Original Research Article

Regime Dependence of the Size Premium: Identification Versus Ex-Ante Forecastability in the Carhart Four-Factor Model

Ethan Wuang

Lake Forest Academy, 1500 West Kennedy Road, Lake Forest, IL 60045, United States

ABSTRACT

Among the four factors of the Carhart four-factor model, which premium is regime-dependent — and can that dependence be exploited in real time? A factor-attribution methodology isolates each factor's regime contribution across 18 size-sorted portfolios and three evaluation windows (January 1927–November 2025). As a regime-identification exercise, the size factor (SMB) is uniquely regime-dependent: forecast improvement scales with SMB loading (Spearman $\rho = 0.948$; pooled dependence-robust $p = 0.12$ at the automatically selected block length, falling to 0.03 at fixed 12-month blocks), and regime parameters show the premium is approximately zero in the calm regime and concentrated in the turbulent regime at 0.5–1.1% per month (calm-stress difference positive in all three windows, one-sided bootstrap $p \leq 0.030$, though the windows are nested and two lean heavily on one historical episode), consistent with countercyclical risk compensation; a parametric Monte Carlo test against a single-state GARCH null indicates this difference exceeds what a volatility-clustered process without switching manufactures ($p = 0.015$–0.090 across windows); high-minus-low (HML) switching is counterproductive. These identification results do not, however, translate into ex-ante forecasting skill. Under a fully ex-ante design — parameters estimated on training data only, combined with one-step-ahead predicted regime probabilities — the regime model's out-of-sample R^2 is 0.07–0.08% depending on the regime-mean estimator, no portfolio-window test is significant even before adjustment (minimum Clark-West $p = 0.081$), and none survives a 5% false-discovery rate across the full 54-test family (minimum adjusted $p = 0.709$). Moving-block bootstrap inference preserving serial and cross-sectional dependence renders even the strongest single-window cross-sectional correlation (2010, naive $p \approx 7\times10^{-9}$) indistinguishable from zero (bootstrap $p = 0.40$; pooled $p = 0.86$). A three-stage decomposition attributes the apparent gains of less disciplined designs to parameter and probability-timing look-aheads, with the timing channel dominating. The size premium is regime-dependent in identification but not exploitably forecastable.

Keywords: Markov Regime Switching; Carhart Four-Factor Model; Size Premium; Look-Ahead Bias; Out-Of-Sample Forecasting; Factor Attribution

INTRODUCTION

Factor asset pricing models decompose portfolio returns into exposures to systematic risk factors. Fama and French (1993) augmented the capital asset pricing model of Sharpe (1964) and Lintner (1965) with size (SMB) and value (HML) factors (1–3), and Carhart (1997) added momentum (UMD), producing the four-factor benchmark used throughout this study (4); the value and

Corresponding author: Ethan Wuang, E-mail: 789wethan@gmail.com.

Accepted August 4, 2026
https://doi.org/10.70251/HYJR2348.4467669l

momentum premia in particular have been documented across asset classes and international markets (5). A core assumption of the standard implementation is that factor premia are constant over time — convenient but empirically questionable, a premise Fama and French (1997) themselves qualified (6). Campbell and Cochrane (1999) showed that habit formation implies countercyclical risk premia (7); Lettau and Ludvigson (2001) demonstrated that the consumption-wealth ratio predicts the equity premium at business-cycle frequencies (8); and Kelly, Pruitt, and Su (2019) used latent factor models to capture continuous time variation in premia (9). Markov regime-switching models (10) provide a natural framework for the complementary possibility of discrete shifts.

Prior applications of regime switching to factor models typically apply switching across all factors simultaneously. Ang and Bekaert (2002) documented that market correlations rise sharply in bear regimes (11), and Guidolin and Timmermann (2007) showed that ignoring regime dynamics leads to significant portfolio welfare losses (12); Ang and Timmermann (2012) survey this literature (13). Such specifications assume uniform regime dependence without testing it. Theory, however, points to one factor in particular: Perez-Quiros and Timmermann (2000) found that small firms are disproportionately sensitive to credit-market tightness and recessions, generating cyclical variation in the size premium specifically (14).

Any forecasting claim must also clear the out-of-sample discipline of Goyal and Welch (2008), who demonstrated that many in-sample predictors fail out of sample against the historical-mean benchmark (15). Forecasts here are therefore evaluated with the Clark and West (2007) test for nested models (16), the Campbell and Thompson (2008) out-of-sample (OOS) R^2 (17), and the Diebold and Mariano (1995) test with the Harvey, Leybourne, and Newbold (1997) small-sample correction (18,19). Multiple testing is controlled with the Benjamini-Hochberg false-discovery-rate procedure over the full family of portfolio-window tests (20), and inference for cross-sectional statistics uses a moving-block bootstrap with automatic block-length selection that preserves both serial and cross-sectional dependence (21).

This paper accordingly asks two targeted questions. First, which factor of the Carhart model is regime-dependent? Second — and separately — does that dependence improve fully ex-ante out-of-sample forecasts? Keeping the two questions distinct turns out to be essential. As a regime-identification exercise, the evidence locates regime dependence in the size factor, with the two strands of that evidence carrying different weight: the parameter-based finding is robust — the size premium is approximately zero in calm regimes and concentrated in turbulent regimes at 0.5–1.1% per month, a difference that survives dependence-robust inference in every window — while the cross-sectional finding that improvement scales with SMB exposure, though descriptively strong and unique to SMB among the four factors, is only marginally distinguishable from zero once the overlap among evaluation windows is respected. The answer to the second question is negative. Under a fully ex-ante design — parameters estimated on training data only, combined with one-step-ahead predicted regime probabilities — out-of-sample gains are negligible and no test is statistically significant, individually or after false-discovery-rate adjustment; dependence-robust bootstrap inference likewise renders the cross-sectional correlation indistinguishable from zero. A three-stage decomposition attributes the apparent forecasting gains of less disciplined designs to two distinct look-ahead channels — full-sample parameter estimation and contemporaneous (filtered) probability timing — with the timing channel contributing more.

The paper makes three contributions: a factor-attribution diagnostic for locating which factor in a multi-factor model carries regime dependence, together with an assessment of how much formal weight that diagnostic can bear; a characterization of the size premium as concentrated in high-volatility regimes, consistent with countercyclical risk compensation; and a look-ahead decomposition that separates regime identification from real-time implementability and quantifies how each look-ahead channel inflates apparent forecasting performance. The remainder of the paper describes the methods and data, presents the identification and ex-ante results, and discusses interpretation, implications, and limitations.

METHODS AND MATERIALS

Baseline Model

The baseline is the Carhart four-factor model (4). For portfolio i at time t:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_{i,mkt}(R_{m,t} - R_{f,t}) + \beta_{i,SMB}\cdot SMB_t + \beta_{i,HML}\cdot HML_t + \beta_{i,UMD}\cdot UMD_t + \varepsilon_{i,t} \quad \text{(Eq. 1)}$$

Factor loadings are estimated by ordinary least squares on training data only, separately for each evaluation window, so loadings for the same portfolio

differ slightly across windows (see the note beneath Table 3). The static forecast uses unconditional training-sample means for all four premia and is constant over the test period.

Factor Attribution Analysis

For each factor k ∈ {Mkt-RF, SMB, HML, UMD}, a partial regime forecast is constructed by applying regime switching to factor k alone while holding all other factors at unconditional training means. The change in mean squared error attributable to factor k is computed across all 18 portfolios and three windows. This isolates the factor through which any regime dependence operates. The two-state specification is applied to all four factors by analogy rather than validated separately for each, so a null attribution result may reflect either the absence of regime dependence or a poorly suited regime specification for that factor.

