A Fresh Look at Return Predictability Using a More Efficient Estimator

Travis L. Johnson
McCombs School of Business, The University of Texas at Austin

I assess time-series return predictability using a weighted least squares estimator that is around 25% more efficient than ordinary least squares (OLS) because it incorporates time-varying volatility into its point estimates. Traditional predictors, such as the dividend yield, perform better in- and out-of-sample when using my estimator, indicating the insignificant OLS estimates may be false negatives driven by a lack of power. Some newer predictors, such as the variance risk premium and the president's political party, are insignificant when using my estimator, indicating the significant OLS estimates may be false positives driven by a few periods with high expected volatility. (*JEL* G10, G11, G12)

Received March 31, 2018; editorial decision September 26, 2018 by Editor Jeffrey Pontiff. Authors have furnished an Internet Appendix and supplementary data and code, which are available on the Oxford University Press Web site next to the link to the final published paper online.

Time-varying market return volatility creates substantial heteroscedasticity in time-series return predictability regressions. Prior literature on predictability typically addresses this heteroscedasticity using ordinary least squares (OLS) with White (1980) heteroscedasticity-consistent standard errors. However, research on the econometrics of predictability regressions (e.g., Singleton 2006; Johannes, Korteweg, and Polson 2014; Westerlund and Narayan 2014) suggests incorporating return heteroscedasticity into point estimates as well as standard errors using the generalized least squares (GLS) insight, resulting in a more efficient estimator that is less noisy and has more power in finite samples.

Thanks to Jeffrey Pontiff (the editor), an anonymous referee, Aydoğan Alt i, Svetlana Bryzgalova, Jonathan Cohn, Andres Donangelo, John Griffin, Michael Halling (Imperial discussant), Yufeng Han (SFA discussant), Michael Johannes (TFF discussant), Bryan Kelly, Arthur Korteweg, Xiang (Nicole) Liu, Zack Liu, Laura Starks, Sheridan Titman, Rasmus Varneskov (AFA discussant), and James Weston and seminar participants at the American Finance Association Meeting, the 11th Annual Hedge Fund Conference at Imperial College, the University of California San Diego, Massachusetts Institute of Technology, the Southern Finance Association Meeting, the Texas Finance Festival, The University of North Carolina at Chapel Hill, and The University of Texas at Austin for their helpful comments. I conducted part of this research while visiting the MIT Sloan School of Management. Data and code are available on my website (http://travislakejohnson.com). Send correspondence to Travis L. Johnson, 2110 Speedway Stop B6600, Austin, TX 78712; telephone: 512-232-6284. E-mail: travis.johnson@mccombs.utexas.edu.

doi:10.1093/rapstu/ray010 Advance Access publication 21 December 2018

Following this suggestion, I assess the predictability afforded by a broad set of variables using an alternative estimator that is more efficient than OLS. The source of these efficiency gains is downweighting observations with low signal-to-noise ratios. For example, in October 2008, the VIX index peaked at 80%, more than four times its median level. At such extremes, realized returns are particularly noisy proxies for expected returns, making the signal-to-noise ratio low and OLS's equal weighting inefficiently high. My estimator, weighted least squares using ex ante variance (WLS-EV), addresses this inefficiency by scaling regression residuals by an estimate of ex ante return volatility, making them comparable in terms of information about expected returns. This weighting represents the standard GLS insight applied to time-series return predictability, a natural setting given the accurate ex ante measures of return volatility and the importance of power in finite samples.[1]

I show that three conclusions about return predictability change when using WLS-EV instead of OLS. First, using WLS-EV strengthens the evidence of predictability for many of the variables studied in Goyal and Welch (2008), including the dividend yield and other theoretically motivated predictors, indicating the insignificant OLS estimates are false negatives stemming from inefficient estimation rather than a failure of return predictability. After adjusting for the Stambaugh (1999) small-sample bias, in-sample WLS-EV estimates indicate 9 of the 16 variables studied in Goyal and Welch (2008) significantly predict next-month returns at the 5% level, compared to only 2 of the 16 for OLS estimates. The in-sample evidence for these predictors is stronger for WLS-EV because it reduces estimation error, and therefore standard errors, while having little impact on point estimates relative to OLS.

One of the criticisms of time-series return predictability made in Goyal and Welch (2008) is that most predictors perform poorly in rolling out-of-sample tests. I show the out-of-sample performance of these predictors substantially improves when using WLS-EV instead of OLS. The improvement is spread across a majority of the predictors and is economically significant, with average out-of-sample R^2 (OOS R^2 hereafter) increasing by 59% and 120% of the average in-sample OLS R^2 for next-month and next-year returns, respectively.

The literature contains other approaches to improving out-of-sample performance that typically generate even higher OOS R^2 than WLS-EV. Examples include imposing economic restrictions on return forecasts (Campbell and Thompson 2008; Pettenuzzo, Timmermann, and Valkanov 2014), allowing for time-varying means (Lettau and Van Nieuwerburgh 2008), and using Bayesian estimates that incorporate estimation risk and time-varying volatility (Johannes, Korteweg, and Polson 2014). Unlike these approaches, WLS-EV is not designed for out-of-sample predictability,

[1] I illustrate WLS-EV's effectiveness in this setting by showing it produces point estimates that are between 25% and 35% less volatile across simulated samples than OLS estimates. See Appendix Appendix B. for details.

though it does outperform OLS. Instead, WLS-EV is designed for inference, providing a more efficient in-sample test of the same "no time-invariant linear predictability" null hypothesis as OLS.

My WLS-EV results indicate four predictors (long-term bond return, term spread, inflation, and the consumption-to-wealth ratio) consistently and significantly predict returns both in-sample and out-of-sample, the Goyal and Welch (2008) standard for evaluating predictability.[2] Three additional predictors (dividend yield, Treasury-bill yield, and payout yield) meet this standard when using simple economic restrictions for out-of-sample tests along with WLS-EV. None of the predictors meet the Goyal and Welch (2008) standard when using OLS. For three other predictors (dividend-to-price ratio, earnings-to-price ratio, and Treasury bond yield), WLS-EV produces evidence for predictability that is stronger than OLS, but not consistently significant across specifications. Both WLS-EV and OLS are pessimistic about the remaining six predictors (dividend-to-earnings ratio, conditional variance, default spread, book-to-market ratio, cross-sectional beta premium, and net equity expansion), with neither approach producing more than fleeting evidence of predictability in my 1927–2015 sample.

In my second application, I show the predictability afforded by proxies for the conditional variance risk premium ($\widehat{\text{VRP}}$) is not robust to the WLS-EV approach. Both Bollerslev, Tauchen, and Zhou (2009) and Drechsler and Yaron (2011) estimate $\widehat{\text{VRP}}$ using the difference between option-implied variance and expected realized variance, and show their measures predict future equity returns in OLS regressions. I show that, regardless of variable construction, forecast horizon, sampling frequency, sample period, or country, WLS-EV estimates of the relation between $\widehat{\text{VRP}}$ and future market returns are not statistically significant.

WLS-EV point estimates of the $\widehat{\text{VRP}}$-return relation are insignificant despite smaller standard errors because they are much closer to zero than their OLS counterparts. This implies the OLS estimates are largely driven by a few observations with extremely positive or negative values of $\widehat{\text{VRP}}$ and high ex ante return volatility. WLS-EV downweights these observations, relying more on the rest of the sample in which volatility is less extreme and $\widehat{\text{VRP}}$ is less predictive of future returns. The combined OLS and WLS-EV results do not indicate statistically significant evidence for a linear relation between the variance risk premium and future equity returns and suggests there may be a nonlinear or time-varying relation that is beyond the scope of both OLS and WLS-EV.

In my third application, I show the WLS-EV approach substantially weakens the surprising predictability afforded by politics, the weather, and—even more puzzlingly—the angle between Mars and Saturn, documented in

[2] See panel A of Table 2 for references to the original research documenting return predictability for each of the Goyal and Welch (2008) variables.

Novy-Marx (2014). All three market return predictors in Novy-Marx (2014) have smaller coefficient estimates and larger *p*-values when using WLS-EV instead of OLS, with two becoming insignificant. Furthermore, WLS-EV estimates indicate the 10 predictors proposed in Novy-Marx (2014) are jointly insignificant.

I examine one of the Novy-Marx (2014) predictors, the party of the United States president, in more detail because it is related to ongoing research connecting stock returns with political preferences and uncertainty. Santa-Clara and Valkanov (2003) documents that average market returns are higher while the president is a Democrat than a Republican, a finding the authors refer to as the "presidential puzzle." Pástor and Veronesi (2017) argues the presidential puzzle is attributable to a correlation between election results and risk premiums, whereby voters elect Democrats during risky time periods with high expected returns.

I provide evidence for an alternative explanation of the presidential puzzle: *un*expected returns in a few time periods with high ex ante volatility happened to be positive under Democrats and negative under Republicans. Consistent with this interpretation, WLS-EV estimates of the relation between the president's party and stock returns are less than half of OLS estimates (4.38% per year vs. 10.20%), and statistically insignificant. I also show that even OLS estimates are insignificant when excluding a small set of observations with extreme ex ante volatility, during which markets had extremely positive average returns under Democrats and negative average returns under Republicans. This pattern is more consistent with unexpected returns than expected returns, which should be positive and moderate. Rather than excluding these observations, WLS-EV downweights them to reflect their noisiness as measures of expected returns, but the conclusion remains that there is no statistically significant evidence of higher returns under Democratic presidents.[3]

For both variance risk premium proxies and the Novy-Marx (2014) predictors, WLS-EV estimates are insignificant, whereas OLS estimates are significant, despite both being asymptotically unbiased tests of the same null hypothesis. One potential reason for this difference is that OLS Newey and West (1987) standard errors are downward biased in small samples with extreme heteroscedasticity. Consistent with this possibility, I find that the variance risk and Novy-Marx (2014) variables often lose significance as a predictor when using OLS combined with *p*-values based on heteroscedastic simulations.

However, small sample biases alone do not explain why WLS-EV coefficients are so much lower, and simulated *p*-values so much higher, than their OLS counterparts. A possible explanation is that some of the predictors were

[3] Powell et al. (2007) also concludes the presidential puzzle is spurious after adjusting standard errors for the extreme persistence of the president's party and moderate autocorrelation in returns.

selected via data mining targeting OLS significance. To illustrate this possibility, I simulate samples under the no predictability null hypothesis and show that when OLS estimates are falsely significant, WLS-EV point estimates are closer to zero on average and statistically insignificant in more than 50% of simulations. Consistent with this data mining interpretation, WLS-EV estimates for the variance risk premium and Novy-Marx (2014) predictors are insignificant despite having smaller confidence intervals than OLS estimates because they are substantially closer to zero. I also show that if data mining targeted WLS-EV significance, OLS estimates would be similarly useful as a partially independent test of the same null hypothesis. However, any data mining in existing literature targeted OLS significance, making WLS-EV a useful diagnostic for revisiting return predictability.

A natural concern about the insignificant WLS-EV estimates of the predictability associated with the variance risk premium and the Novy-Marx (2014) variables is that they downweight the times we may care about most economically, those occurring after a market crash.[4] However, it is important to note WLS-EV downweights these observations *econometrically* and not *economically*. Unlike economically distinct alternatives, ex ante volatile observations have the same linear relation between X_t and $\mathbb{E}(r_{t+1})$ in WLS-EV like in OLS, they are just downweighted econometrically to produce more efficient point estimates. Neither OLS nor WLS-EV address economically distinct alternatives such as time-varying predictability that is stronger after market crashes.

Applying the GLS insight to return predictability regressions using ex ante variance weights is not new. Singleton (2006) discusses the econometric basis for this approach in Section 3.6.2. French, Schwert, and Stambaugh (1987) and Campbell et al. (2018) use this procedure in the context of the risk-return trade-off. The GARCH-in-mean framework estimated in Engle, Lilien, and Robins (1987) and Glosten, Jagannathan, and Runkle (1993), the GARCH-X framework in Brenner, Harjes, and Kroner (1996), and the MIDAS framework in Ghysels, Santa-Clara, and Valkanov (2005) are all structural approaches to incorporating conditional variance in estimating the risk-return trade-off. The stochastic volatility model in Johannes, Korteweg, and Polson (2014) embeds the ex ante variance weighting idea in a structural Bayesian learning framework. Finally, Westerlund and Narayan (2014) derives the asymptotic properties of weighted least squares with potentially misspecified ex ante variance measures. They also implement this approach using structural variance models such as ARCH, use small-sample simulations to illustrate its effectiveness, and apply it in-sample to the Goyal and Welch (2008) predictors.

