is uninterpretable at high persistence; (ii) **AR(1) subtraction** from any level-on-forward-change correlation, with $-\sqrt{(1-\varphi^h)/2}$ as the reference; (iii) **three fill lines** with the decision taken on full cross; (iv) **MTM-daily, full-calendar marking** (exit-day marking inflated Sharpe 2.2–2.4× in related internal audits); (v) **per-year concentration disclosure**; (vi) **leg attribution with a constant-side negative control**—if always-short beats the signal, there is no signal. Items (i)–(ii) kill the illusion of speed; (iii)–(v) kill the mean-for-edge substitution; (vi) kills the premium-for-alpha substitution. None requires proprietary data or unusual computation, and item (vi) in particular would, we suspect, retire a nontrivial fraction of backtested option "signals" at the cost of one extra run.

9 On the friction basis

Our go/no-go line charges the *full quoted spread*. That is a floor, and deliberately a conservative one: Muravyev and Pearson (2021) show that traders who time marketable executions pay effective spreads below 40% of conventional measures, and Cao and Han (2013) assume 25% of quoted in the single-name cross-section. A referee is entitled to ask whether the negative result is an artifact of an unrealistically harsh assumption. It is not, and the arithmetic is short: the tape runs 0.77 at mid and 0.34 at full cross, so any effective-spread fraction ϕ places it at $0.77 - \phi(0.77 - 0.34)$ — **0.62 at** $\phi = 0.35$**, 0.66 at** $\phi = 0.25$ — against a deployability bar of 1.5. The conclusion is invariant across the whole plausible range. More importantly, the two findings that carry the paper are untouched by ϕ altogether: the concentration of 98% of P&L in a single year, and the side-randomized control in which a constant short position outscores the signal five-fold. Friction sets the level; neither of those is a level.

10 Limitations

One underlying (SPY); one fit family (SVI per snapshot-expiry, so residuals are fit-relative: a different parameterization relabels individual residuals, though the persistence and toll structure are properties of the quotes, not the fit); hourly snapshots (intra-hour reversion is unobservable; the companion execution study finds the intra-hour clock is dominated by theta in any case); 5.7 years containing one dominant trending-volatility year, which is exactly why the per-year table is load-bearing. The tape tests one expression—single-leg fade, nine-day hold—though the constant-side control's verdict applies to any side-symmetric expression of this signal. Cross-sectional constructions (residual *ranking* within snapshot, traded market-neutral across contracts) are a different object and are not closed by this paper.

A Pre-registration summary and deviations

Both gates were registered before execution: (A1) the convergence-null design—universe, thresholds, clock split, cell-shuffle null, 97.5th-percentile bar, and both write-ups (confirmation and non-confirmation) committed in advance; (A2) the tape—entry rule, next-snap fills, nine-day calendar holds, force-close, hedge cadence and cost, marking library, the 1.5 bar, and the three committed outcomes. Recorded deviations, in full: (D1) the predecessor's absolute 0.003 threshold was replaced by scale-free percentiles after the scale error of §3 was discovered—the bar itself was unchanged; (D2) the panel's DTE span (25–45) made a registered 7–60-day secondary arm unrunnable; it was dropped, not approximated; (D3) no re-entry while a contract is open (the registered "one trade per signal" would multiply-count persistent dislocations; the change is conservative). The registration documents are available with the replication materials.

B Additional tables

Available with the replication materials: p50 null arms; per-year leg decomposition for all years; the side split by year; the volatility-quintile interaction; hold-distribution statistics (median realized hold 5 calendar days; force-close binds on DTE and panel exit); and the bitwise-reproducibility manifest for the surface chain.

References

Arnott, R., Harvey, C., Markowitz, H. (2019). A backtesting protocol in the era of machine learning. *Journal of Financial Data Science* 1(1).

Bakshi, G., Kapadia, N. (2003). Delta-hedged gains and the negative market volatility risk premium. *Review of Financial Studies* 16(2).

Bernales, A., Guidolin, M. (2014). Can we forecast the implied volatility surface dynamics of equity options? *Journal of Banking & Finance* 46.

Bollen, N., Whaley, R. (2004). Does net buying pressure affect the shape of implied volatility functions? *Journal of Finance* 59(2).

Carr, P., Wu, L. (2009). Variance risk premiums. *Review of Financial Studies* 22(3).

Cont, R., Da Fonseca, J. (2002). Dynamics of implied volatility surfaces. *Quantitative Finance* 2(1).

Coval, J., Shumway, T. (2001). Expected option returns. *Journal of Finance* 56(3).

Daglish, T., Hull, J., Suo, W. (2007). Volatility surfaces: theory, rules of thumb, and empirical evidence. *Quantitative Finance* 7(5).

Dumas, B., Fleming, J., Whaley, R. (1998). Implied volatility functions: empirical tests. *Journal of Finance* 53(6).

Fengler, M. (2009). Arbitrage-free smoothing of the implied volatility surface. *Quantitative Finance* 9(4).

Figlewski, S. (1989). Options arbitrage in imperfect markets. *Journal of Finance* 44(5).

Gatheral, J., Jacquier, A. (2014). Arbitrage-free SVI volatility surfaces. *Quantitative Finance* 14(1).

Gârleanu, N., Pedersen, L., Poteshman, A. (2009). Demand-based option pricing. *Review of Financial Studies* 22(10).

Goncalves, S., Guidolin, M. (2006). Predictable dynamics in the S&P 500 index options implied volatility surface. *Journal of Business* 79(3).

Muravyev, D. (2016). Order flow and expected option returns. *Journal of Finance* 71(2).

Cao, J., Han, B. (2013). Cross section of option returns and idiosyncratic stock volatility. *Journal of Financial Economics* 108(1).

Muravyev, D., Pearson, N. (2021). Options trading costs are lower than you think. *Review of Financial Studies* 33(11).

Politis, D., Romano, J. (1994). The stationary bootstrap. *Journal of the American Statistical Association* 89(428).