Markov Regime-Switching Specification

SMB is modeled as a two-state Markov regime-switching process (10):

$$\mathrm{SMB}_t = \mu_{s(t)} + \varepsilon_t,\ \ \varepsilon_t \sim \mathrm{N}(0, \sigma^2_{s(t)}) \qquad \text{(Eq. 2)}$$

where $s_t \in \{0,1\}$ is the latent regime (calm = lower variance, stress = higher variance), evolving as a first-order Markov chain with transition probabilities $p_{jj} = \Pr(s_t = j \mid s_{t-1} = j)$. Model selection is reported in Table 8, Panel A. A single-state (constant mean and variance) specification is decisively rejected by the Bayesian information criterion in every training window, by 186 to 229 BIC units. The apparent in-sample preference for three states (BIC favors three states in the 1995 and 2010 windows, is indifferent in 2000; AIC favors three states in all three windows) does not reflect a genuine competing model. An unconstrained three-state fit is not well-identified for this series: in every training window the added state collapses onto a handful of outlier months with an implied mean of roughly +6% to +16% per month, the well-known unbounded-likelihood degeneracy of maximum-likelihood estimation for mixture and regime-switching models with unrestricted variances (22) (Table 8, Panel A note). Repeated random-restart search does not correct this and can worsen it, since the degenerate solution has an unbounded likelihood. Both the information criteria and the reversed three-state cross-sectional pattern (Table 4) inherit this artifact. The two-state specification is retained because it is the only well-identified model considered; the identification claim rests on it alone. Testing the two-state model against the null of no switching by a likelihood-ratio statistic is not valid, because the transition probabilities are unidentified under that null and the statistic does not have its standard asymptotic distribution (23,24). A parametric Monte Carlo test is used instead. For each training window, a single-state null model — constant mean with GARCH(1,1) errors and Student-t innovations, reproducing volatility clustering and fat tails but containing no regime process and no time-varying mean — is fitted to the SMB series; 1,000 series of the same length are simulated from the fitted null; and the identical two-state model is estimated on each. The Monte Carlo p-value is the fraction of null series whose fitted calm-stress mean difference is at least as large as the observed difference. Because Markov-switching likelihoods can be multimodal, each training-window fit was additionally re-estimated with 30 randomized starting values; in every window the default-start solution attained the highest log-likelihood found (agreement to within 10^{-10}), so the reported estimates are not local optima.

Forecast Designs and Look-Ahead Structure

Two regime-probability sequences are distinguished. Filtered probabilities $P(s_t \mid \mathscr{F}_t)$ condition on SMB observations through month t and therefore embed contemporaneous information about the month being forecast. Predicted probabilities $P(s_t \mid \mathscr{F}_{t-1})$ are obtained from the prediction step of the Hamilton filter — the previous month's filtered state propagated through the transition matrix — and use information through month t−1 only (10). Similarly, the parameters governing the filter can be estimated on the training sample only or on the full sample; the latter introduces a parameter-level look-ahead.

The *regime-identification design* combines full-sample parameters with filtered probabilities. It is the sharpest available lens for characterizing regimes ex post, but it is not a valid forecasting test, and it is labeled as an ex-post exercise throughout. The *fully ex-ante design* — training-only parameters with predicted probabilities — is the primary forecasting specification of the paper. In either design, the regime forecast replaces the static SMB premium with the regime-conditional expectation while holding all other premia at training means:

$$\hat{\mathrm{E}}[R_{i,t} - R_{f,t}]_{\text{regime}} = \hat{\alpha}_i + \hat{\beta}_{i,\mathrm{mkt}}\cdot\bar{\mu}_{\mathrm{mkt}} + \hat{\beta}_{i,\mathrm{SMB}}\cdot\hat{\mathrm{E}}[\mathrm{SMB}_t \mid \mathscr{F}_\cdot] + \hat{\beta}_{i,\mathrm{HML}}\cdot\bar{\mu}_{\mathrm{HML}} + \hat{\beta}_{i,\mathrm{UMD}}\cdot\bar{\mu}_{\mathrm{UMD}} \qquad \text{(Eq. 3)}$$

where $\mathcal{F}_\cdot$ denotes $\mathcal{F}_t$ (filtered) or $\mathcal{F}_{t-1}$ (predicted) according to the design, and the regime-conditional premium is $\hat{E}[SMB_t \mid \mathcal{F}_\cdot] = \Sigma_j P(s_t = j \mid \mathcal{F}_\cdot)\cdot\hat{\mu}_j$. Two estimators of $\hat{\mu}_j$ appear in this paper: the fitted regime intercepts of Eq. 2 (Tables 2, 6 Panels A and C, and 7) and the probability-weighted average of training-sample SMB returns (Tables 1, 3, 5, Figure 1, and Table 6 Panel B), which differ by about 0.1 percentage point per month here. Both are reported because the conclusions do not depend on the choice: the ex-ante mean out-of-sample R^2 is 0.080% and 0.068% respectively, with no significant test under either. A three-stage decomposition re-runs the full analysis under (A) full-sample parameters with filtered probabilities, (B) training-only parameters with filtered probabilities, and (C) training-only parameters with predicted probabilities, isolating the parameter channel (A→B) and the timing channel (B→C) in turn.

Estimation And Evaluation

The training samples run from January 1927 through December 1994, December 1999, and December 2009; the corresponding test periods begin January 1995, January 2000, and January 2010 and all end in November 2025 (371, 311, and 191 monthly observations, respectively). Hidden-Markov-model parameters are estimated once per training window — not recursively — by maximum likelihood using statsmodels 0.14.6 (MarkovRegression) under Python 3.13 with package defaults; predicted probabilities are the package's predicted marginal probabilities. All model fits converged — including the 15,000 re-fits inside the dependence bootstrap and the 45,000 inside the parameter-difference bootstrap described below — so no failed-fit handling was required.

For portfolio i, forecast improvement is measured by $\Delta MSE_i = MSE_i(\text{regime}) - MSE_i(\text{static})$, with negative values indicating that the regime forecast is more accurate (Table 1); equivalently, by the mean of the monthly loss differential $d_t = e^2_{\text{static},t} - e^2_{\text{regime},t}$, which is positive when the regime model improves (Tables 6–7). The Campbell-Thompson out-of-sample R^2 is 1 − MSE(regime) / MSE(historical mean) (17). The Clark-West statistic tests the null of equal predictive accuracy for nested models using the adjusted mean-squared-prediction-error (MSPE) correction, with one-sided p-values (16). The Diebold-Mariano statistic is computed on the squared-error loss differential at horizon one with the Harvey-Leybourne-Newbold correction and compared to a Student-t distribution with n−1 degrees of freedom, two-sided (18,19). The multiple-testing family is defined ex ante as all 54 portfolio-window tests (18 portfolios × 3 windows), and Benjamini-Hochberg adjustment is applied across this full family (20); defining the family over every test examined, rather than a favorable subset, follows the multiple-testing discipline urged for cross-sectional asset pricing by Harvey, Liu, and Zhu (2016) (25).