[4] WLS-EV does not tend to downweight the crashes themselves, because ex ante volatility is often moderate before a crash and spikes upward only after.

Relative to this literature, my contribution is applying the GLS insight to a large set of predictors and showing it changes OLS-based conclusions about return predictability. Overall, significant evidence suggests that the equity risk premium relates to traditional theory-based predictors with long sample periods, such as the dividend yield and the Lettau and Ludvigson (2001) "cay" variable, whereas there is no significant evidence for the predictability afforded by more-recent predictors with weak theoretical grounding or short sample periods, such as the Novy-Marx (2014) predictors, the president's party, or the variance risk premium.

1. Weighted Least Squares with Ex Ante Return Variance

The WLS-EV approach estimates the linear regression:

$$r_{t+1} = X_t \cdot \beta + \epsilon_{t+1}. \tag{1}$$

The returns r_{t+1} can be raw or log returns, can be overlapping or non-overlapping and can be adjusted for the risk-free rate or unadjusted. There can be multiple return predictors along with an optional constant in the X_t vector.

I follow two steps to estimate β in Equation (1) using WLS-EV:

1. Estimate σ_t^2, the conditional variance of next-period unexpected returns ϵ_{t+1}.
2. Estimate $\hat{\beta}_{\text{WLS-EV}}$ using

$$\hat{\beta}_{\text{WLS-EV}} = \text{argmin}_\beta \sum_{t=1}^{T} \left(\frac{r_{t+1} - X_t \cdot \beta}{\hat{\sigma}_t} \right)^2, \tag{2}$$

where $\hat{\sigma}_t$ is the square root of the estimate of σ_t^2 from step 1. This estimator can be implemented using any OLS package by regressing $\frac{r_{t+1}}{\hat{\sigma}_t}$ on $\frac{X_t}{\hat{\sigma}_t}$. Note that, since any constant is in X_t, this OLS implementation has no constant term.

Many different potential approaches can be taken from the literature for estimating σ_t^2, the conditional variance of next-period returns, any of which can be used to estimate WLS-EV. I describe my approach in Section 1.1.

Standard errors for WLS-EV are the same as OLS standard errors when regressing the scaled returns $\frac{r_{t+1}}{\hat{\sigma}_t}$ on the scaled constant and regressors $\frac{X_t}{\hat{\sigma}_t}$. These standard errors should be heteroscedasticity and autocorrelation consistent (HAC) to adjust for any remaining heteroscedasticity and autocorrelation. I use the Newey and West (1987) procedure and a simulation approach described in Appendix Appendix B. as two alternative HAC standard errors.

A natural concern is that my standard errors are downward biased, because they do not adjust for estimation error in the first-stage variance prediction regression. I address this concern by using heteroscedasticity consistent standard errors in the second-stage return prediction regression, which would not be necessary with perfectly specified variance measures. With misspecified measures, heteroscedasticity remains in the second stage or is even exacerbated. Fortunately, Wooldridge (2010), Westerlund and Narayan (2014), and Romano and Wolf (2017) show that the typical heteroscedasticity consistent (HC) standard errors remain asymptotically consistent when applied to WLS, even in the presence of first-stage estimation errors. Romano and Wolf (2017) therefore conclude "combining WLS with HC standard errors allows for valid inference, even if the conditional variance model is misspecified."

My simulated standard errors address the possibility of estimation error in my ex ante variance measures more directly by using simulations in which the $\hat{\sigma}_t^2$ used for WLS-EV differ from true return variance, as detailed in Appendix Appendix B.

1.1 Estimating $\hat{\sigma}_t^2$

I use two different sampling frequencies, monthly for predictors with longer sample periods and daily for the variance risk premium. My approach to handling overlapping regressions, described in Section 1.2, assures I only need variance forecasts for one sample period ahead even when the return forecasting horizon spans multiple sampling periods. I therefore need a next-month variance forecast $\hat{\sigma}_m^2$ and a next-day variance forecast $\hat{\sigma}_d^2$.[5]

For monthly sampling frequencies, the left-hand side of my first-stage regressions is RV_{m+1}, the sum of squared daily log market returns in month $m+1$. I use different combinations of $RV_{m-s,m}$, realized variance in months $m-s$ through m, as potential variance predictors on the right-hand side of my first-stage regressions.

Panel A of Table 1 shows that past realized variance strongly predicts future realized variance, with R^2 between 25% and 40%, suggesting WLS-EV could provide substantial efficiency gains relative to OLS. I use Column (5) of panel A, which includes all potential predictors, to guide my choice of specification for $\hat{\sigma}_m^2$. Columns (5) and (6) show only prior-month and prior-year realized variance are statistically significant predictors, and together they provide nearly all the predictability afforded by the four lags of realized variance. For this reason, I use the more parsimonious specification in

[5] Throughout the paper I use subscript m to denote monthly observations, d for daily observations, and t for generic time periods.

Table 1
Effectiveness of ex ante variance proxies

	(1)	(2)	(3)	(4)	(5)	(6)
A. Predicting next-month variance RV_{m+1}						
Const ($\times 10^2$)	10.22***	8.63***	7.07***	5.70**	4.76***	4.73***
	(2.38)	(2.13)	(2.05)	(2.28)	(1.72)	(1.72)
RV_m	0.60***				0.46***	0.46***
	(0.09)				(0.14)	(0.10)
$\mathrm{RV}_{m-2,m}$		0.66***			–0.03	
		(0.09)			(0.13)	
$\mathrm{RV}_{m-5,m}$			0.72***		0.08	
			(0.09)		(0.10)	
$\mathrm{RV}_{m-11,m}$				0.78***	0.31**	0.36***
				(0.12)	(0.14)	(0.11)
Adj. R^2 (%)	35.9	30.2	28.2	26.3	39.3	39.4
						$\rightarrow$RV $\hat{\sigma}^2_m$
B. Predicting next-day variance FutRV_{d+1}						
Const ($\times 10^2$)	0.34***	0.55***	0.17	–0.59***	–0.29***	
	(0.08)	(0.07)	(0.13)	(0.18)	(0.07)	
$\mathrm{RV}d-20,d$	0.80***				0.40	
	(0.13)				(0.27)	
$\mathrm{RV}d-251,d$	–0.05				–0.31***	
	(0.07)				(0.09)	
FutRV_d		0.58***			0.18***	0.24***
		(0.05)			(0.05)	(0.04)
$\mathrm{FutRV}_{d-20,d}$			0.86***		–0.49*	
			(0.12)		(0.27)	
$(\mathrm{VIX}^2_d)/252$				1.04***	1.03***	0.67***
				(0.12)	(0.16)	(0.07)
Adj. R^2 (%)	37.3	34.0	36.7	46.2	50.1	47.1
	$\rightarrow$RV $\hat{\sigma}^2_d$					$\rightarrow$VIXF $\hat{\sigma}^2_d$

This table presents regressions of realized return variance on potential ex ante variance predictors. For each month m in panel A, the left-hand side is RV_{m+1}, the realized variance in month $m+1$, where $\mathrm{RV}m = \sum_{d\in m} r^2_d$, and r_d is the log dividend-inclusive excess return of the CRSP value-weighted index on day d. Predictors in panel A are $\mathrm{RV}_{m-a,m} = \frac{1}{a+1}\sum_{s=0}^{a} \mathrm{RV}_{m-s}$. For each day d in panel B, the left-hand side is FutRV_{d+1}, the realized variance on day $d+1$, where $\mathrm{FutRV}_d = \sum_{i\in d} r^2_{i,\mathrm{fut.}}$ and $r_{i,\mathrm{fut.}}$ is the log return of the front-maturity S&P 500 futures contract in 5-minute interval i. Predictors in panel B are $\mathrm{RV}_{d-a,d} = \frac{1}{a+1}\sum_{s=0}^{a} r^2_{d-s}$, $\mathrm{FutRV}_{d-a,d} = \frac{1}{a+1}\sum_{s=0}^{a} \mathrm{FutRV}_{d-s}$, and VIX^2_d, the square of the VIX index on day d. The sample is 1,062 monthly observations from 1927 to 2015 in panel A and 6,552 daily observations from 1990 to 2015 in panel B. Standard errors are in parentheses and are computed using Newey and West (1987) with 12 lags. *$p < .1$; **$p < .05$; ***$p < .01$.

Column (6) to produce fitted values for my monthly ex ante variance proxy, which I refer to as RV $\hat{\sigma}^2_m$ hereafter:

$$\mathrm{RV}\ \hat{\sigma}^2_m \equiv \hat{a}_m + \hat{b}_m \cdot \mathrm{RV}_m + \hat{c}_m \cdot \mathrm{RV}_{m-11,m}, \tag{3}$$

where $\hat{a}_m$, $\hat{b}_m$, and $\hat{c}_m$ are the estimated coefficients in a regression of RV_{m+1} on a constant, RV_m, and $RV_{m-11,m}$. When using the full 1927–2015 sample, these coefficients are those in Column (6) of panel A in Table 1.

I use daily sampling in my analysis of the variance risk premium as a predictor, which is only available starting in 1990. Therefore, I can use intra-day futures return variance, FutRV_{d+1}, instead of squared daily returns as

the dependent variable in my first-stage regression. To remain consistent with the rest of the paper, my primary $\hat{\sigma}_d^2$ uses past-month and past-year realized variances from daily returns as independent variables in the first-stage regression. I also consider an alternative measure using past FutRV and the VIX as variance predictors.

Panel B of Table 1 shows the results of various possible first-stage regressions for predicting FutRV_{d+1} in a 1990–2015 sample. Column (1) shows that the same predictors used for RV $\hat{\sigma}_m^2$ also strongly predict next-day next-day variance, and with a similar 37.3% R^2. Unsurprisingly, next-day variance is primarily related to last-month variance rather than last-year variance. I use fitted values from Column (1) as my main daily ex ante variance proxy, which I refer to as RV $\hat{\sigma}_d^2$ hereafter:

$$\text{RV } \hat{\sigma}_d^2 \equiv \hat{a}_d + \hat{b}_d \cdot \text{RV}_{d-20,d} + \hat{c}_d \cdot \text{RV}_{d-251,d}, \tag{4}$$

where $\hat{a}_d$, $\hat{b}_d$, and $\hat{c}_d$ are the estimated coefficients in a regression of FutRV_{d+1} on a constant, $\text{RV}_{d-20,d}$, and $RV_{d-251,d}$, as presented in Column (1) of panel B in Table 1.

Column (4) shows that the VIX index by itself is an extremely good predictor, achieving a higher R^2 than any of the approaches based on past realized variances alone. Column (5) indicates that only realized variance on day d, FutRV_d, incrementally and positively predicts next-day variance relative to the VIX. For this reason, I use the more parsimonious specification in Column (6) to produce fitted values for my alternative ex ante variance proxy, which I refer to as VIXF $\hat{\sigma}_d^2$ hereafter:

$$\text{VIXF}\hat{\sigma}_d^2 = \hat{b}_v \cdot \text{FutRV}_d + \hat{c}_v \cdot \frac{\text{VIX}_d^2}{252}, \tag{5}$$

where $\hat{b}_v$ and $\hat{c}_v$ are the estimated coefficients in a regression of FutRV_{d+1} on FutRV_d and $\frac{\text{VIX}_d^2}{252}$. I omit a constant from this first-stage regression to avoid negative fitted values for next-day variance. I use VIXF $\hat{\sigma}_d^2$ to illustrate the robustness of WLS-EV to alternative $\hat{\sigma}_d^2$ and the potential efficiency gains resulting from better ex ante variance proxies.

Figure 1 plots RV $\hat{\sigma}_m$ for 1927–2015, and both RV $\hat{\sigma}_d$ and VIXF $\hat{\sigma}_d$ for 1990–2015, all of which are displayed as annualized standard deviations. Like other conditional volatility estimates, RV $\hat{\sigma}_m$ is small and steady in normal times but spikes upward during market downturns, particularly in 1929, 1987, and 2008. These episodes have conditional return volatility higher than 50%, approximately 3 times the typical values, which are between 15% and 20%. The more-recent daily sample shows similar patterns but with even more extreme values during the 2008 crisis. They also show that RV $\hat{\sigma}_d$ and VIXF $\hat{\sigma}_d$ are highly correlated (83%, result untabulted), with VIXF $\hat{\sigma}_d$ being slightly more volatile because it explains more variation in FutRV_{d+1}.

