

Agentic Replication of Asset Pricing

Gu, Kelly & Xiu (2020) linear models re-run on Bloomberg data

Bloomberg-Columbia Machine Learning in Finance Conference 2026

Kai Horkan

BQuant Research (CTO Office) · Columbia University
Work performed during summer internship at BloombergMentor: Paulo Hermanny, Senior Quant Researcher
Manager: Songyun Duan, Head of BQuant Research

01 · MOTIVATION

The replication crisis in finance

Hundreds of published factors; few reproducible out of sample.

t > 3.0Multiple-testing adjusted
significance threshold

Harvey, Liu & Zhu (2016)

26% / 58%Return decay out-of-sample
and post-publication

McLean & Pontiff (2016)

~65%of 452 anomalies
fail retesting

Hou, Xue & Zhang (2020)

02 · MY ROLE

Agentic Replication with Human-in-the-Loop

Faithful, not favorable. Every implementation choice documented.

Bloomberg point-in-time data

Model only sees temporally consistent data.

A real out-of-sample test

Different vendor and window. The paper's sample ends 2016. Numbers should not match.

Defensible, auditable baseline

Never tuned for an optimized result.

03 · THE PROCESS

One reusable pipeline, rebuilt for each paper



THE TWO REPLICATION METRICS · GKX (2020) EQ. 21, 19

DESIGN · EQ. 21

POTENTIAL PREDICTORS · OOS R² · EQ. 19

PORTFOLIO · ANNUALIZED SHARPE

$$z_{i,t} = (1, x_t)' \otimes c_{i,t}$$

$$R^2 = 1 - \frac{\sum (r - \hat{y})^2}{\sum r^2}$$

$$SR = \sqrt{12} \frac{\mu_H - L}{\sigma_H - L}$$

118 chars × 9 macro states = 1,062 columns
paper: 94 × 9 + 74 SIC dummies = 920pooled over 776,169 stock-months
vs. a zero forecast, not the meanrank on $\hat{y}$; equal-weight D10 ~ D1
gross 2.17 → net 1.82 at 25 bp/leg

JUDGMENT CALLS WITH DOMAIN KNOWLEDGE

Decision	What we locked
α-selection regime	GKX val-block vs. 2-fold TS-CV — both reported
Standardization	center + scale (the weakest of three)
Loss	L2 — the paper's "OLS" is really Huber
Weighting	EW headline, VW reported alongside

DATA & SCOPE · WHAT EACH REPLICATION RUNS ON

	GKX (2020)	Ours · CRSP	Ours · Bloomberg
Universe	CRSP / Compustat, all US	GKX datashare (same)	Bloomberg JKP · Russell 3000
Potential Predictors	94 × 9 + 74 SIC = 920	920	118 × 9 = 1,062
Train / validate	1957–74 / 1975–86	same	2000–05 expanding / 2006–09 rolling
Out-of-sample	1987–2016	1987–2021	2010–2025 · 191 months
Bloomberg panel: 776,169 stock-months · ~4,060 names/mo · 16 annual refits			

04 · THE HEADLINE RESULT

The GKX pattern replicates on Bloomberg

Method locked on GKX's CRSP/Compustat, re-run on Bloomberg JKP with nothing re-tuned.

+0.39%	2.17	1.82	155 bp	8.7
pooled OOS R ²	gross EW Sharpe	net Sharpe	breakeven cost	FF3 α t-stat
Ridge (TS-CV) · Bloomberg JKP 2010–2025 · decile H–L · net 25 bp/leg · equal-weight (VW net SR 0.18)				

WHAT WAS RUN

1,062-column Kronecker panel · four linear models · 16 yearly refits

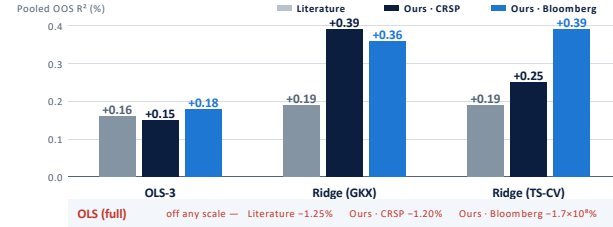
OLS-3 size · value · momentum

Ridge (GKX) L2 · held-out val block

OLS (full) 1,062 cols · no penalty

Ridge (TS-CV) L2 · 2-fold time-series CV

FIGURE 1 · OUT-OF-SAMPLE ACCURACY — POTENTIAL EQUITY PREMIUM PREDICTORS



Ridge ≥ OLS-3 > 0 > OLS (full), both datasets.