Because the 18 portfolios overlap in their underlying securities and the three evaluation windows overlap in time, standard rank-correlation and sign-test p-values overstate precision. Inference for cross-sectional statistics therefore uses a moving-block bootstrap (5,000 replications; random seed 20260706) that resamples calendar months in blocks, moving the entire 18-portfolio cross-section together, so that both serial and cross-sectional dependence are preserved. Because the three test periods overlap in calendar time — the 1995 window contains the 2000 and 2010 windows — pooled statistics are bootstrapped by resampling the superset (1995-window) calendar once per replication and evaluating every window on the same drawn months, so that cross-window dependence is preserved rather than treated as independent information. Block lengths are selected automatically following Politis and White (2004) — one to two months on the loss-differential series — with fixed lengths of 6 and 12 months as robustness checks (21). The stationary bootstrap (26) is an alternative, but its randomized block lengths would not address the block-length sensitivity documented below, which is reported at all three lengths rather than dismissed. Bootstrap p-values are percentile (confidence-interval-inversion) p-values: they report the smallest level at which the resampling distribution of the statistic excludes zero, and therefore test whether the estimated statistic is distinguishable from zero given the dependence structure of the data, rather than testing against a simulated no-regime-dependence null. Explicitly, write $\hat{P}(\cdot)$ for the empirical proportion across the B = 5,000 resampled statistics $\hat{\theta}^b$. Two-sided p-values are $p = \min\{1, 2 \cdot \min[\hat{P}(\hat{\theta}^b \le 0), \hat{P}(\hat{\theta}^b \ge 0)]\}$; the one-sided p-value, used only for the calm–stress mean difference in Table 6, Panel C, is $\hat{P}(\mu_{\text{stress}} - \mu c_{\text{alm}} \le 0)$. For continuous statistics the two-sided form is the smallest α at which the equal-tailed (1 − α) percentile interval excludes zero, so p-values and the reported 95% intervals are inversions of one another. Because the inversion uses the empirical distribution rather than a symmetric approximation, it assumes no symmetry and remains well defined under skew or boundary pile-up, with the asymmetry carried by the intervals. Two cases

warrant note. The sign statistic is discrete, so resamples at exactly zero enter both tail proportions and its two-sided p is conservative relative to strict inversion; this is immaterial here, as every sign-statistic p-value exceeds 0.42. Degenerate re-fits (stress occupancy below 2% of training months) are retained without reweighting, so p-values and intervals use the same resample set, and no resample failed to converge in any window. Because these re-fits inflate fitted stress means they load the right tail of the calm–stress difference, whereas the one-sided p-value depends on the left tail, so retaining them barely moves it. For p_{11} they concentrate the distribution far below the full-sample estimate; the inversion is uninformative and neither interval nor p-value is reported. Reported intervals are percentile intervals, and the same bootstrap yields confidence intervals for the regime parameters, ΔMSE, out-of-sample R^2, and the rank correlations. In every bootstrap re-fit of the regime model, regimes are relabeled by variance ordering before parameters are recorded; re-fits in which the stress regime is degenerate (smoothed occupancy below 2% of training months) are retained in the primary intervals, with trimmed intervals reported as a sensitivity. Two companion applications complete the inference program: the same moving-block resampling applied to each training sample, with full re-estimation of the regime model on every resample, provides intervals and a one-sided test for the difference $\mu_{stress} - \mu_{calm}$ (Table 6, Panel C); and the same resampling applied to the identification design's loss differentials provides dependence-robust inference for the identification-design correlations (Table 6, Panel B), so a single inferential standard is applied symmetrically to both designs.

Robustness analyses within the identification design repeat the cross-sectional test with five split dates, with the 2008–2009 financial crisis excluded, under proportional transaction costs of 10 bp, and under an attempted three-state specification (found to be non-identified; see Results and Discussion); a Welch two-sample t-test comparing mean SMB returns before and after a fixed January 2008 split date, together with a Levene test for variance equality across the same split, is also reported; no formal break-point search (e.g., Quandt-Andrews) is performed, and the fixed split date is pre-specified.

Data

Factor and portfolio data are from the Kenneth R. French Data Library (27): Portfolios_Formed_on_ME.csv for the test portfolios and F-F_Research_Data_Factors.csv and F-F_Momentum_Factor.csv for the four factors (all accessed June 2026). Returns are converted from percent to decimal, and all series are aligned on their common monthly dates. The sample spans January 1927 through November 2025, yielding 1,187 monthly observations; the sample ends in November 2025 because, in the June 2026 vintage, the size-portfolio file lags the factor files, which extend through January 2026. Test portfolios are the 18 size-sorted portfolios formed on market capitalization of NYSE, AMEX, and NASDAQ stocks at the end of the prior June; the file's nineteenth category — firms with market equity less than or equal to zero, labeled "Negative (not used)" by the source — is excluded because it is not sorted on size. All returns are value-weighted and in excess of the one-month Treasury-bill rate. SMB factor loadings range from −0.21 (Hi 10) to 1.67 (Lo 10) across training windows, providing the cross-sectional variation needed for the monotonicity diagnostic.

RESULTS

Factor Attribution (Regime Identification)

Figure 1 and Table 1 present the attribution results under the regime-identification design; Figure 2

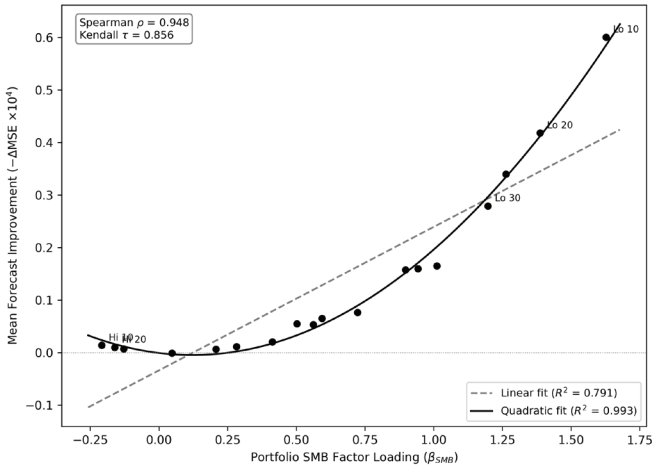

Figure 1. *Mean forecast improvement (−ΔMSE × 10^4) versus portfolio SMB factor loading across 18 size-sorted portfolios under the regime-identification design, averaged across the three evaluation windows (1995, 2000, 2010). Dashed line: linear OLS fit (R^2 = 0.791). Solid curve: quadratic fit (R^2 = 0.993). Spearman ρ = 0.948, Kendall τ = 0.856 (pooled dependence-robust p = 0.12 at the automatically selected block length, 0.03 at fixed 12-month blocks; Table 6, Panel B).*

summarizes the attribution by factor. SMB is the only factor whose improvement is systematically related to factor exposure: improvement occurs in 89% of portfolio-window combinations, with a Spearman rank correlation of ρ = +0.948 between SMB loading and improvement magnitude (Table 1) — the signature of genuine factor-level regime dependence. Market and momentum switching reduce MSE in many cases (67% and 87%) but with negative or zero rank correlations, indicating those gains are unrelated to the corresponding loadings. HML switching is actively counterproductive, increasing MSE in 87% of cases (ρ = −0.994). These results identify SMB as the locus of regime dependence and motivate the single-factor specification. Dependence-robust inference for the SMB correlation is presented in Table 6, Panel B.

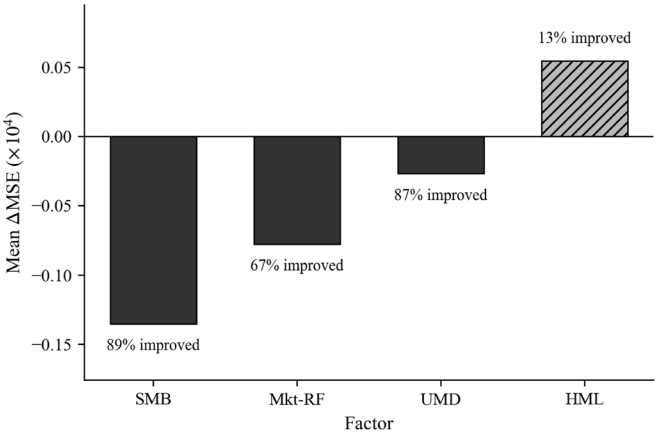

Figure 2. *Factor attribution bar chart: mean ΔMSE (×10⁴) by factor across all 54 portfolio-window combinations under the regime-identification design. Negative values indicate improvement. The HML channel is actively counterproductive.*

Regime Parameter Estimates

Table 2 reports the estimated two-state parameters for each training window, and Table 6 (Panel C) reports dependence-robust 95% confidence intervals. The regimes are persistent in-sample (calm duration 30–36 months; stress duration 6–23 months) and clearly separated by volatility (stress-regime volatility 2.3–2.8× the calm regime). Most strikingly, the stress-regime mean is substantially larger than the calm-regime mean: the size premium is close to zero in calm markets (0.0–0.1% per month) and is concentrated almost entirely in the turbulent regime (0.5–1.1% per month, or roughly 6–14%