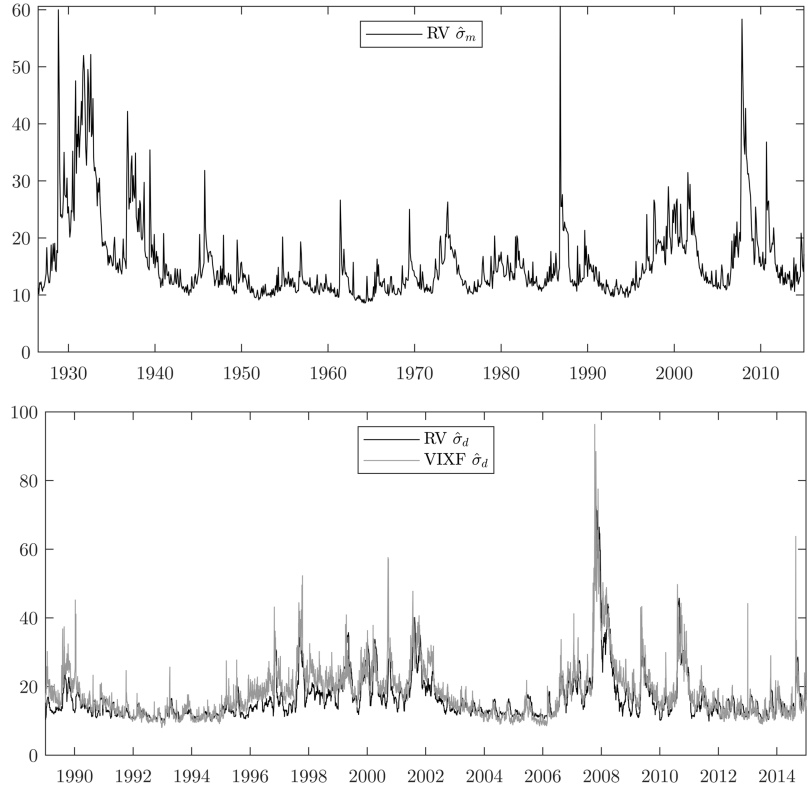

Figure 1
Conditional volatility measures
The first plot presents RV $\hat{\sigma}_m$, the estimates of the volatility of next-month equity market returns conditional on past realized variance, estimated using regressions described in Section 1. The second plot presents RV $\hat{\sigma}_d$ and VIXF $\hat{\sigma}_d$, the estimates of the volatility of next-day equity market returns conditional on past realized variance, and the VIX and past intraday futures variance, respectively. All volatilities are displayed as an annualized percentage. The monthly sample consists of 1,062 observations from 1927 to 2015 and the daily sample of 6,552 observations from 1990 to 2015.

While my ex ante variance proxies are effective empirically, other proxies may predict realized variance as well or even better. As discussed above, any of these can be used with WLS-EV as long as they are constructed from ex ante information. Fortunately, these proxies are strongly correlated with each other, and in untabulated tests, I find my results are not sensitive to using other predictors from Table 1, MIDAS estimates following Ghysels, Santa-Clara, and Valkanov (2005), or RV_m without a first-stage regression.

1.2 Overlapping returns

To maximize power in relatively short samples, many return predictability studies use sampling frequencies shorter than their forecast horizon h,

resulting in overlapping returns. The standard approach in this case is to estimate $\hat{\beta}$ using OLS and adjust the standard errors using the procedures suggested in Newey and West (1987) or Hodrick (1992).

The WLS-EV equivalent of the standard approach would be to estimate $\hat{\beta}$ using least squares weighted by conditional next-h period variance to account for heteroscedasticity, and standard errors from Newey and West (1987) or simulations to account for overlap-driven autocorrelation. This approach suffers from two problems. The first is that conditional next-h period variance measures do not predict realized variance as well as conditional next-period variance measures, reducing WLS-EV's efficiency gains. The second is that the small-sample bias in Newey and West (1987) standard errors for overlapping return regressions, documented in Hodrick (1992) and Ang and Bekaert (2006), applies here and cannot be addressed using Hodrick (1992) standard errors because they have no natural translation to weighted regressions. In Online Appendix E, I use simulations to illustrate these shortcomings and show that the paper's conclusions are nevertheless unchanged when using overlapping regressions with next-h period variance weights.

To apply WLS-EV with overlapping observations while avoiding the problems associated with next-h period ex ante variances, I instead rely on the insight in Hodrick (1992) that overlapping return predictability regressions can be mapped to equivalent (after scaling) nonoverlapping regressions of r_{t+1} on $\bar{X}_t \equiv \sum_{s=0}^{h-1} X_t$, as detailed in Appendix A.1.

I use this insight to estimate $\hat{\beta}$ in overlapping samples using OLS or WLS-EV as follows:

1. Estimate the nonoverlapping regression $r_{t+1} = (\sum_{s=0}^{h-1} X_{t-s}) \cdot \beta + \epsilon_{t+1}$ using either OLS or WLS-EV. Use Newey and West (1987) standard errors to adjust for remaining heteroscedasticity or autocorrelation, for example, because of omitted predictor variables.
2. Scale the resultant coefficients and standard errors by $\mathbb{E}_T(X_t'X_t)^{-1}\mathbb{E}_T(\bar{X}_t'\bar{X}_t)$, which simplifies to $\frac{\mathrm{Var}_T(\bar{x}_t)}{\mathrm{Var}_T(x_t)}$ when X_t has a constant and univariate predictor x_t.

Appendix A.1 shows this approach produces OLS estimates approximately equal to those from a standard overlapping regression in finite samples, and identical asymptotically.

2. Traditional Predictors

My first application of the WLS-EV estimator is to reassess the return predictability afforded by the 16 variables studied in Goyal and Welch (2008). Overall, I find the WLS-EV evidence for return predictability is substantially stronger than the insignificant OLS evidence documented in Goyal and Welch (2008).

As summarized by panel A of Table 2, the 16 predictors Goyal and Welch (2008) and I study are the log dividend-to-current-price ratio (dp), log

Table 2
In-sample return predictability

A. List of Goyal and Welch (2008) predictors

Name	Summary	First paper(s) to document predictability
dp	log dividend-to-current-price ratio	Rozeff 1984; Shiller, Fischer, and Friedman 1984
dy	log dividend-to-lagged-price ratio	Rozeff 1984; Shiller, Fischer, and Friedman 1984
ep	log earnings-to-price ratio	Shiller, Fischer, and Friedman 1984
de	log dividend-to-earnings ratio	Lamont 1998
$\hat{\sigma}^2_m$	Conditional variance	French, Schwert, and Stambaugh 1987; Campbell 1987
tbl	Treasury-bill yield	Fama and Schwert 1977
lty	Treasury bond yield	Campbell 1987; Fama and French 1989
ltr	Treasury bond return	Campbell 1987; Fama and French 1989
tms	Term spread	Campbell 1987; Fama and French 1989
dfy	Default spread	Keim and Stambaugh 1986
infl	Inflation rate	Lintner 1975
bm	log book-to-market ratio	Kothari and Shanken 1997; Pontiff and Schall 1998
csp	Cross-sectional beta premium	Polk, Thompson, and Vuolteenaho 2006
ntis	Net equity expansion	Boudoukh et al. 2007
lpy	log payout yield	Boudoukh et al. 2007
cay	Consumption-to-wealth ratio	Lettau and Ludvigson 2001

B. Predicting next-month returns

	OLS	WLS-EV	OLS	WLS-EV	OLS	WLS-EV	OLS	WLS-EV
Predictor:	dp		dy		ep		de	
Stambaugh $\hat{b}_{adj}$	0.15	0.20	0.64	0.60**	0.56	0.51	–0.26	–0.06
Unadjusted b	0.54	0.59	0.68	0.63	0.80	0.75	–0.23	–0.02
SE (Asy)	(0.48)	(0.33)	(0.50)	(0.33)	(0.48)	(0.37)	(0.84)	(0.57)
p-value (Asy %)	74.9	54.3	19.6	7.3	25.0	16.9	75.4	92.2
SE (Sim)	(0.43)	(0.29)	(0.42)	(0.29)	(0.43)	(0.32)	(0.74)	(0.49)
p-value (Sim %)	71.9	48.8	13.0	4.1	19.5	11.4	72.1	91.0
Predictor:	RV $\hat{\sigma}^2_m$		tbl		lty		ltr	
Stambaugh $\hat{b}_{adj}$	–0.71	–0.60	–0.09*	–0.12***	–0.08	–0.10**	0.12	0.18***
Unadjusted b	–0.69	–0.58	–0.09	–0.12	–0.07	–0.09	0.12	0.18
SE (Asy)	(1.18)	(1.03)	(0.05)	(0.05)	(0.06)	(0.05)	(0.06)	(0.06)
p-value (Asy %)	54.7	55.7	7.6	1.5	18.7	5.9	3.5	0.2
SE (Sim)	(1.19)	(0.94)	(0.05)	(0.04)	(0.05)	(0.04)	(0.07)	(0.05)
p-value (Sim %)	55.2	52.2	5.6	0.3	12.2	1.8	11.9	0.1
Predictor:	tms		dfy		infl		bm	
Stambaugh $\hat{b}_{adj}$	0.20	0.19**	0.13	–0.01	–0.36	–1.00***	1.03	0.19
Unadjusted b	0.20	0.18	0.16	0.03	–0.36	–1.00	1.42	0.58
SE (Asy)	(0.12)	(0.11)	(0.56)	(0.39)	(0.44)	(0.29)	(0.86)	(0.60)
p-value (Asy %)	10.9	8.7	81.7	98.6	41.2	0.1	22.7	74.6
SE (Sim)	(0.13)	(0.09)	(0.43)	(0.27)	(0.45)	(0.29)	(0.86)	(0.52)
p-value (Sim %)	11.6	4.7	76.8	98.0	41.4	0.1	23.3	70.8
Predictor:	csp		ntis		lpy		cay	
Stambaugh $\hat{b}_{adj}$	2.12***	1.85***	–0.16	–0.12*	1.65*	1.56**	0.19**	0.22***
Unadjusted b	2.14	1.87	–0.16	–0.12	1.78	1.70	0.20	0.23
SE (Asy)	(0.68)	(0.66)	(0.09)	(0.08)	(0.85)	(0.76)	(0.10)	(0.09)
p-value (Asy %)	0.2	0.5	8.1	12.1	5.1	3.9	5.6	1.1
SE (Sim)	(0.69)	(0.55)	(0.10)	(0.07)	(0.96)	(0.71)	(0.08)	(0.07)
p-value (Sim %)	0.2	0.1	10.1	8.8	8.5	2.7	2.3	0.2

Summary statistics

	# significant (Sim)				
	10%	5%	1%	Mean$\left(\frac{\text{WLS SE (Sim)}}{\text{OLS SE (Sim)}}\right)$:	# (WLS *p*-val (Sim) < OLS):
OLS	4	2	1	73.2	13
WLS-EV	10	9	5		

C. Predicting next-year returns

	OLS	WLS-EV	OLS	WLS-EV	OLS	WLS-EV	OLS	WLS-EV
Predictor:	dp		dy		ep		de	
Stambaugh $\hat{b}_{\text{adj}}$	3.17	3.19	7.60	7.51**	6.54	6.43*	–0.39	–0.83
Unadjusted $\hat{b}$	7.69	7.71	8.01	7.91	9.30	9.19	0.08	–0.35
SE (Asy)	(5.37)	(4.01)	(5.27)	(4.00)	(4.91)	(3.99)	(9.03)	(5.97)
p-value (Asy %)	55.5	42.6	14.9	6.1	18.3	10.7	96.5	88.9
SE (Sim)	(5.08)	(3.42)	(5.05)	(3.43)	(4.32)	(3.52)	(7.84)	(5.34)
p-value (Sim %)	53.3	35.1	13.0	2.8	13.0	6.7	96.0	87.8
Predictor:	RV $\hat{\sigma}_m^2$		tbl		lty		ltr	
Stambaugh $\hat{b}_{\text{adj}}$	0.22	–5.72	–0.87	–1.05**	–0.46	–0.73	0.69***	0.63***
Unadjusted $\hat{b}$	0.42	–5.52	–0.82	–1.00	–0.35	–0.61	0.68	0.62
SE (Asy)	(9.39)	(8.04)	(0.68)	(0.62)	(0.73)	(0.66)	(0.19)	(0.17)
p-value (Asy %)	98.1	47.7	20.0	9.2	52.3	26.6	0.0	0.0
SE (Sim)	(9.30)	(6.87)	(0.57)	(0.48)	(0.58)	(0.50)	(0.20)	(0.17)
p-value (Sim %)	98.1	40.5	12.7	2.9	42.7	14.6	0.0	0.0
Predictor:	tms		dfy		infl		bm	
Stambaugh $\hat{b}_{\text{adj}}$	3.03**	2.41**	2.00	–0.09	–1.78	–5.03**	16.34*	6.80
Unadjusted $\hat{b}$	3.02	2.40	2.38	0.29	–1.78	–5.03	20.70	11.16
SE (Asy)	(1.36)	(1.17)	(4.81)	(3.87)	(5.44)	(3.55)	(8.96)	(7.10)
p-value (Asy %)	2.6	4.0	67.8	98.1	74.3	15.7	6.8	33.8
SE (Sim)	(1.38)	(1.03)	(4.77)	(2.84)	(3.97)	(2.45)	(9.67)	(6.01)
p-value (Sim %)	2.9	1.9	67.4	97.4	65.5	4.0	9.1	25.8
Predictor:	csp		ntis		lpy		cay	
Stambaugh $\hat{b}_{\text{adj}}$	5.79	3.29	–2.52**	–1.66**	28.05***	22.65***	1.90*	2.05**
Unadjusted $\hat{b}$	5.98	3.47	–2.50	–1.64	29.61	24.20	1.96	2.11
SE (Asy)	(7.45)	(7.65)	(1.08)	(0.92)	(9.22)	(9.22)	(1.14)	(1.01)
p-value (Asy %)	43.7	66.8	2.0	7.0	0.2	1.4	9.7	4.3
SE (Sim)	(7.40)	(6.24)	(1.08)	(0.80)	(10.33)	(7.80)	(1.04)	(0.87)
p-value (Sim %)	43.5	60.0	1.9	3.8	0.7	0.4	6.6	1.8