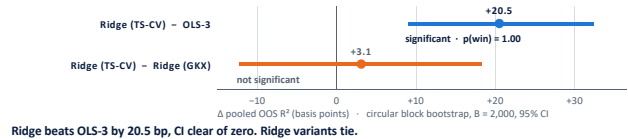
TABLE 1 · GROSS EW DECILE H–L SHARPE

Model	Bloomberg JKP 2010–2025	GKX CRSP 1987–2021	Pooled OOS R ² Bloomberg
Ridge (TS-CV)	2.17	1.75	+0.39%
Ridge (GKX)	1.78	1.64	+0.36%
OLS (full)	1.47	1.49	–1.7×10 ⁶ %
OLS-3	1.00	0.78	+0.18%

Rank order identical in both columns.

- Levels shift with the panel; the ordering survives the swap
- OLS (full) beats OLS-3 on Sharpe with an R² eight orders worse

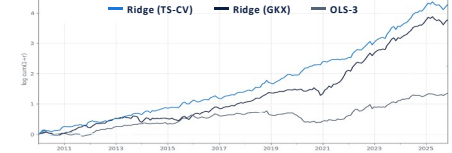
FIGURE 3 · IS THE GAP REAL?



Ridge beats OLS-3 by 20.5 bp, CI clear of zero. Ridge variants tie.

05 · ECONOMIC RELEVANCE

Investable after real costs

FIGURE 4 · NET-OF-COST CUMULATIVE PERFORMANCE
Cumulative log long–short return, net of 25 bp/leg

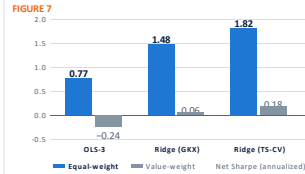
Ridge (TS-CV) leads throughout. All tuned models survive costs.

TABLE 2 · DECILE H–L · EQUAL-WEIGHT · NET OF 25 bp/leg

Model	Net bp/mo	Net SR
OLS-3	77	0.77
Ridge (GKX)	212	1.48
Ridge (TS-CV)	236	1.82
Ridge (TS-CV) H–L +2.82%/mo · Newey–West t = +8.9 (lag 4) · 191 months · ~4,060 names/mo		
Breakeven 155 bp/leg = 6× assumed cost		
FF3 α 288 bp/mo (t 8.7) Max drawdown ~21% Turnover 45.6%/leg/mo		

06 · ROBUSTNESS

Standard checks do not move the result

Alpha sits in the equal-weight book
VW net Sharpe 0.18. Research baseline, not production.

ALSO CHECKED

- 85% buffer: turnover 45.6 → 34.0%/leg/mo, net Sharpe unchanged at 1.82
- Newey–West lag 4, 191 months. a grid interior in all 16 refits

VALIDATED AGAINST THE PAPER · JUDGE COMPOSITE 4.67 / 5 = (D1+D2+D3+2·D4+D5) / 6

FIGURE 6 · OURS vs. an independent implementation

Ridge gross SR · Ours and SX_ML agree within 0.04



Standardization swings SR by +1.05

Center-only 2.83 vs. canonical 1.78. We report the weakest.

07 · CONCLUSION

Impact and Next Steps

Delivered

- Independent replication of the GKX linear models
- Seed-deterministic notebook demo reproduces end-to-end
- Judgment calls documented per decision

Learnings

- Data is the hard part, not the model
- Preprocessing beats model choice: 2 chars moved R² 8×
- Replication broke on join keys, dates, signs

Next

- Manual work from months to days (projected)
- Encode the human judgment into the pipeline
- Hand the agent a research agenda, get a baseline back

Independent Bloomberg replication of the GKX (2020) linear models → a reusable agent for downstream research.

© 2026 Bloomberg Finance L.P. All rights reserved.