Table 1. *Factor Attribution Summary (regime-identification design). Mean ΔMSE (×10⁴) and fraction improved across 18 portfolios × 3 evaluation windows; ΔMSE = MSE(regime) − MSE(static), negative indicates improvement. Spearman ρ is between the window-averaged factor loading and mean improvement across the 18 portfolios (the convention of Figure 1); dependence-robust inference for the SMB correlation is reported in Table 6, Panel B. Values are computed on the corrected 18-portfolio sample under the same conventions as Table 5, stage A, with regime labels aligned by variance ordering across fits.*

Factor	Mean ΔMSE (×10⁴)	% Improved	Spearman ρ	Loading–improvement relationship
SMB	−0.135	88.9%	+0.948	Positive
Mkt-RF	−0.078	66.7%	−0.470	Negative
UMD	−0.027	87.0%	−0.955	Negative
HML	+0.054	13.0%	−0.994	Negative

Table 2. *Regime Parameter Estimates. Two-state Markov switching parameters for monthly SMB returns, estimated on each training window (training-only). Regimes ordered by variance (calm = lower). μ and σ in percent per month; p_{jj} = probability of remaining in regime j; expected duration = 1/(1−p_{jj}) in months. Dependence-robust 95% confidence intervals for these parameters, and for the difference $\mu_{stress} - \mu_{calm}$, are reported in Table 6, Panel C.*

Split	μ calm	μ stress	σ calm	σ stress	p_{00}	p_{11}	Dur. calm	Dur. stress
1995	0.02%	0.65%	1.88%	4.59%	0.967	0.938	30.4	16.1
2000	−0.02%	0.48%	1.83%	4.34%	0.968	0.956	31.0	22.9
2010	0.09%	1.08%	2.34%	6.54%	0.972	0.820	35.7	5.6

annualized, compounded). A static model averages these into a single unconditional mean accurate in neither regime. The difference $\mu_{stress} - \mu_{calm}$ — the quantity on which the concentration claim rests — is positive under the one-sided bootstrap test in every window (p between 0.013 and 0.030 across windows and block lengths; Table 6, Panel C), though the two-sided 95% interval for the 2000 window includes zero at its lower edge (−0.01% per month). Excluding the minority of bootstrap re-fits with degenerate stress regimes (up to 14.5% of replications) leaves the one-sided conclusion unchanged. Interval lower bounds are nearly fixed (1995 window: 0.045% to 0.028% per month), while the degenerate-driven upper tails contract sharply (12.20% to 8.74%). The contrast with the uninformative persistence intervals reported below is not a contradiction: when few stress months are drawn the likelihood is nearly flat in p_{11} and the estimate is driven to the boundary, so persistence is only weakly identified in those resamples, whereas the regime means remain estimable, if imprecisely; the mean difference therefore retains its sign in exactly the resamples where p_{11} collapses. The bootstrap intervals add an important qualification: calm-regime parameters are estimated precisely, but stress-regime parameters are not — stress-regime persistence in particular cannot be assigned a meaningful percentile interval, because a substantial minority of bootstrap resamples contain too few stress months to estimate persistence reliably, leaving the likelihood nearly flat in p_{11} (Table 6, Panel C note), reflecting the small number of stress episodes any training sample contains. Figure 3 shows the filtered stress-regime probability over the full sample: probability spikes cluster around recessions as dated by the National Bureau of Economic Research (NBER) (28) and other periods of elevated volatility, consistent with the business-cycle interpretation of the stress regime.

Test Against A No-Switching Null

Table 8 (Panel B) reports the parametric Monte Carlo test. Fitting the two-state model to series simulated from a single-state GARCH null produces calm-stress mean differences centered on zero (median −0.004% to +0.004% per month, positive in 50% of replications), so the two-state specification does not systematically manufacture a positive difference. Its tails are wide, however: differences of 0.44–0.47% per month arise in 10% of null series and 0.62–0.66% in 5%. Against this null the observed differences are large enough to be unusual in all three windows, though not uniformly decisive — Monte Carlo p = 0.050 (1995), 0.090 (2000), and 0.015 (2010). The three training samples are nested, so these are not independent tests and cannot be combined into a joint statistic. The 2010 window is additionally distinctive in regime shape: its stress state

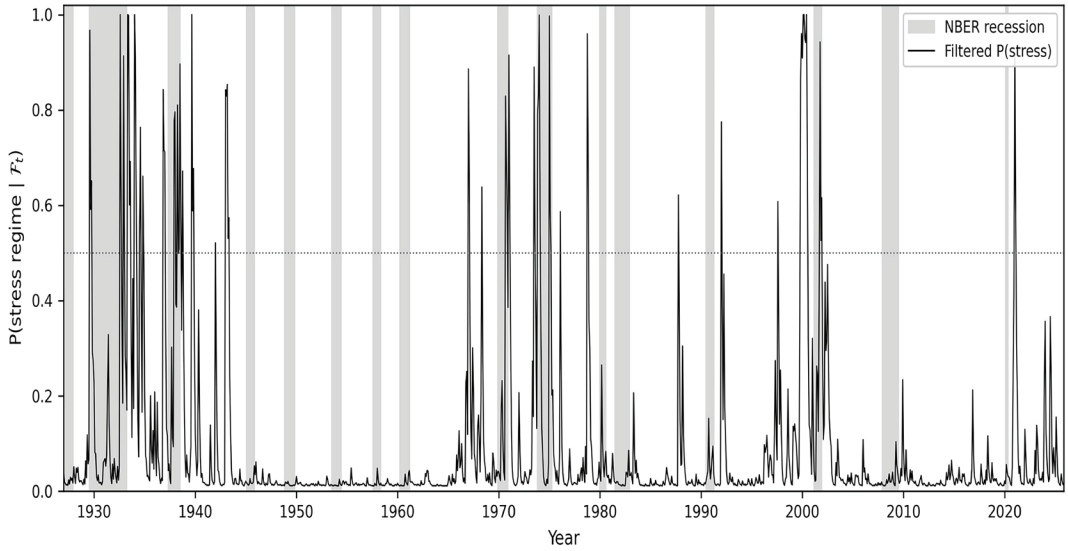

Figure 3. *Filtered SMB stress-regime probability, January 1927–November 2025, estimated on the full sample (regime-identification design; full-sample parameters). Grey shading indicates NBER-dated recessions, from the NBER Business Cycle Dating Chronology (28); the dotted line marks p = 0.5. Stress-regime probability rises around NBER-dated recessions — the periods in which Table 2 indicates the size premium is concentrated.*

is far shorter and rarer than the null typically produces (expected duration 5.6 months against a null median of 20.8, second percentile; occupancy 13.5% against 30.3%, fourth percentile), whereas the 1995 and 2000 stress states are close to what the null generates. The regime structure of the size premium therefore exceeds what a volatility-clustered process without switching would manufacture, most clearly in the most recent window, and the evidence in the 1995 and 2000 windows is correspondingly weaker.