Summary statistics

	# significant (Sim.)				
	10%	5%	1%	Mean$\left(\frac{\text{WLS SE (Sim)}}{\text{OLS SE (Sim)}}\right)$:	# (WLS *p*-val (Sim) < OLS):
OLS	6	4	2	74.4	12
WLS-EV	9	8	2		

This table presents estimates of in-sample return predictability regressions of the form:

$$r_{m+1,m+h} = a + b \cdot x_m + \epsilon_{m+h},$$

where $r_{m+1,m+h}$ is the log dividend-inclusive excess return of the CRSP value-weighted index from months $m+1$ through $m+h$, and x_m is a candidate return predictor. The predictors x_m are summarized in panel A. The forecast horizons are $h = 1$ month in panel B and $h = 12$ months in panel C. To improve the readability of the coefficients, I divide dp, ep, de, bm, and lpy by 100. For each predictor, I estimate b using OLS and WLS-EV, detailed in Section 1, using RV $\hat{\sigma}_m^2$. I also adjust b for the Stambaugh bias using a simulation procedure. I compute asymptotic (Asy) errors and p-values for the bias-adjusted coefficients by mapping to equivalent nonoverlapping regressions, as described in Section 1, and then using Newey and West (1987) with 12 lags and the simulated (Sim) standard errors and p-values, using the procedure described in Appendix Appendix B. The sample is 1,062 monthly observations from 1927 to 2015. *$p < .1$; **$p < .05$; ***$p < .01$.

dividend-to-lagged-price ratio (dy), log earnings-to-price (ep) ratio, log dividend-to-earnings (de) ratio, conditional variance of returns (RV $\hat{\sigma}_m^2$), Treasury-bill yield (tbl), long-term Treasury bond yield (lty), return of long-term bonds (ltr), term spread (tms), default yield spread (dfy), inflation (infl), log book-to-market (bm) ratio, cross-sectional beta premium (csp), net equity expansion (ntis), log net payout yield (lpy), and consumption-to-wealth ratio (cay). Panel A of Table 2 references the original papers documenting each variable's relation with future market returns.

To improve the readability of the coefficients, I divide dp, dy, ep, de, bm, and lpy by 100. I compute RV $\hat{\sigma}_m^2$ as described above and retrieve lpy from Michael Roberts' website, cay from Martin Lettau's website, and the remaining predictors from Amit Goyal's website. Detailed definitions of the predictors are in Boudoukh et al. (2007) for lpy, Lettau and Ludvigson (2001) for cay, and Goyal and Welch (2008) for the remaining predictors.

2.1 In-sample predictability

For each of the 16 predictors, I estimate univariate predictability regressions of the form:

$$r_{m+1,m+h} = a + b \cdot x_m + \epsilon_{m+1,m+h}, \tag{6}$$

where $r_{m+1,m+h}$ is the log excess return of the CRSP value-weighted index in months $m+1$ through $m+h$. I use both OLS and WLS-EV to estimate the coefficients a and b. I assess next-month ($h = 1$) and next-year ($h = 12$) predictability and adjust for the overlap when $h = 12$ using the procedure in Section 1.2. I also compute simulated standard errors and adjust for the Stambaugh (1999) bias using procedures described in Appendix Appendix B.

The results of my in-sample tests are in Table 2, beginning with a 1-month forecast horizon ($h = 1$) in panel B. The WLS-EV estimates have simulated standard errors smaller than their OLS counterparts by an average of 26.8%, indicating WLS-EV results in substantial efficiency gains relative to OLS.

Furthermore, the WLS-EV point estimates are consistent with the OLS point estimates in most cases, and substantially larger for tbl, lty, ltr, and infl. Combining these features strengthens the overall in-sample evidence of return predictability, with WLS-EV p-values smaller than OLS p-values for 13 of the 16 predictors. Using 10%, 5%, and 1% critical values, WLS-EV estimates are statistically significant for 10, 9, and 5 of the predictors, respectively, compared to only 4, 2, and 1 for OLS.

I assess the predictive power of these 16 variables for next-year returns ($h = 12$) in panel C of Table 2. The results are consistent with the next-month return results in panel A, indicating stronger in-sample evidence of return predictability. The WLS-EV approach yields 25.6% smaller simulated standard errors and largely unchanged point estimates, making the WLS-EV evidence for return predictability stronger than the OLS evidence for 12 of the 16 predictors. Using 10%, 5%, and 1% critical values, WLS-EV estimates are statistically significant for 9, 8, and 2 predictors, respectively, compared to 6, 4, and 2 for OLS.

2.2 Out-of-sample predictability

The evidence supporting return predictability in Table 2 has two potential concerns. The first is data mining: the predictive variables are not chosen at random but instead are selected by the literature, among many potential predictors, based on their in-sample OLS statistical significance. The second concern is a bias in the standard errors not captured by the asymptotic HAC or simulated standard errors I use to test the no-predictability null hypothesis.

To address these concerns, I examine the out-of-sample predictive power of these regressors using both OLS and WLS-EV. As discussed in Goyal and Welch (2008), out-of-sample tests provide an additional falsifiable implication of the no-predictability null hypothesis that was not itself the target of data mining in most return predictability research. While some debate (e.g., Cochrane 2008 or Campbell and Thompson 2008) swirls about the power of out-of-sample tests for rejecting the null, making a failure to reject difficult to interpret, out-of-sample success provides strong evidence of predictability, because it cannot be explained by in-sample data mining or biased standard errors.

In addition to providing researchers with an alternative test of the no-predictability null, out-of-sample predictability provides a simple measure of the practical value a predictor offers to investors. As discussed in Campbell and Thompson (2008), Johannes, Korteweg, and Polson (2014), and elsewhere, investors may use more sophisticated techniques in forming expectations about future market returns and their portfolios. Nevertheless, out-of-sample R^2 provides a simple indicator of which predictors would have benefited investors if used in "real-time" over the past century.

I compute the OOS R^2 for each predictor using a procedure very similar to the one in Goyal and Welch (2008). Specifically, for each month τ in my 1927–2015 sample,[6] starting 20 years after the first month the predictor is available, I compute the conditional expected future return over the next h months, $\mathbb{E}_\tau(r_{\tau+1,\tau+h}|x_\tau)$ as follows.

1. Estimate coefficients $\hat{a}_\tau$ and $\hat{b}_\tau$ in the regression:

$$r_{m+1,m+h} = a_\tau + b_\tau \cdot x_m + \epsilon_{m+1,m+h}, \tag{7}$$

   using OLS or WLS-EV, and only data available as of τ, that is, $m \leq \tau - h$. To maximize power, I use overlapping monthly regressions, instead of the annual regressions used in Goyal and Welch (2008). For WLS-EV, I reestimate the first-stage variance prediction regression for each τ:

$$RV_{m+1} = c_\tau + d_\tau \cdot RV_m + e_\tau \cdot RV_{m-11,m} + \gamma_{m+1}, \tag{8}$$

   using only data available as of τ.
2. Use estimated coefficients and current predictor values to compute:

$$\mathbb{E}_\tau(r_{\tau+1,\tau+h}|x_\tau) \equiv \hat{a}_\tau + \hat{b}_\tau x_\tau. \tag{9}$$

As a benchmark, I also compute an out-of-sample return prediction ignoring x_τ:

$$\mathbb{E}_\tau(r_{\tau+1,\tau+h}) \equiv \hat{\mu}_\tau, \tag{10}$$

where $\hat{\mu}_\tau$ is the coefficient in a regression of future returns on only a constant (Equation (7) restricted so $b_\tau = 0$).

Given time series of out-of-sample return predictions $\mathbb{E}_\tau(r_{\tau+1,\tau+h}|x_\tau)$ and $\mathbb{E}_\tau(r_{\tau+1,\tau+h})$, I compute the out-of-sample R^2 and adjusted R^2 like in Goyal and Welch (2008):

$$R^2 \equiv 1 - \frac{\mathrm{MSE}_A}{\mathrm{MSE}_N}, \quad \mathrm{Adj.R}^2 \equiv \mathrm{R}^2 - (1 - \mathrm{R}^2)\frac{\mathrm{K}}{\mathrm{T} - \mathrm{K} - 1}, \tag{11}$$

$$\mathrm{MSE}_A \equiv \frac{1}{T}\sum_{\tau=1}^{T} e_A(\tau, x)^2 \quad \mathrm{MSE}_N \equiv \frac{1}{T}\sum_{\tau=1}^{T} e_N(\tau)^2 \tag{12}$$

$$e_A(\tau, x) \equiv r_{\tau+1,\tau+h} - \mathbb{E}_\tau(r_{\tau+1,\tau+h}|x_\tau) \quad e_N(\tau) \equiv r_{\tau+1,\tau+h} - \mathbb{E}_\tau(r_{\tau+1,\tau+h}), \tag{13}$$

[6] Unlike that of Goyal and Welch (2008), my sample starts in 1927, because I require daily return data to compute the RV $\hat{\sigma}^2_m$, and ends in 2015, rather than 2005. The exceptions are csp (available 1937–2002), lpy (available 1927–2010), and caya (the ex ante version of cay, available 1952–2013).

Table 3
Out-of-sample return predictability

A. Predicting next-month returns

	OLS	WLS-EV	OLS	WLS-EV	OLS	WLS-EV	OLS	WLS-EV
Predictor:	dp		dy		ep		de	
OOS R^2 (%)	–0.01	0.13*	–0.28	0.27**	–0.94	–0.11	–1.08	–1.30
CT R^2 (%)	0.16**	0.33**	0.19**	0.47**	–0.16	0.25**	0.00	–0.16
PTV R^2 (%)	0.25**	0.24**	0.41**	0.38**	0.49**	0.33**	–0.23	–0.28
Predictor:	RV$t-11,t$		tbl		lty		ltr	
OOS R^2 (%)	–0.11	–1.24	–0.05*	–0.42	–0.86	–1.23	–0.46	0.25**
CT R^2 (%)	0.00	0.00	0.20**	0.28**	0.18**	0.29**	0.28**	0.13*
PTV R^2 (%)	–0.05	–0.15	0.59***	0.64***	0.46**	0.55***	0.19*	0.57**
Predictor:	tms		dfy		infl		bm	
OOS R^2	0.21**	0.37**	–0.16	–0.42	0.15	0.59**	–1.36	–0.04
CT R^2 (%)	0.21**	0.45***	–0.15	–0.03	0.17*	0.70***	–0.82	–0.04
PTV R^2 (%)	0.40**	0.46***	–0.15	–0.06	0.10	0.12*	0.04*	0.11*
Predictor:	csp		ntis		lpy		caya	
OOS R^2 (%)	–0.49	–0.04	–0.71	–0.47	–0.62	–0.08	0.15*	0.29**
CT R^2 (%)	0.54**	0.42**	–0.70	–0.47	0.02	0.24**	–0.06	0.09*
PTV R^2 (%)	0.60**	0.56**	0.02	–0.02	0.24**	0.32**	0.50***	0.53***

Summary statistics

	IS R^2	OOS R^2				CT R^2				PTV R^2			
	Mean (%)	Mean (%)	#>0	# sig. 10%	5%	Mean (%)	#>0	# sig. 10%	5%	Mean (%)	#>0	# sig. 10%	5%
OLS	0.34	–0.41	3	3	1	0.00	9	8	7	0.24	13	11	9
WLS-EV	0.24	–0.21	6	6	5	0.18	11	11	9	0.27	12	12	10
Diff	–0.09	0.20	11	4	0	0.18	12	2	2	0.03	9	0	0

B. Predicting next-year returns

	OLS	WLS-EV	OLS	WLS-EV	OLS	WLS-EV	OLS	WLS-EV
Predictor:	dp		dy		ep		de	
OOS R^2	–4.38	0.69*	–5.75	0.57*	–8.77	–0.90	–3.90	–7.66
CT R^2	1.88**	4.66**	2.18**	4.73**	1.55*	4.15**	–0.34	–0.82
PTV R^2	2.62**	4.74**	2.71**	4.87**	4.34**	4.49**	–2.83	–3.28
Predictor:	RV$t-11,t$		tbl		lty		ltr	
OOS R^2	–0.92	–8.85	–12.46	–11.89	–16.52	–20.32	0.81***	0.81***
CT R^2	–0.15	0.00	–1.11	2.51**	0.38*	0.95**	1.16***	1.06***
PTV R^2	–0.92	–1.41	2.40**	3.83**	–0.73	0.38**	0.90**	1.16***
Predictor:	tms		dfy		infl		bm	
OOS R^2	–0.93	3.04**	–3.23	–1.17	–0.45	1.60**	–19.55	–2.35
CT R^2	0.43*	2.83**	–2.34	–0.19	0.04	1.89**	–7.88	–1.20
PTV R^2	3.86***	4.60***	–2.40	–0.78	–0.34	1.22**	–0.61	1.73**

(continued)

Table 3
Continued

B. Predicting next-year returns

Predictor:	csp		ntis		lpy		caya	
OOS R^2	–3.76	–4.67	–14.97	–7.54	–17.99	0.71	1.45*	3.37**
CT R^2	–1.83	–2.54	–14.99	–7.54	2.17*	4.93**	1.14*	2.34**
PTV R^2	–3.74	–4.67	0.13	–0.24	4.52**	5.61**	6.55***	6.88***

Summary statistics

	IS R^2	OOS R^2				CT R^2				PTV R^2			
				# sig.				# sig.				# sig.	
	Mean (%)	Mean (%)	#>0	10%	5%	Mean (%)	#>0	10%	5%	Mean (%)	#>0	10%	5%
OLS	2.96	–6.96	2	2	1	–1.11	9	8	3	1.03	9	8	8
WLS-EV	2.11	–3.41	7	6	4	1.11	10	10	10	1.82	11	11	11
Diff	–0.84	3.55	12	5	1	2.22	13	2	1	0.79	12	1	0

This table presents statistics on the out-of-sample predictability afforded by 16 candidate predictors from Table 2. In panel A, I predict next-month log dividend-inclusive excess returns of the CRSP value-weighted index. In panel B, the forecast horizon is 1 year. The predictors are identical to those in Table 2 with two exceptions: since cay and RV $\hat{\sigma}^2_m$ require the full-sample of data to construct, I replace them with a rolling estimate of cay, caya, and past realized variance $RV_{m-11,m}$. For each predictor, I compute out-of-sample return forecasts starting 20 years after the sample begins, using both OLS and WLS-EV with out-of-sample variance forecasts, as detailed in Section 1. Given these out-of-sample forecasts, I compute the out-of-sample R^2 (OOS R^2) using the procedure described in Section 2. I also compute the out-of-sample R^2 using the Campbell and Thompson (2008) (CT R^2) and Pettenuzzo, Timmermann, and Valkanov (2014) (PTV R^2) approaches, described in Section 2. I compute p-values using the simulations described in Appendix Appendix B. *$p < .1$; **$p < .05$; ***$p < .01$. Both panels use 1,062 monthly observations from 1927 to 2015.

where T is the number of observations in the post-training sample period and K is the number of regressors (including the constant). Following Goyal and Welch (2008), I focus my analysis on in-sample and out-of-sample adjusted R^2 (OOS R^2 hereafter).

I use OOS R^2 as both a measure of economic value to investors and as a test statistic for the no-predictability null hypothesis. I compute the small-sample distribution of OOS R^2 under the null hypothesis using the same simulation procedure I use for in-sample standard errors, as described in Appendix Appendix B. The p-values in Table 3 are the fraction of simulated samples with larger OOS R^2 than I find in the observed sample.

Table 3 presents the OOS R^2 afforded by the 16 traditional predictors for next-month returns in panel A and next-year returns in panel B. The OLS results echo the conclusion in Goyal and Welch (2008) that these predictors, when combined with OLS, do not produce significant OOS R^2. Only one predictor has OOS R^2 with p-value below 5% for next-month returns (tms) and for next-year returns (ltr).

Table 3 shows the out-of-sample performance of WLS-EV is substantially better than OLS. WLS-EV OOS R^2 are higher than their OLS counterparts

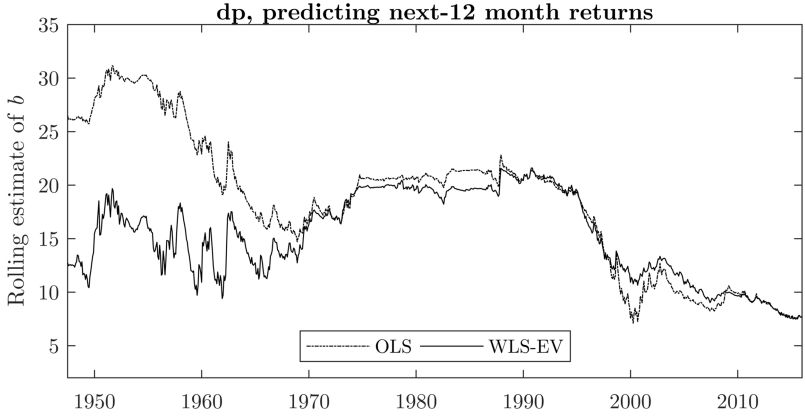

Figure 2
Rolling estimates of the predictability coefficient for the dp ratio
This figure presents coefficients from rolling regressions of next-year returns on the log dividend-to-price (dp) ratio. Specifically, for each date τ in my monthly sample following a 20-year training period, using only data available at time τ, that is, $m \leq \tau - 12$, I estimate the regression:

$$r_{m+1,m+12} = a + b \cdot dp_m + \epsilon_{m+1,m+h},$$

where $r_{m+1,m+12}$ is the log dividend-inclusive excess return of the CRSP value-weighted index over the 12 months starting with $m + 1$, and dp_m is the log dividend-to-price ratio in month m. For each τ, I compute point estimates $\hat{a}_\tau$ and $\hat{b}_\tau$ using OLS and WLS-EV, detailed in Section 1. I plot the resultant OLS and WLS-EV estimates $\hat{b}_\tau$ for each month from 1947 to 2015.

for 11 of the predictors using next-month and 12 for next-year returns. The mean OOS R^2 across predictors is -0.21% and -3.41% for next-month and next-year returns for WLS-EV, compared to -0.41% and -6.96% for OLS. These increases are economically large relative to the average in-sample OLS R^2 of 0.34% and 2.96%, respectively.

Finally, and most importantly, four predictors offer positive OOS R^2 with p-values below 5% using WLS-EV for both next-month returns and next-year returns (ltr, tms, infl, and caya). Two predictors, dp and dy, have p-values below 10% for both horizons. As was the case for in-sample tests presented in Table 2, out-of-sample tests produce much stronger evidence for return predictability by these variables when using WLS-EV instead of OLS.

To illustrate the source of the out-of-sample performance gains, I examine the dividend-price ratio (dp) in more detail. Figure 2 shows the evolution of $\hat{b}_\tau$ over the post-training period for both OLS and WLS-EV estimates. For next-year returns, although the full-sample estimates are very similar for OLS and WLS-EV (both around 7.7, as presented in panel B of Table 2), the rolling WLS-EV estimates are closer to the full-sample estimate early in the sample and more stable over time, both reflecting greater efficiency. Online Appendix A shows that the rolling coefficients for the other predictors follow a similar pattern, with WLS-EV rolling $\hat{b}_\tau$ closer to full-sample $\hat{b}_T$ than their OLS

counterparts. This pattern results in WLS-EV OOS R^2 closer to the IS R^2 than OLS OOS R^2.

Despite the improved OOS performance of WLS-EV relative to OLS, many predictors that are significant in-sample still have negative or insignificant OOS R^2 when using WLS-EV. Campbell and Thompson (2008) provides a method for improving the OOS performance of these predictors. Specifically, Campbell and Thompson (2008) suggests two economically motivated restrictions on the $\hat{b}_\tau$ and $\mathbb{E}_\tau(r_{\tau+1,\tau+h}|x_\tau)$:

1. For each predictor, economic theory suggests the correct sign of b. If $\hat{b}_\tau$ has the economically incorrect sign, set $\hat{b}_\tau = 0$ and $\mathbb{E}_\tau(r_{\tau+1,\tau+h}|x_\tau) = \mathbb{E}_\tau(r_{\tau+1,\tau+h})$.
2. The expected equity risk premium $\mathbb{E}_\tau(r_{\tau+1,\tau+h}|x_\tau)$ should always be positive. If it is not, use $\mathbb{E}_\tau(r_{\tau+1,\tau+h}|x_\tau) = 0$.

I apply these restrictions for each of the return predictors, using both OLS and WLS-EV, and compute the resultant adjusted OOS R^2, as defined in Equation (12).

Table 3 shows OOS R^2 using the Campbell and Thompson (2008) approach (CT R^2) for each predictor. For both OLS and WLS-EV, CT R^2 are substantially higher than OOS R^2, with 9 (9) predictors offering positive CT R^2 for next-month (next-year) returns using OLS. While the Campbell and Thompson (2008) approach improves the economic magnitude of out-of-sample performance, this does not necessarily imply it more-strongly rejects the no predictability null because CT R^2 are larger even under the null, meaning they have higher simulation-based critical values than standard OOS R^2. Nevertheless, I find that CT R^2 using OLS are statistically significant for 7 (3) predictors of next-month (next-year) returns.

WLS-EV still outperforms OLS when paired with the Campbell and Thompson (2008) restrictions. WLS-EV CT R^2 are higher than OLS CT R^2 for 12 (13) predictors of the next-month (next-year), and the mean CT R^2 is 0.18% (1.11%) for next-month (next-year) returns, compared to 0.00% (−1.11%) for OLS. The statistical evidence of predictability is also much stronger using WLS-EV instead of OLS alongside Campbell and Thompson (2008), with 9 (10) predictors having p-values below 5% for next-month (next-year) returns.

Another method for improving out-of-sample performance is to use the economic restrictions in Pettenuzzo, Timmermann, and Valkanov (2014), which require conditional annualized Sharpe ratios for the market be bounded between zero and one. Therefore, I estimate coefficients $\hat{a}_\tau$ and $\hat{b}_\tau$, using the regression in Equation (7) constrained so that

$$0 \leq \frac{\hat{a}_\tau + \hat{b}_\tau x_m}{\hat{\sigma}_m} \leq 1 \quad \forall m \leq \tau - h. \tag{14}$$

This differs from the Campbell and Thompson (2008) approach by requiring the conditional risk premium is positive for all $m \leq \tau - h$, rather than just τ, and adds an upper bound on the conditional Sharpe ratio.

Table 3 shows that, as demonstrated in Pettenuzzo, Timmermann, and Valkanov (2014), these economic restrictions result in out-of-sample R^2 (PTV R^2) substantially larger than even the CT R^2. Because the Pettenuzzo, Timmermann, and Valkanov (2014) approach already dampens the influence of extreme observations on OLS point estimates, the additional improvement afforded by WLS-EV is smaller for PTV R^2 than for OOS R^2 or CT R^2. Nevertheless, using WLS-EV instead of OLS results in slightly higher average PTV R^2 and more predictors with positive and significant PTV R^2.

As an alternative measure of out-of-sample performance, I also compute certainty equivalents (CEs) for an investor optimizing their portfolio using estimated conditional means and variances, an approach used in Campbell and Thompson (2008) and Johannes, Korteweg, and Polson (2014). Compared to OOS R^2, CEs have the advantage of a natural economic interpretation but the disadvantage of being dependent on the investor's utility function. Online Appendix B shows that CEs follow the same pattern as OOS R^2, with WLS-EV offering a 20- to 50-bp increase in per-year CE over OLS.