Cross-Sectional Pattern (Regime Identification)

Under the identification design, the Spearman rank correlation between portfolio SMB loading and mean forecast improvement is 0.948 and Kendall τ is 0.856, implying 92.8% of pairwise portfolio comparisons are concordant; the pattern holds in each window separately (ρ: 0.92, 0.99, 0.91). A quadratic fit yields $R^2 = 0.993$ versus $R^2 = 0.791$ for the linear fit, and a sign test shows 48 of 54 portfolio-window combinations (88.9%) with positive improvement. Naive p-values for these statistics — reported as $p < 0.001$ in an earlier version of this analysis — treat the portfolios and windows as independent and therefore overstate precision. Under the dependence-preserving bootstrap (Table 6, Panel B) — with pooled statistics resampled on the joint calendar so that the overlap among the three test periods is respected — the pooled correlation is at best marginally distinguishable from zero ($\rho = 0.948$; bootstrap $p = 0.121$ at the automatic block length, 0.063 and 0.034 at fixed 6- and 12-month blocks), and no individual window is significant at the 5% level (minimum $p = 0.076$). The monotone pattern is thus a strong descriptive regularity whose formal significance is sensitive to the block length, reflecting the concentration of improvements in a small number of high-volatility months; the parameter-based evidence for regime dependence — the calm-stress mean difference reported above — does not share this sensitivity. Table 3 reports the corresponding portfolio-level Campbell-Thompson OOS R^2 and mean-squared-forecast-error (MSFE)-adjusted statistics (against the prevailing-mean benchmark) for the three smallest size deciles under the identification design; these are ex-post statistics, inflated by construction by the design's look-ahead (quantified in Table 5), and are reported descriptively.

Fully Ex-Ante Forecasting (Primary Specification)

Table 7 reports the primary forecasting results. Under training-only parameters and predicted probabilities, the regime model's average out-of-sample R^2 across the 18 portfolios is 0.050%, 0.030%, and 0.160% in the 1995, 2000, and 2010 windows — 0.080% overall. Under the alternative regime-mean estimator (Methods), the

Table 3. *Campbell-Thompson OOS R^2 and Clark-West Statistics — Small-Cap Portfolios, regime-identification design (ex post). OOS R^2 relative to the prevailing historical mean. The MSFE-adjusted (Clark-West form) statistic tests H_0: the prevailing-mean benchmark is as accurate as the regime model (one-sided p, nominal). These statistics are inflated by construction by the design's look-ahead and are reported descriptively; the decomposition in Table 5 quantifies the inflation, and valid ex-ante inference is reported in Table 7. Values are generated by the unified pipeline of Table 5 (stage A), so the two tables agree by construction. Note: factor loadings are re-estimated separately on each window's training sample; values for the same portfolio therefore differ across evaluation windows by construction.*

Portfolio	β SMB	Split	OOS R^2	MSFE-adj. stat	p (nominal)
Lo 10	1.67	1995	1.46%	2.518	0.006
Lo 10	1.63	2000	1.15%	2.252	0.012
Lo 10	1.59	2010	2.00%	2.465	0.007
Lo 20	1.41	1995	1.02%	2.151	0.016
Lo 20	1.38	2000	0.83%	1.956	0.025
Lo 20	1.37	2010	1.44%	2.034	0.021
Lo 30	1.20	1995	0.75%	1.914	0.028
Lo 30	1.19	2000	0.61%	1.699	0.045
Lo 30	1.19	2010	1.02%	1.972	0.024

Table 4. *Robustness Summary (regime-identification design). Spearman rank correlation between SMB loading and forecast improvement across 18 portfolios. Baseline: ρ = 0.948. Nominal p-values treat portfolios and windows as independent and overstate precision; dependence-robust inference for the identification design is reported in Table 6, Panel B. The three-state row is non-identified: the maximum-likelihood fit collapses onto 4–10 outlier months per window (0.7–1.8% occupancy) with an implausible fitted mean of roughly +6% to +16% per month, the unbounded-likelihood degeneracy of unrestricted-variance mixture and regime-switching models; see Methods.*

Test	Spearman ρ	Nominal p	Frac. Improved	Interpretation
Baseline (3 splits)	0.948	< 0.001	88.9%	Baseline
Five split dates	0.971	< 0.001	100%	Pattern persists descriptively
Exclude crisis 2008–2009	0.901	< 0.001	88.9%	Pattern persists descriptively
Transaction costs 10 bp	0.969	< 0.001	61.1%	Fewer net improve
Three-state model	n/a	n/a	n/a	Non-identified

Table 5. *Look-Ahead Decomposition (three stages, unified pipeline; stage A replicates the identification design). Regime OOS R^2 is the mean Campbell-Thompson OOS R^2 of the regime model across all 54 portfolio-window evaluations, computed using the probability-weighted regime-mean estimator (Methods); stage C therefore differs slightly from the window means of Table 6, Panel A, which use the fitted regime intercepts — 0.068% versus 0.080% overall, with no significant test under either; mean MSFE stat is the mean MSFE-adjusted test statistic; frac. sig. is the fraction of the 54 tests nominally significant at 5% (one-sided, unadjusted). All MSFE-adjusted statistics in this table test the regime model against the prevailing-mean benchmark and are therefore not identical to the Clark-West statistics of Table 7, which test the regime model against the nested static model. Panel B reports the mean MSFE-adjusted p-value for the three smallest deciles at each stage.*

Panel A: Aggregate

Design	Regime OOS R^2	Mean MSFE stat	Frac. sig. 5%
A: Full-sample params + filtered	0.41%	1.334	38.9%
B: Training-only params + filtered	0.37%	0.957	22.2%
C: Training-only params + predicted (fully ex ante)	0.07%	0.418	0.0%

Panel B. Small-cap MSFE-adjusted p-values

Portfolio	A: full + filtered	B: train + filtered	C: train + predicted
Lo 10	0.008	0.059	0.179
Lo 20	0.021	0.065	0.213
Lo 30	0.032	0.090	0.289

corresponding overall figure is 0.068%, with a negative mean in the 2000 window; both are economically negligible and neither yields a significant test. No portfolio in any window is significant even at the unadjusted 5% level: the strongest of the 54 tests is Lo 10 in the 2010 window, with a Clark-West statistic of 1.396 (one-sided p = 0.081). After Benjamini-Hochberg adjustment across the full 54-test family, the minimum adjusted p-value is 0.709. The Diebold-Mariano test agrees: no test is significant (minimum two-sided p = 0.141, Lo 10 in the 2010 window). Because the static model is nested in the regime model, the DM test is expected to be conservative — the motivation for the Clark-West adjustment (16) — but the two tests lead to the same conclusion here: there is no statistically detectable ex-ante forecasting gain.

Dependence-Robust Inference

Table 6 (Panel A) reports the moving-block bootstrap results for the fully ex-ante design. The cross-sectional rank correlation is 0.940 in the 2010 window but −0.059

and −0.189 in the 1995 and 2000 windows, with a pooled value of 0.104. Under the dependence-preserving bootstrap, none is distinguishable from zero: the 2010-window correlation — which a naive calculation would assign $p \approx 7\times10^{-9}$ — has a bootstrap p-value of 0.40 with a 95% confidence interval of [−0.87, 1.00]. Results for the ex-ante design are insensitive to the block length (automatic selection of 1–2 months versus fixed 6 and 12 months), and the intervals for the mean loss differential and out-of-sample R^2 straddle zero in every window. The wide, left-skewed intervals are consistent with a correlation driven by a small number of high-volatility months rather than by a stable cross-sectional relationship.

Look-Ahead Decomposition

Table 5 presents the three-stage decomposition, estimated within a unified pipeline so that the stages are directly comparable (stage A replicates the identification design). Moving from full-sample to training-only parameters (A→B) reduces the mean MSFE-adjusted statistic (regime model versus the prevailing-mean benchmark) from 1.334 to 0.957 and the fraction of nominally significant tests from 38.9% to 22.2%; small-cap p-values cross from significant to marginal. Replacing filtered with predicted probabilities (B→C) is the larger correction, reducing the mean statistic to 0.418 and the significant fraction to zero; overall, the regime out-of-sample R^2 falls from 0.41% to 0.07% across the

Table 6. *Dependence-Robust Bootstrap Inference. Moving-block bootstrap, 5,000 replications, resampling calendar months in blocks with the entire 18-portfolio cross-section moved together; block lengths selected automatically per Politis and White (2004) (21) (1–2 months), with fixed 6- and 12-month blocks as robustness checks (ex-ante results insensitive; the pooled identification statistic is not — see Panel B). p-values are two-sided percentile bootstrap p-values; intervals are 95% percentile intervals. S is the sign-test statistic across the 18 portfolios. Pooled rows compute each portfolio's improvement as the mean across the three evaluation windows before forming the cross-sectional statistic, so pooled values need not lie between the per-window values.* $\bar{d}$ *is the mean monthly loss differential (static − regime), positive when the regime model improves.*

Panel A: Cross-sectional statistics and forecast-accuracy measures

Window	Statistic	Estimate	Boot p	95% CI
1995	Spearman ρ	−0.059	1.00	[−0.996, 0.994]
1995	Kendall τ	−0.150	0.90	[−0.974, 0.961]
1995	Sign statistic S	0	0.87	[−14, 16]
1995	$\bar{d}$ (×10⁴)	−0.006	0.95	[−0.152, 0.150]
1995	OOS R^2 (mean)	0.050%	0.80	[−0.32%, 0.43%]
2000	Spearman ρ	−0.189	0.90	[−0.996, 0.996]
2000	Kendall τ	−0.255	0.76	[−0.974, 0.974]
2000	Sign statistic S	−8	0.71	[−12, 18]
2000	$\bar{d}$ (×10⁴)	−0.013	0.85	[−0.157, 0.131]
2000	OOS R^2 (mean)	0.030%	0.88	[−0.37%, 0.41%]
2010	Spearman ρ	0.940	0.40	[−0.870, 0.998]
2010	Kendall τ	0.830	0.46	[−0.778, 0.987]
2010	Sign statistic S	18	0.57	[−12, 18]
2010	$\bar{d}$ (×10⁴)	0.047	0.38	[−0.054, 0.159]
2010	OOS R^2 (mean)	0.160%	0.38	[−0.21%, 0.53%]
Pooled	Spearman ρ	0.104	0.86	[−0.971, 0.996]
Pooled	Kendall τ	−0.046	0.99	[−0.908, 0.974]
Pooled	Sign statistic S	−2	0.95	[−12, 18]
Pooled	$\bar{d}$ (×10⁴)	0.009	0.89	[−0.104, 0.123]

Continued Table 6. *Dependence-Robust Bootstrap Inference. Moving-block bootstrap, 5,000 replications, resampling calendar months in blocks with the entire 18-portfolio cross-section moved together; block lengths selected automatically per Politis and White (2004) (21) (1–2 months), with fixed 6- and 12-month blocks as robustness checks (ex-ante results insensitive; the pooled identification statistic is not — see Panel B). p-values are two-sided percentile bootstrap p-values; intervals are 95% percentile intervals. S is the sign-test statistic across the 18 portfolios. Pooled rows compute each portfolio's improvement as the mean across the three evaluation windows before forming the cross-sectional statistic, so pooled values need not lie between the per-window values.* $\bar{d}$ *is the mean monthly loss differential (static − regime), positive when the regime model improves.*

Panel B: Identification design (stage A), cross-sectional statistics. Per-window rows use the automatic block length. Pooled rows use joint-calendar resampling (see Methods) and are shown at all three block lengths because the pooled result is block-length-sensitive; pooled Kendall τ = 0.856 behaves identically (p = 0.148, 0.079, 0.043).

Window	Statistic	Estimate	Boot p	95% CI
1995	Spearman ρ	0.924	0.21	[−0.697, 1.000]
1995	Kendall τ	0.804	0.24	[−0.647, 1.000]
2000	Spearman ρ	0.990	0.25	[−0.823, 1.000]
2000	Kendall τ	0.948	0.29	[−0.725, 1.000]
2010	Spearman ρ	0.911	0.076	[−0.226, 0.994]
2010	Kendall τ	0.804	0.09	[−0.229, 0.974]
Pooled (auto)	Spearman ρ	0.948	0.121	[−0.445, 1.000]
Pooled (6-mo)	Spearman ρ	0.948	0.063	[−0.121, 1.000]
Pooled (12-mo)	Spearman ρ	0.948	0.034	[0.141, 1.000]

Panel C: Regime parameters and calm-stress difference, point estimate with 95% interval (percent per month for μ and σ). One-sided p is the bootstrap $P(\mu_{stress} - \mu_{calm} \leq 0)$ ***at the automatic block length; across all windows and block lengths it ranges 0.013–0.030. † Percentile intervals for stress-regime persistence are not reported: in a substantial minority of resamples (327, 723, and 462 of 5,000 replications in the 1995, 2000, and 2010 windows) too few stress months are drawn to estimate persistence reliably — the likelihood is nearly flat in*** p_{11} ***and the estimate is driven to the boundary — concentrating the bootstrap distribution of*** p_{11} ***far below the full-training-sample estimate; the resulting intervals fail to cover the point estimates and are uninformative as uncertainty bands. This fragility is itself the finding: stress-regime persistence is the least-identified parameter in every window. The same degenerate resamples drive the extreme upper tails of the μ stress and σ stress intervals; trimming them leaves lower bounds essentially unchanged (1995 difference: 0.045 → 0.028% per month) while pulling the upper bound of the calm-stress difference from 12.2% to 8.7% per month.***

Parameter	1995 window	2000 window	2010 window
μ calm	0.02 [−0.25, 0.30]	−0.02 [−0.34, 0.26]	0.09 [−0.19, 0.27]
μ stress	0.65 [0.15, 12.28]	0.48 [0.09, 14.63]	1.08 [0.24, 11.02]
σ calm	1.88 [1.57, 2.68]	1.83 [1.45, 2.79]	2.34 [1.92, 2.78]
σ stress	4.59 [3.48, 14.14]	4.34 [3.32, 14.48]	6.54 [3.88, 14.49]
p_{00} (calm)	0.967 [0.715, 0.991]	0.968 [0.739, 0.994]	0.972 [0.806, 0.991]
p_{11} (stress)	0.938 [n.r.†]	0.956 [n.r.†]	0.820 [n.r.†]
μ stress − μ calm	0.63 [0.04, 12.20]	0.50 [−0.01, 13.99]	1.00 [0.14, 11.16]
One-sided p (diff ≤ 0)	0.020	0.026	0.017

three stages (0.41% → 0.37% → 0.07%), with the timing step accounting for most of the decline. For the smallest decile, the mean MSFE-adjusted p-value rises from 0.008 (A) to 0.059 (B) to 0.179 (C). Both look-ahead channels therefore materially inflated the apparent small-cap forecasting gains, with the probability-timing channel — the use of contemporaneous month-t information in filtered probabilities — contributing more than the parameter channel. This ordering is consistent with the regime dynamics: because stress episodes are short-lived, the current month's own realization carries most of the regime information, and removing it removes most of the apparent skill.

Robustness (Identification Design)

Table 4 summarizes robustness of the cross-sectional pattern within the identification design. Five split dates strengthen the correlation to ρ = 0.971 with 100% of portfolios improving; excluding the 2008–2009 financial crisis yields ρ = 0.901, indicating the pattern is not driven by a single episode. Proportional transaction costs of 10 bp leave the rank ordering intact (ρ = 0.969) but reduce the fraction of portfolios with positive net improvement from 88.9% to 61.1%, indicating the economic margin is considerably thinner than the statistical one even before the ex-ante correction. The attempted three-state specification does not provide a usable check: as discussed in Methods, its maximum-likelihood fit is non-identified in every training window, collapsing onto a handful of outlier months rather than describing a genuine third regime, so no comparable three-state correlation exists to report. A Welch two-sample t-test finds no significant shift in mean SMB returns across the fixed January 2008 split (p = 0.262; Levene variance test p = 0.292), consistent with the regime characterization: the SMB process alternates between recurring states rather than undergoing a one-time discrete break. These checks characterize the stability of the ex-post pattern; they do not alter the ex-ante conclusions above.