2.3 Discussion

Goyal and Welch (2008) comes to a pessimistic conclusion about return predictability: "despite extensive search, we were unsuccessful in identifying any models on annual or shorter frequency that systematically had both good IS and OOS performance" (Goyal and Welch 2008, p. 1504).

I show that using my more efficient estimator results in a more optimistic conclusion about return predictability. While my results echo Goyal and Welch (2008) in that none of the predictors systematically and significantly predict returns both in- and out-of-sample when using OLS, I show that four predictors meet the Goyal and Welch (2008) criteria when using WLS-EV: long-term bond return (ltr), term spread (tms), inflation (infl), and consumption-to-wealth ratio (cay).[7] Three more predictors meet the criteria when OOS performance uses WLS-EV with the Campbell and Thompson (2008) economic restrictions: log dividend-to-lagged price ratio (dy), Treasury-bill rate (tbl), and log payout yield (lpy).

For three of the predictors not meeting the stringent Goyal and Welch (2008) criteria—the dividend-to-price (dp) ratio, earnings-to-price (ep) ratio, and long-term bond yields (lty)—optimism is merited, because WLS-EV

[7] I define the Goyal and Welch (2008) criteria as having simulated p-values less than 5% for both in-sample and out-of-sample tests using both next-month and next-year returns.

evidence is stronger than OLS evidence, albeit being inconsistent across specifications. Finally, the remaining six predictors have no convincing evidence of predictability based on WLS-EV or OLS estimates: the dividend-to-earnings (de) ratio, conditional variance (RV $\hat{\sigma}_m^2$), default spread (dfy), book-to-market (bm) ratio, cross-sectional beta premium (csp), and net equity issuance (ntis).

The stronger evidence for predictability when using my more-efficient estimator is consistent with returns being predictable in the time series, as suggested by modern asset pricing and macro-theory models, but OLS estimates being inefficiently noisy. When using the more-efficient WLS-EV estimator, in-sample point estimates resemble OLS estimates in most cases, remaining economically substantial. However, weighting observations by ex ante volatility improves efficiency, reducing simulated standard errors by an average of 27% for next-month returns and 26% for next-year returns (see summary statistics in Table 2). This added efficiency allows more-frequent rejection of the no-predictability null using the same data, and also improves out-of-sample performance by reducing the estimation error embedded in out-of-sample return forecasts.

Alternative approaches to improving the out-of-sample performance of return predictors (e.g., those in Campbell and Thompson 2008; Lettau and Van Nieuwerburgh 2008; Johannes, Korteweg, and Polson 2014; Pettenuzzo, Timmermann, and Valkanov 2014) often improve out-of-sample performance as much or more than using WLS-EV. As a methodology for improving out-of-sample performance, using WLS-EV has the advantage of being a minimal extension to OLS, making it easier to understand and implement, and the disadvantage of not being designed to maximize out-of-sample performance. Instead, WLS-EV is designed to provide a more efficient in-sample test of the "no time-invariant linear predictability" null hypothesis tested by OLS. Alternative out-of-sample approaches either do not address in-sample estimation or test a different null hypothesis, precluding the type of in-sample inference that is the focus of this paper and much of the literature.

3. The Variance Risk Premium as a Predictor

3.1 Methodology

As a second application of WLS-EV, I revisit the empirical relation between future returns and the variance risk premium proxies in Bollerslev, Tauchen, and Zhou (2009) and Drechsler and Yaron (2011), BTZ and DY hereafter, and show it is not robust to WLS-EV. BTZ and DY show that the difference between VIX^2 and an estimate of statistical-measure variance positively predicts equity returns. Both papers motivate this result by modeling equity and variance risk premiums in a setting with stochastic volatility and

volatility-of-volatility, resulting in a positive correlation between equity and variance risk premiums.

BTZ and DY use slightly different empirical proxies for the variance risk premium, both of which I replicate. The BTZ proxy is

$$\text{BTZ } \widehat{\text{VRP}}_d \equiv \text{VIX}_d^2 - \text{IndRV}_{d-20,d}, \tag{15}$$

where VIX_d is the CBOE VIX index on day d and $\text{IndRV}_{d-20,d}$ is the realized variance of S&P 500 index returns over the 21 trading days ending on day d. I follow BTZ and compute IndRV from realized 5-minute log S&P 500 index returns and scale both VIX_d^2 and $\text{IndRV}_{d-20,d}$ to monthly percentages squared. The DY proxy for the variance risk premium is

$$\text{DY } \widehat{\text{VRP}}_d \equiv \text{VIX}_d^2 - \hat{\mathbb{E}}_d(\text{FutRV}_{d+1,d+21}), \tag{16}$$

where $\text{FutRV}_{d+1,d+21}$ is the sum of squared 5-minute log S&P 500 futures returns in the 21 trading days following d. I follow DY and use the fitted value from a full-sample time-series regression of $\text{FutRV}_{d+1,d+21}$ on $\text{IndRV}_{t-20,d}$ and VIX_d^2 as $\hat{\mathbb{E}}_d(\text{FutRV}_{d+1,d+21})$.

Unlike BTZ and DY, I use a daily sampling frequency for $\widehat{\text{VRP}}_d$ rather than monthly. The reason is the meaningful and observable day-to-day variation in $\widehat{\text{VRP}}_d$, with the half-lives of BTZ and DY VRP_d being only 4.6 and 5.2, respectively. Overlapping regressions with daily sampling use this variation to maximize power in a relatively short 1990–2015 sample period. 5 repeats my analysis with a monthly sampling frequency.

I use $\widehat{\text{VRP}}_d$ to predict $r_{d+1,d+h}$, the log excess return of the CRSP value-weighted index over the h days following the measurement of $\widehat{\text{VRP}}_d$. Because the results in BTZ and DY indicate that these proxies predict returns at 1-month and one-quarter horizons, I consider $h = 21$ and $h = 63$. Also following BTZ and DY, I scale log returns to annualized percentages. I adjust the point estimates for the Stambaugh (1999) bias, using the simulation procedure described in Appendix Appendix B., and account for the overlap using the approach described in Section 1.2. I also compute simulated standard errors and p-values using the heteroscedastic simulations (Sim), described in Appendix Appendix B. For both the observed and simulated samples, I compute WLS-EV using the RV $\hat{\sigma}_d^2$ and VIXF $\hat{\sigma}_d^2$, defined in Section 1.

3.2 Results

Table 4 presents the results for all 12 combinations of variance risk premium proxy, forecast horizon, and estimator. In all cases, the OLS coefficients are much larger than the corresponding asymptotic standard errors, resulting in asymptotic p-values of 5.8%, 4.6%, 3.1%, and 0.3%. This indicates that the OLS estimates documented in BTZ and DY remain significant when using an overlapping daily sample that also includes 2008–2015.

Table 4

Predicting returns using the variance risk premium

A. Drechsler and Yaron (2011) approach

$\hat{\text{VRP}}_d = \text{VIX}_d^2 - \hat{\mathbb{E}}_d(\text{FutRV}_{d+1,d+21}^2)$

Forecast horizon:	1 month (h = 12)			3 months (h = 63)		
	OLS	WLS-RV	WLS-VIXF	OLS	WLS-RV	WLS-VIXF
Stambaugh $\hat{b}_{\text{adj}}$	0.409*	0.215	0.194	0.324*	0.182	0.170
Unadjusted b	0.418	0.223	0.202	0.331	0.189	0.177
SE (Asy)	(0.216)	(0.162)	(0.155)	(0.162)	(0.123)	(0.118)
p-value (Asy %)	5.8	18.6	21.3	4.6	13.9	15.0
SE (Sim)	(0.246)	(0.180)	(0.162)	(0.183)	(0.141)	(0.131)
p-value (Sim %)	9.6	23.2	23.3	7.7	19.9	19.4

B. Bollerslev, Tauchen, and Zhou (2009) approach

$\hat{\text{VRP}}_d = \text{VIX}_d^2 - \text{IndRV}_{d-20,d}^2$

Forecast horizon:	1 month (h = 12)			3 months (h = 63)		
	OLS	WLS-RV	WLS-VIXF	OLS	WLS-RV	WLS-VIXF
Stambaugh $\hat{b}_{\text{adj}}$	0.506**	0.233	0.240	0.396**	0.202	0.201*
Unadjusted b	0.513	0.241	0.247	0.404	0.210	0.208
SE (Asy)	(0.234)	(0.159)	(0.154)	(0.136)	(0.108)	(0.104)
p-value (Asy %)	3.1	14.2	11.9	0.4	6.0	5.3
SE (Sim)	(0.242)	(0.175)	(0.157)	(0.157)	(0.127)	(0.117)
p-value (Sim %)	3.7	18.3	12.6	1.2	11.2	8.5

This table presents estimates of return predictability regressions of the form:

$$r_{d+1,d+h} = a + b \cdot \hat{\text{VRP}}_d + \epsilon_{d+h},$$

where $r_{d+1,d+h}$ is the log dividend-inclusive excess return of the CRSP value-weighted index over the h days starting with $d+1$, annualized and as percentages. $\hat{\text{VRP}}_d$ is one of two proxies for the variance risk premium, both expressed as monthly percentages squared. The first, DY VRP_d, is from Drechsler and Yaron (2011). The second, BTZ VRP_d, is from Bollerslev, Tauchen, and Zhou (2009). For each predictor, I estimate b using OLS and WLS-EV, detailed in Section 1, using RV $\hat{\sigma}_d^2$ and VIXF $\hat{\sigma}_d^2$. I also adjust b for the Stambaugh bias using a simulation procedure. I compute asymptotic (Asy) errors and p-values for the bias-adjusted coefficients by mapping to equivalent nonoverlapping regressions, as described in Section 1, and then using Newey and West (1987) with 21 lags and the simulated (Sim) standard errors and p-values using the procedure described in Appendix Appendix B. The sample is 6,552 daily observations from 1990 to 2015. *$p < .1$; **$p < .05$; ***$p < .01$.

The WLS-EV results, which account for heteroscedasticity in point estimates as well as standard errors, are much more pessimistic than the OLS results. In all cases, the WLS-RV and WLS-VIXF point estimates are statistically insignificant at the 5% level despite standard errors that are as much as 35% smaller than their OLS counterparts. This insignificance is a result of WLS-EV estimates being around half of OLS estimates, indicating OLS estimates are driven primarily by observations with high ex ante volatility.

To help understand why WLS-EV estimates are much smaller than OLS estimates, Figure 3 plots the observed $\hat{\text{VRP}}_d$ and next-month returns $r_{d+1,d+21}$, along with the OLS and WLS-VIXF regression lines. The darkness of each point represents its weight in the WLS-VIXF regressions, where the observation with the highest weight is black and other points are on the gray scale based on what fraction of the maximum weight the corresponding

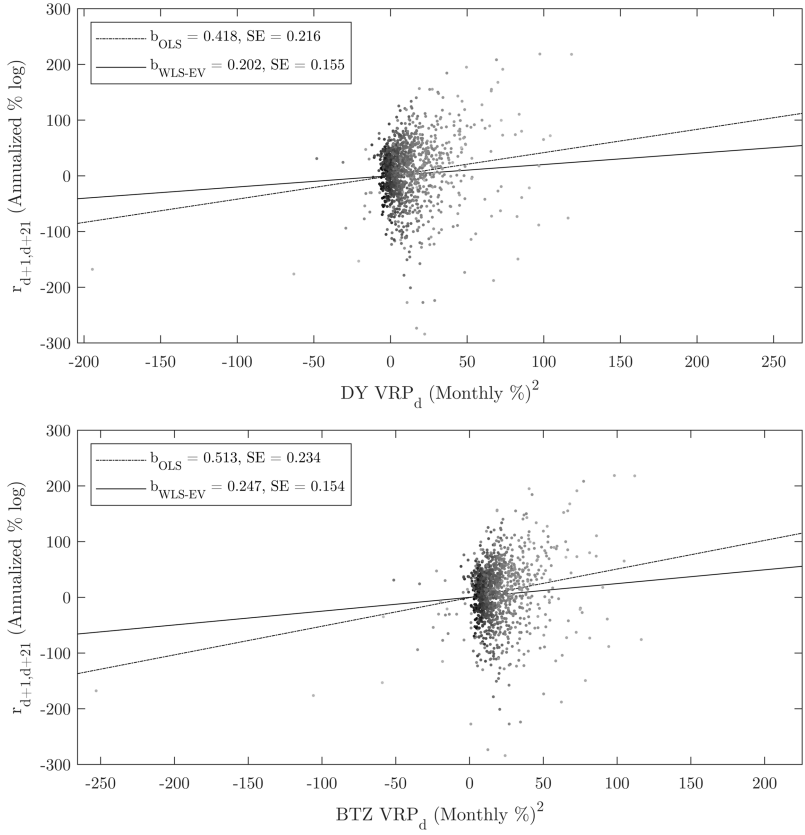

Figure 3
Predicting returns using variance risk premiums
This figure illustrates regressions of $r_{d+1,d+21}$, the log dividend-inclusive excess return of the CRSP value-weighted index over the 21 days starting with $d + 1$, annualized and as percentages, on one of two proxies for the variance risk premium, both expressed as monthly percentages squared. The first, DY VRP_d, is from Drechsler and Yaron (2011). The second, BTZ VRP_d, is from Bollerslev, Tauchen, and Zhou (2009). For each VRP_d, I compute point estimates of a and b using OLS and WLS-EV, as detailed in Section 1, using VIXF $\hat{\sigma}_d^2$. The lines represent the predicted values from the two regressions. The points represent the 6,552 daily observations from 1990 to 2015, with only every fifth observation plotted to improve readability. The darkness of each point represents its weight in the WLS-EV regressions, where the observation with the highest weight is black.