DISCUSSION

Economic Interpretation

The regime parameter estimates in Table 2 give the identification result its economic content. The size premium is not earned smoothly: it is approximately zero in calm regimes and is concentrated in the turbulent regime, averaging 0.5–1.1% per month. This is the signature of countercyclical risk compensation. Small firms are disproportionately exposed to credit conditions and business-cycle risk (14); investors bearing that risk through high-volatility states earn a premium concentrated precisely where the risk materializes, consistent with habit-formation models in which risk

Table 7. *Fully Ex-Ante Forecasting Results — Small-Cap Portfolios (primary specification: training-only parameters, predicted probabilities). CW is the Clark-West statistic (one-sided p); DM is the Diebold-Mariano statistic with the Harvey-Leybourne-Newbold correction (two-sided p); BH-adj. p applies Benjamini-Hochberg across the full family of all 54 portfolio-window tests. Adjusted values are identical across these nine cells because the step-up procedure monotonizes upward from the rank minimising* $(54/j)p_{(j)}$*, which occurs at rank 34 (raw p = 0.446, adjusted = 54 × 0.446 / 34 = 0.709); the 34 tests ranked at or above this cutoff share the adjusted value 0.709, the remaining 20 receive larger values, and six distinct adjusted values arise across the full family. Across the full family, the minimum unadjusted CW p is 0.081, the minimum BH-adjusted p is 0.709, and no DM test is significant (minimum p = 0.141). Note: the stress-regime bootstrap intervals in Table 6 (Panel C) are consistent with this null.*

Portfolio	β SMB	Split	CW stat	CW p	DM stat	DM p	BH-adj. p
Lo 10	1.67	1995	0.709	0.239	0.351	0.726	0.709
Lo 10	1.63	2000	0.824	0.205	0.587	0.558	0.709
Lo 10	1.59	2010	1.396	0.081	1.477	0.141	0.709
Lo 20	1.41	1995	0.561	0.287	0.234	0.815	0.709
Lo 20	1.38	2000	0.466	0.321	0.231	0.818	0.709
Lo 20	1.37	2010	1.119	0.132	1.121	0.264	0.709
Lo 30	1.20	1995	0.270	0.394	−0.035	0.972	0.709
Lo 30	1.19	2000	0.140	0.444	−0.083	0.934	0.709
Lo 30	1.19	2010	0.868	0.193	0.845	0.399	0.709

premia rise as economic conditions deteriorate (7).

The ex-ante results sharpen rather than contradict this interpretation. The premium is concentrated in turbulent months, but those months are identified in real time only about as they occur: the automatic block lengths of one to two months on the loss differentials, the fragility of stress-regime persistence under resampling (Table 6, Panel C), and the dominance of the timing channel in the decomposition all indicate that last month's regime information is largely uninformative about this month. The size premium therefore appears as compensation for bearing turbulence, not as a timing signal — a reading consistent with, though not uniquely implied by, the risk-based interpretation: a premium that could be reliably timed would sit uneasily with compensation for risk, whereas a premium accruing in states identifiable only contemporaneously is what risk-based models predict. The absence of forecastability is not itself evidence for that interpretation, and it is the Monte Carlo evidence of Table 8, rather than the absence of forecastability, that distinguishes a genuine regime structure from one manufactured by fitting two states to a volatility-clustered series. The alternative reading of that window — that a stress state this rare reflects the model locking onto a few severe months rather than a recurring regime — is testable and is not supported: its fitted stress state comprises 17 discrete episodes, the largest holding 27% of stress months. The concentration caveat applies instead to the earlier windows, where a single Depression-era episode accounts for roughly 40% of stress months in each.

The convex cross-sectional structure (quadratic R^2 = 0.993 versus linear R^2 = 0.791) follows mechanically from the attribution design rather than indicating non-linear economic amplification. Because the regime and static forecasts differ only in the SMB term, a portfolio's forecast-error improvement equals its squared SMB loading times the gain in SMB-premium forecast accuracy, and is therefore quadratic in SMB loading by construction. What is not mechanical is the sign of the curvature: it is positive only when switching improves the factor-premium forecast, positive for SMB and negative

Table 8. *Model Selection and Test Against a No-Switching Null. Panel A: log-likelihood and Bayesian information criterion for one-, two-, and three-state specifications fitted to SMB on each training sample (lower BIC preferred). Panel B: parametric Monte Carlo test. A constant-mean GARCH(1,1) model with Student-t innovations — volatility clustering and fat tails, no regime process — is fitted per training window; 1,000 series are simulated from it and the identical two-state model estimated on each. Monte Carlo p is the fraction of null series whose fitted calm-stress mean difference equals or exceeds the observed difference. A likelihood-ratio test is not used because the transition probabilities are unidentified under the no-switching null (23,24). Training samples are nested across windows, so the three tests are not independent. The final column counts discrete stress episodes in each training window (maximal runs with smoothed P(stress) > 0.5) and the share of all stress months falling in the largest single episode.*

Panel A: Model selection (log-likelihood; BIC; AIC). Lower BIC and AIC preferred. The three-state column's information-criterion advantage reflects the same non-identification discussed in Methods and Table 4: an unconstrained third state can always raise the likelihood by collapsing onto a small number of outlier months, and multi-start search does not correct this, since the degenerate solution has an unbounded likelihood and a wide basin of attraction. The three-state comparison is reported for completeness, not as evidence of a genuine higher-order model.

Window	One state	Two states	Three states
1995	1667.9; −3322; −3332	1774.2; −3508; −3536	1816.3; −3552; −3609
2000	1784.4; −3555; −3565	1887.8; −3735; −3764	1907.2; −3733; −3790
2010	1996.1; −3978; −3988	2124.7; −4208; −4237	2175.0; −4267; −4326

Panel B: Monte Carlo test against the single-state null (1,000 replications per window; $\mu_{stress} - \mu_{calm}$ in percent per month)

Window	Observed	Null median	Null 90th	Null 95th	Monte Carlo p	Episodes (largest share)
1995	0.630	−0.004	0.463	0.627	0.050	14 (42%)
2000	0.499	0.001	0.466	0.664	0.090	14 (37%)
2010	0.998	0.004	0.439	0.618	0.015	17 (27%)

for HML — which is why HML switching raises MSE and reverses the rank correlation. The cross-sectional pattern thus functions as an attribution diagnostic, confirming that the identification-design gains operate through the SMB channel specifically. The failure of HML switching is consistent with the value premium being better captured by continuous state variables such as the consumption-wealth ratio (8) or latent factor models (9) than by discrete regimes.

Methodological Implications

The decomposition carries a message beyond this application. Apparent out-of-sample gains in regime-switching forecasts can arise through two separable look-ahead channels: full-sample parameter estimation, which embeds knowledge of future stress episodes in the filter, and filtered probability timing, which embeds the forecast month's own realization in the regime signal. Either alone inflates performance; in combination they inflated the regime out-of-sample R^2 by a factor of 5.9 here and manufactured nominally significant small-cap results from what is, ex ante, a null. Because filtered probabilities are standard in this literature, forecasting studies built on them warrant particular scrutiny, and the three-stage design here provides a template for quantifying each channel. The bootstrap addresses a complementary channel: statistics computed over overlapping portfolios and windows can carry vanishing naive p-values (here 7×10^{-9}) that dissolve (to 0.40) once dependence is preserved.

Implications And Limitations

For practitioners the results are cautionary: the regime structure is real but does not translate into an implementable edge. Even under the identification design, 10 bp transaction costs cut the fraction of portfolios with positive net improvement from 89% to 61%; under the ex-ante design there is no improvement to implement. For researchers, the attribution diagnostic offers a way to locate regime dependence in multi-factor models — test each factor in isolation and verify that improvements scale with exposure — while the look-ahead decomposition and dependence-robust inference test whether apparent gains survive real-time discipline (15).

Several limitations warrant acknowledgment. First, because the test portfolios are sorted on size, SMB loadings span −0.22 to 1.67 across portfolios and windows, while the market, value, and momentum loadings span only 0.96 to 1.08, −0.04 to 0.78, and −0.16 to 0.02 respectively; the SMB attribution is therefore the best-powered of the four factor tests, and the null attribution results for market, value, and momentum should be read with that asymmetry in mind — portfolio sets sorted on value or momentum would provide more symmetric tests of those channels. Second, the stress-regime parameters are imprecisely estimated in every training window, which bounds how much any real-time filter built on them can know; and in the 1995 and 2000 windows a single Depression-era episode accounts for roughly 40% of all stress months, so the calm-stress difference in those windows rests more heavily on one historical period than it does in the 2010 window (Table 8). Third, the identification results are ex post by construction and make no real-time claim. Fourth, the analysis uses single-factor two-state switching with a fixed information set; richer conditioning information — credit spreads, volatility indices, or macroeconomic indicators — might identify stress states faster than returns alone, and the ex-ante null does not preclude such extensions.

CONCLUSION

Among the four factors of the Carhart model, regime dependence resides in the size channel. Factor attribution across 18 size-sorted portfolios and three evaluation windows locates it in SMB alone, and the regime parameters show why: the size premium is approximately zero in calm regimes and is earned almost entirely in turbulent regimes. Dependence-robust inference supports the parameter-based pillar of this identification result. Because the training windows are nested and two of the three lean on a single Depression-era episode for roughly 40% of stress months, this is best read as one finding replicated in two additional, overlapping checks rather than three independent confirmations: the calm-stress mean difference is positive in all three windows (one-sided $p \leq 0.030$) and exceeds what a no-switching volatility-clustered null manufactures (Monte Carlo $p = 0.015$–0.090, decisively only in the 2010 window). The cross-sectional monotonicity, though descriptively strong, is only marginally distinguishable from zero once the calendar overlap among test periods is respected (pooled $p = 0.12$ at the automatically selected block length, 0.03 at fixed 12-month blocks).

That regime dependence is an identification result, not a forecasting result: under the fully ex-ante design no portfolio-window test is significant, before or after false-discovery-rate adjustment, by either test. The three-stage

decomposition attributes the apparent forecastability of less disciplined designs to two measurable look-ahead channels, with probability timing contributing more than parameter estimation. The apparent forecastability of the size premium is, in this setting, an artifact of look-ahead; its regime dependence is not.

Future research could extend the attribution approach to the Fama-French five-factor model and international markets, test whether conditioning information beyond the return series can identify stress states quickly enough to convert regime structure into ex-ante skill, and estimate a joint multivariate regime-switching specification across all four factors simultaneously rather than the per-factor univariate approach applied here.

ACKNOWLEDGMENTS

The author thanks academic mentors and peers for guidance on model specification and feedback throughout the project.

AI DISCLOSURE STATEMENT

During the preparation of this work, the author used Anthropic's Claude for language editing, coding assistance in implementing and cross-checking the statistical procedures, and manuscript drafting support. Following the use of this tool, the author reviewed, verified, and edited the content as necessary and takes full responsibility for the accuracy, originality, and integrity of the published work.

DATA AND CODE AVAILABILITY

Factor and portfolio data are publicly available from the Kenneth R. French Data Library at mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. Replication code is available at https://github.com/789wethan-wq/ajsr-size-premium-regimes.

CONFLICT OF INTEREST

The author declares no conflicts of interest related to this work.

REFERENCES

1. Fama EF, *et al*. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*. 1993; 33 (1): 3-56. https://doi.org/10.1016/0304-405X(88)90020-7
2. Sharpe WF Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *The Journal of Finance.* 1964; 19 (3): 425-442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x
3. Lintner J The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets: A Reply. *The Review of Economics and Statistics.* 1969; 51 (2): 222-224. https://doi.org/10.2307/1926735
4. Carhart MM On Persistence in Mutual Fund Performance. *The Journal of Finance*. 1997; 52 (1): 57-82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x
5. Asness CS, *et al*. Value and Momentum Everywhere. *The Journal of Finance.* 2013; 68 (3): 929-985. https://doi.org/10.1111/jofi.12021
6. Fama EF, *et al*. Industry costs of equity. *Journal of Financial Economics*. 1997; 43 (2): 153-193. https://doi.org/10.1016/S0304-405X(96)00896-3
7. Campbell JY, *et al*. By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior. *Journal of Political Economy*. 1999; 107 (2): 205-251. https://doi.org/10.1086/250059
8. Lettau M, *et al*. Consumption, Aggregate Wealth, and Expected Stock Returns. *The Journal of Finance*. 2001; 56 (3): 815-849. https://doi.org/10.1111/0022-1082.00347
9. Kelly BT, *et al*. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*. 2019; 134 (3): 501-524. https://doi.org/10.1016/j.jfineco.2019.05.001
10. Hamilton JD A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*. 1989; 57 (2): 357-384. https://doi.org/10.2307/1912559
11. Ang A, *et al*. International Asset Allocation With Regime Shifts. *The Review of Financial Studies*. 2002; 15 (4): 1137-1187. https://doi.org/10.1093/rfs/15.4.1137
12. Guidolin M, *et al*. Asset allocation under multivariate regime switching. *Journal of Economic Dynamics and Control.* 2007; 31 (11): 3503-3544. https://doi.org/10.1016/j.jedc.2006.12.004
13. Ang A, *et al*. Regime Changes and Financial Markets. *Annual Review of Financial Economics*. 2012; 4 (Volume 4, 2012): 313-337. https://doi.org/10.1146/annurev-financial-110311-101808
14. Perez-Quiros G, *et al*. Firm Size and Cyclical Variations in Stock Returns. *The Journal of Finance*. 2000; 55 (3): 1229-1262. https://doi.org/10.1111/0022-1082.00246
15. Welch I, *et al*. A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*. 2008; 21 (4): 1455-1508. https://doi.org/10.1093/rfs/hhm014

16. Clark TE, *et al*. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics.* 2007; 138 (1): 291-311. https://doi.org/10.1016/j.jeconom.2006.05.023
17. Campbell JY, *et al*. Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *The Review of Financial Studies*. 2008; 21 (4): 1509-1531. https://doi.org/10.1093/rfs/hhm055
18. Diebold FX, *et al*. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*. 1995; 13 (3): 253-263. https://doi.org/10.1080/07350015.1995.10524599
19. Harvey D, *et al*. Testing the equality of prediction mean squared errors. *International Journal of Forecasting.* 1997; 13 (2): 281-291. https://doi.org/10.1016/S0169-2070(96)00719-4
20. Benjamini Y, *et al*. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological).* 1995; 57 (1): 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
21. Politis DN, *et al*. Automatic Block-Length Selection for the Dependent Bootstrap. *Econometric Reviews*. 2004; 23 (1): 53-70. https://doi.org/10.1081/ETC-120028836
22. Day N Estimating the components of a mixture of normal distributions. *Biometrika*. 1969; 56 (3): 463-474. https://doi.org/10.1093/biomet/56.3.463
23. Hansen BE The likelihood ratio test under nonstandard conditions: Testing the markov switching model of gnp. *Journal of Applied Econometrics.* 1992; 7 (S1): S61-S82. https://doi.org/10.1002/jae.3950070506
24. Garcia R Asymptotic Null Distribution of the Likelihood Ratio Test in Markov Switching Models. *International Economic Review.* 1998; 39 (3): 763-788. https://doi.org/10.2307/2527399
25. Harvey CR, *et al*. … and the Cross-Section of Expected Returns. *The Review of Financial Studies*. 2016; 29 (1): 5-68. https://doi.org/10.1093/rfs/hhv059
26. Politis DN, *et al*. The Stationary Bootstrap. *Journal of the American Statistical Association*. 1994; 89 (428): 1303-1313. https://doi.org/10.1080/01621459.1994.10476870
27. Kenneth R. French - Data Library. Available from: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html (accessed on 2026-7-23).
28. US Business Cycle Expansions and Contractions. Available from: https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions (accessed on 2026-7-23).