observation receives. Figure 3 indicates that $\hat{VRP}_d$ is near zero for most observations but extremely positive or negative for a small subset. These extreme observations also have high VIXF $\hat{\sigma}_d^2$, resulting in a low weight in the WLS-VIXF regressions. The observations with extremely negative $\hat{VRP}$ happened to have negative future return realizations, and those with extremely positive $\hat{VRP}$ happened to have positive future return realizations. Thus, when these points receive full weight in OLS, the coefficient is strongly positive. However, WLS-EV downweights these points because their return

realizations are particularly poor proxies for expected returns, and instead fits mostly on the darker points toward the middle of the distribution, which do not significantly support return predictability.[8]

3.3 Robustness

To ensure the failure of $\widehat{\text{VRP}}$ to significantly predict returns in Table 4 is not driven by the extended sample or overlapping daily returns, I repeat my analysis on a monthly sample from 1990 to 2007, the sampling frequency frequency and sample period used in BTZ and DY. For the BTZ analysis, I use the proxy on Hao Zhou's website to assure that my results are not driven by an error in my calculation of $\widehat{\text{VRP}}$.[9] To match BTZ and DY, I also use S&P 500 returns, rather than the CRSP index returns I use throughout the paper.

The results of my replication analysis are in Table 5, along with the point estimates and standard errors from the original DY and BTZ papers for comparison. In both cases, my replication is quite close to the original papers in terms of *t*-stats and *p*-values. I extend this replication by estimating heteroscedastic simulated (Sim) standard errors as well as WLS-EV using RV $\hat{\sigma}^2_m$ on the original BTZ and DY samples.

Unlike for my daily approach, simulated standard errors and *p*-values for the monthly samples in Table 5 are much higher for OLS than their asymptotic counterparts, reflecting the influence of severe heteroscedasticity in a short monthly sample. As a result, even the OLS point estimates are no longer significant at the 5% level in the original monthly samples when *p*-values are computed using heteroscedastic simulations.

More importantly, even in the original papers' sample, the WLS-EV estimates are only about half as large as the OLS estimates, making them statistically insignificant despite lower simulated standard errors. This failure holds across both VRP proxies and both prediction horizons, in all cases with simulated *p*-values above 29%. Table 5 also shows results for simple and log S&P 500 returns, 1990–2007 daily samples, extended monthly samples, and combinations thereof. In none of the 24 alternative procedures are WLS-EV estimates statistically significant at the 5% level, and in only 4 of the 24 are OLS estimates significant at the 5% level when *p*-values are computed using heteroscedastic simulations.

Finally, Table 5 shows $\widehat{\text{VRP}}$'s predictability indicated by calendar-monthly sampling is stronger than the predictability indicated by "offset" monthly sampling where each "offset month" is defined as starting on the

[8] Online Appendix C shows that other approaches to mitigating the influence of these observations, for example, by using deciles of $\widehat{\text{VRP}}_d$ or winsorizing $\widehat{\text{VRP}}_d$ below at zero, result in even weaker evidence of return predictability in both OLS and WLS-EV regressions.

[9] The proxy I use in Table 4 differs from the downloadable version only because I measure it daily using a rolling 21-day window, rather than measuring it for calendar months.

Table 5

Predicting returns using the variance risk premiums: Alternative procedures

A. Drechsler and Yaron (2011) approach

$\hat{\text{VRP}}_d = \text{VIX}_d^2 - \hat{\mathbb{E}}_d(\text{FutRV}_{d+1,d+21}^2)$

$\hat{\text{VRP}}_d$ measure:	Original	Replicated		Replicated		Replicated		Replicated		Replicated		Replicated	
Sample period:	1990-2007	1990-2007		1990-2007		1990-2007		1990-2015		1990-2015		1990-2015	
Sample frequency:	Monthly	Monthly		Monthly		Daily		Monthly		Offset Monthly		Daily	
Index:	S&P	S&P		log S&P		log S&P		log S&P		log S&P		log S&P	
	OLS	OLS	WLS	OLS	WLS	OLS	WLS	OLS	WLS	OLS	WLS	OLS	WLS
Forecast horizon: 1 month													
Stambaugh $\hat{b}_{\text{adj}}$		0.378	0.204	0.344	0.164	0.435	0.196	0.539	0.386	0.347	0.444	0.403	0.221
Unadjusted $\hat{b}$	0.760***	0.402	0.228	0.370	0.190	0.460	0.221	0.547	0.393	0.354	0.452	0.411	0.230
SE (Asy)	(0.350)	(0.178)	(0.214)	(0.179)	(0.216)	(0.239)	(0.232)	(0.174)	(0.175)	(0.219)	(0.296)	(0.212)	(0.157)
p-value (Asy %)	2.9	3.4	34.1	5.4	44.7	6.8	39.7	0.2	2.7	11.3	13.3	5.7	15.9
SE (Sim)		(0.386)	(0.325)	(0.387)	(0.325)	(0.325)	(0.270)	(0.333)	(0.262)	(0.348)	(0.274)	(0.246)	(0.180)
p-value (Sim %)		33.2	52.9	37.8	61.6	18.0	46.9	10.5	14.2	32.1	10.5	10.2	21.9
Forecast horizon: 3 months													
Stambaugh $\hat{b}_{\text{adj}}$		0.423	0.230	0.399	0.204	0.394	0.159	0.481**	0.305*	0.314	0.272	0.305*	0.168
Unadjusted $\hat{b}$	0.860***	0.444	0.251	0.421	0.225	0.413	0.178	0.488	0.312	0.321	0.279	0.312	0.175
SE (Asy)	(0.270)	(0.141)	(0.141)	(0.142)	(0.142)	(0.178)	(0.180)	(0.096)	(0.103)	(0.127)	(0.150)	(0.159)	(0.120)
p-value (Asy %)	0.1	0.3	10.3	0.5	15.2	2.7	37.8	0.0	0.3	1.3	7.0	5.5	16.2
SE (Sim)		(0.259)	(0.230)	(0.261)	(0.232)	(0.243)	(0.214)	(0.207)	(0.172)	(0.221)	(0.181)	(0.183)	(0.142)
p-value (Sim %)		10.1	31.6	12.5	37.9	10.4	46.0	2.0	7.6	15.6	13.3	9.5	23.5

B. Bollerslev, Tauchen, and Zhou (2009) approach

$\hat{\mathrm{VRP}}_d = \mathrm{VIX}_d^2 - \mathrm{IndRV}_{d-20,d}^2$

$\hat{\mathrm{VRP}}_d$ measure:	Original	Website		Replicated		Replicated		Replicated		Replicated		Replicated	
Sample period:	1990-2007	1990-2007		1990-2007		1990-2007		1990-2015		1990-2015		1990-2015	
Sample interval:	Monthly	Monthly		Monthly		Daily		Monthly		Offset Monthly		Daily	
Index:	log S&P	log S&P		log S&P		log S&P		log S&P		log S&P		log S&P	
	OLS	OLS	WLS	OLS	WLS	OLS	WLS	OLS	WLS	OLS	WLS	OLS	WLS
Forecast horizon: 1 month													
Stambaugh $\hat{b}_{\mathrm{adj}}$		0.364	0.190	0.383	0.213	0.456	0.156	0.561	0.408	0.435	0.484*	0.497**	0.218
Unadjusted $\hat{b}$	0.760*	0.381	0.206	0.400	0.230	0.478	0.178	0.564	0.410	0.439	0.488	0.505	0.225
SE (Asy)	(0.222)	(0.186)	(0.228)	(0.191)	(0.235)	(0.241)	(0.228)	(0.111)	(0.183)	(0.242)	(0.313)	(0.228)	(0.148)
p-value (Asy %)	7.8	5.1	40.5	4.5	36.5	5.9	49.2	0.0	2.6	7.2	12.2	2.9	14.1
SE (Sim)		(0.404)	(0.348)	(0.414)	(0.357)	(0.327)	(0.245)	(0.397)	(0.278)	(0.383)	(0.287)	(0.242)	(0.157)
p-value (Sim %)		37.1	58.4	36.1	55.3	16.4	52.4	15.8	14.3	25.8	9.2	4.0	16.7
Forecast horizon: 3 months													
Stambaugh $\hat{b}_{\mathrm{adj}}$		0.455*	0.248	0.477*	0.269	0.432*	0.167	0.464**	0.302*	0.377*	0.290*	0.382**	0.176
Unadjusted $\hat{b}$	0.470***	0.469	0.262	0.491	0.283	0.451	0.186	0.466	0.304	0.381	0.293	0.389	0.183
SE (Asy)	(0.164)	(0.133)	(0.135)	(0.133)	(0.136)	(0.174)	(0.171)	(0.071)	(0.089)	(0.102)	(0.139)	(0.131)	(0.101)
p-value (Asy %)	0.4	0.1	6.7	0.0	4.7	1.3	32.9	0.0	0.1	0.0	3.7	0.4	8.0
SE (Sim)		(0.265)	(0.235)	(0.268)	(0.239)	(0.241)	(0.197)	(0.221)	(0.156)	(0.211)	(0.172)	(0.157)	(0.117)
p-value (Sim %)		8.5	29.5	7.4	26.1	7.3	39.7	3.5	5.1	7.4	9.1	1.5	13.1

This table presents variations of the regressions in Table 4 but with different measures, sample periods, sample frequencies, and return indices. Columns with "Original" measures contain estimates copied from the original papers, columns with "Website" measures contain my estimates using data uploaded by the authors, and the remaining columns contain my estimates using replicated $\hat{\mathrm{VRP}}$ measures. The two potential indices are the CRSP value-weighted index and the S&P 500 index, both including dividends and net of the risk-free rate. Offset monthly samples redefine months as beginning of the 15th day of month m and ending on the 14th day of month $m+1$. For each column, I estimate b using OLS and WLS-EV, detailed in Section 1, using RV $\hat{\sigma}_t^2$. I also adjust b for the Stambaugh bias using a simulation procedure. I compute asymptotic (Asy) errors and p-values for the bias-adjusted coefficients by mapping to equivalent nonoverlapping regressions, as described in Section 1, and then using Newey and West (1987) with 21 lags for daily sampling and 12 lags for monthly sampling, and the simulated (Sim) standard errors and p-values using the procedure described in Appendix Appendix B. Monthly sampling results in 215 monthly observations from 1990 to 2007 and 311 from 1990 to 2015, while daily sampling results in 4,537 observations from 1990 to 2007 and 6,552 from 1990 to 2015. *$p < .1$; **$p < .05$; ***$p < .01$.

15th of calendar month m and ending on the 14th of month $m + 1$. For each of these offset months, I compute excess market returns, realized variances, variance risk premium proxies, and everything else I need to assess the variance risk premiums's predictability exactly as I do for calendar-monthly sampling. This procedure preserves the sample size, degree of overlap, and economic basis of calendar-monthly sampling. The columns labelled "Offset Monthly" in Table 5 show that OLS point estimates are between 18% and 35% smaller in offset-monthly samples compared to calendar-monthly samples, and none are significant at the 5% level when using OLS or WLS-EV. This reduction has two potential intrepretations: either variance and equity risk premiums are more strongly related at the end of calendar months or the statistical relation in this sample happens to be stronger when measured only at the end of the month. I account for this pattern by using daily sampling for my main analysis, which effectively averages the predictive coefficient from calendar-monthly observations with the coefficients from all the other potential offsets (1, 2, ... 30 days).

Online Appendix D revisits the analysis of VRP's predictability around the world in Bollerslev et al. (2014) using WLS-EV. While I replicate the OLS evidence of predictability in many countries for certain forecast horizons, this evidence disappears when using small-sample standard errors based on heteroscedastic simulations or when using WLS-EV. With either methodology and a longer daily sample, none of the 56 country-forecast horizon pairs in Bollerslev et al. (2014) yield statistically significant evidence of return predictability.

3.4 Discussion

The results in Tables 4 and 5, Figure 3, and Online Appendices C and D do not indicate significant evidence for a linear relation between conditional equity and variance risk premiums when using more-efficient WLS-EV regressions. While this nonresult could indicate the OLS evidence of a relation is spurious, many other interpretations are also possible. One possibility is the relation between equity and variance risk premiums is positive but weaker than indicated by OLS, meaning that we cannot detect it in the relatively short 1990–2015 sample period. Consistent with this possibility, the WLS-EV estimates are often economically significant, especially for U.S. data. Another potential interpretation is that the relation between equity and variance risk premiums is time-varying or nonlinear, meaning linear regressions are misspecified.

Given these possible interpretations, I view my variance premium results as indicating a need for further analysis with additional data, or alternative nonlinear or time-varying specifications, to reach a conclusion about the predictive value of variance risk premium proxies for equity returns.

4. Politics, the Weather, and the Stars as Predictors

Novy-Marx (2014) discusses 10 potential return predictors: the political party of the president of the United States, the monthly highest temperature in New York City, the global average temperature, the rolling average global temperature, the quasiperiodic Pacific temperature anomaly (El Niño), the rolling average Pacific Ocean temperature, the observed number of sunspots, the rolling average number of sunspots, the angle between Mars and Saturn, and the angle between Jupiter and Saturn. Using OLS, Novy-Marx (2014) shows these 10 variables predict returns for 22 factors or anomalies, forcing readers to either accept implausible predictive relations or consider "rejecting the standard methodology on which the return predictability literature is built" (Novy-Marx 2014, p. 144).

As a final application of the WLS-EV methodology, I revisit the surprising predictability evidence in Novy-Marx (2014) and show it is weaker when using WLS-EV instead of the standard OLS methodology. For each of the 10 predictors Novy-Marx (2014) considers, I estimate monthly predictive regressions using data from 1961 to 2012.[10] I find that small-sample p-values based on the simulation approach in Appendix Appendix B. indicate OLS estimates are significant for the same three variables Novy-Marx (2014) finds predict market returns: the president's political party, New York City weather, and the Mars/Saturn angle. However, WLS-EV estimates of these variables' predictability are all closer to zero, and all have higher p-values than their OLS counterparts, with only Mars/Saturn remaining statistically significant. Moreover, WLS-EV estimates of the other seven variables remain insignificant despite smaller simulated standard errors.

I also estimate the joint significance of these variables in a multivariate predictive regression using an asymptotic χ^2 test, as well as a small sample χ^2 test that employs the covariance matrix of multivariate point estimates across simulated samples. I compute p-values for small sample χ^2 using the distribution of this statistic across the same simulations. Table 6 presents the resultant statistics for both OLS and WLS-EV estimators. The 10 variables are jointly significant with a p-value of 3.30% when using OLS and an asymptotic χ^2 test. However, my heteroscedastic simulation approach and WLS-EV both result in the 10 predictors no longer being jointly significant, with p-values between 11% and 22%.

4.1 The presidential puzzle

To illustrate what drives the difference between WLS-EV and OLS estimates for the Novy-Marx (2014) predictors, I examine the predictability afforded by the president's party in more detail. As documented in Santa-Clara and

[10] I follow Novy-Marx (2014) by limiting my sample to 1961–2012 and not adjusting for the Stambaugh (1999) bias.

Table 6
Predicting returns using politics, the weather, and the stars

Predicting next-month returns

	OLS	WLS-EV	OLS	WLS-EV	OLS	WLS-EV
Predictor:	Dem		NYC weather		Global temp.	
$\hat{b}$	0.767**	0.491	−0.026**	−0.016	0.112	0.250
p-value (Asy %)	3.3	12.5	3.7	14.9	80.4	50.9
p-value (Sim %)	3.9	12.7	4.3	15.6	83.8	59.2
Predictor:	Roll. global temp.		El Nino		Roll. El Nino	
$\hat{b}$	0.106	0.327	0.000	0.026	−0.303	0.039
p-value (Asy %)	84.6	44.8	99.8	87.9	52.5	92.5
p-value (Sim %)	85.8	52.5	99.8	87.9	57.1	93.0
Predictor:	Sunspots		Roll. sunspots		Mars/Saturn angle	
$\hat{b}$	−0.002	−0.003	0.005	0.006	0.513***	0.398**
p-value (Asy %)	44.9	25.2	55.3	47.1	1.3	1.8
p-value (Sim %)	46.1	25.7	65.4	52.0	0.9	1.6
Predictor:	Jupiter/Saturn angle					
$\hat{b}$	−0.009	−0.113				
p-value (Asy %)	96.5	56.9				
p-value (Sim %)	96.5	54.1				
Joint significance:	OLS	WLS-EV			OLS	WLS-EV
χ^2 statistic (Asy)	19.602**	14.84	χ^2 statistic (Sim)		15.51	13.19
p-value (Asy %)	3.30	13.80	p-value (Sim %)		11.30	21.40

This table presents estimates of return predictability regressions of the form:

$$r_{m+1} = a + b \cdot x_m + \epsilon_t + m,$$

where r_{m+1} is the log dividend-inclusive excess return of the CRSP value-weighted index in month $m + 1$, as percentages, and x_m is one of 10 predictors from Novy-Marx (2014): an indicator for whether the president of the United States is a Democrat (Dem), the monthly highest temperature in New York City (NYC weather), the global temperature anomaly (global temp.), the rolling average global temperature (roll. global temp.), the quasiperiodic Pacific temperature anomaly (El Niño), the rolling average Pacific Ocean temperature (roll. El Niño), the number of sunspots (sunspots), the rolling average number of sunspots (Roll. sunspots), the angle between Mars and Saturn (Mars/Saturn angle), and the angle between Jupiter and Saturn (Jupiter/Saturn angle). For each predictor, I compute point estimates of b using OLS and WLS-EV, detailed in Section 1, using RV $\hat{\sigma}_m^2$. I compute p-values for the coefficients as well as for joint significance using the heteroscedastic simulation procedure (Sim) described in Appendix Appendix B. and Section 4. The sample is 624 monthly observations from 1961 to 2012. *$p < .1$; **$p < .05$; ***$p < .01$.

Valkanov (2003) and Novy-Marx (2014), average market returns are higher under Democratic presidents than Republican ones. Pástor and Veronesi (2017) provides a model rationalizing this "presidential puzzle" in which voters prefer redistributive Democrats during bad times, which also feature high risk premiums.

I provide an alternative explanation for the presidential puzzle, that it stems from the realizations of unexpected returns in a few months with high ex ante volatility, rather than variations in expected returns as

hypothesized in Pástor and Veronesi (2017). To support this hypothesis, I revisit the presidential puzzle using the same sample period as Pástor and Veronesi (2017) but applying my WLS-EV estimator. Specifically, in a 1927–2015 monthly sample I regress annualized log excess market returns on an indicator for whether the current U.S. president is a Democrat at the end of month m:

$$r_m = a + b \cdot \text{Dem}_m + \epsilon_m. \tag{17}$$

This analysis differs from the Novy-Marx (2014) analysis by using a longer sample period and not lagging Dem_m because a new president is inaugurated in January, more than 2 months after the election, meaning Dem_m is always known prior to month m.

Panel A of Table 7 presents estimates of Equation (17) in the full sample and for the three subsamples. My OLS results replicate those in Santa-Clara and Valkanov (2003) and Pástor and Veronesi (2017), with a 10.19% difference in average annualized log returns between Democratic and Republican presidents. The magnitude of this difference is consistent in the first and second half of the sample, and the asymptotic p-value for the full sample is 1.7%.

Table 7 also shows that the presidential puzzle shrinks substantially and is insignificant when using WLS-EV. Specifically, the estimated difference in returns between Democratic and Republican presidents shrinks to 4.38%, and the p-value rises to 22.6%. The difference is also smaller and insignificant when using WLS-EV in both halves of the sample.

To help understand why WLS-EV estimates of the presidential effect are insignificant, I examine how the magnitude of the return difference varies with ex ante volatility RV $\hat{\sigma}_{m-1}$. By using ex ante volatility instead of realized volatility, I avoid the well known negative correlation between realized returns and volatility. Instead, months with high RV $\hat{\sigma}_{m-1}$ have volatile realized returns r_m that can be positive or negative and are particularly noisy measures of expected returns.[11] Panel B of Table 7 shows that among observations with high ex ante volatility, realized returns happened to be strongly positive under Democratic presidents and negative under Republican ones. This return difference is most dramatic among the top 1% of months by ex ante volatility, in which average market returns were 21.7% (260.4% annualized) under Democrats and –6.6% (–79.0% annualized) under Republicans.

Without these few observations with high ex ante volatility, the presidential puzzle disappears even when using OLS. Table 7 demonstrates this in two ways. First, panel A shows that without two post-crash periods (1929:11 through 1934:11 and 2008:11 through 2009:12), the difference in average

[11] For example, panel C of Table 7 illustrates that the top 1% of months by RV $\hat{\sigma}_{m-1}$ are not the months of market crashes, like those in October of 1929, 1987, or 2008, but rather are the months *after* market crashes.

Table 7
The presidential puzzle

A. Subsamples by time

Sample:	1927-2015		1927-1971		1972-2015		No post-crash	
Observations:	1,062		535		527		988	
	OLS	WLS-EV	OLS	WLS-EV	OLS	WLS-EV	OLS	WLS-EV
Democrat mean	10.69	8.63	10.70	8.53	10.68	8.79	9.77	8.56
Republican mean	0.49	4.24	0.24	7.06	0.66	2.02	3.80	5.06
Difference	10.19***	4.38	10.46*	1.47	10.01**	6.765*	5.97*	3.51
p-value (Asy %)	1.7	22.6	18.2	80.6	2.7	10.9	11.5	33.2
p-value (Sim %)	0.6	11.1	7.7	69.7	4.6	9.9	7.4	20.5
p-value (FSS Sim %)	1.3	17.0	10.1	74.1	6.6	14.4	11.4	27.6

B. Subsamples by ex ante volatility (RV $\hat{\sigma}_{t-1}$)

Sample:	Top 1 of σ_{m-1}		Top 10 σ_{m-1}		Top 20 σ_{m-1}		Bottom 80 σ_{m-1}	
Observations:	11		106		213		849	
	OLS	WLS-EV	OLS	WLS-EV	OLS	WLS-EV	OLS	WLS-EV
Democrat mean	260.41	266.81	11.10	3.14	15.07	12.70	9.64	8.39
Republican mean	−78.98	−73.52	−25.51	−20.33	−17.11	−13.63	5.14	5.52
Difference	339.38***	340.33***	36.61	23.47	32.18**	26.33**	4.50	2.87
p-value (Asy %)	0.0	0.0	16.1	29.4	2.6	2.4	20.8	38.9
p-value (Sim %)	0.4	0.5	13.6	27.9	1.9	2.2	15.3	30.9
p-value (FSS Sim %)	0.4	0.4	13.9	27.9	2.0	2.4	21.2	38.1

C. Top 1 of months by RV $\hat{\sigma}_{m-1}$

Democratic president			
Month	RV $\hat{\sigma}_{m-1}$ (%)	Monthly r_m (%)	Annualized r_m (%)
1933:04	14.3	32.0	384.5
1933:08	15.1	11.4	136.3
Mean	14.7	21.7	260.4
Republican president			
Month	RV $\hat{\sigma}_{m-1}$ (%)	Monthly r_m (%)	Annualized r_m (%)
1929:11	17.3	−13.6	−162.7
1929:12	14.2	1.5	17.5
1931:11	13.7	−9.6	−114.6
1932:09	14.1	−3.2	−38.5
1932:10	15.0	−14.0	−167.5
1932:11	14.2	−5.8	−69.4
1987:11	17.5	−8.1	−97.1
2008:11	16.8	−8.2	−98.2
2008:12	14.4	1.6	19.8
Mean	15.8	−6.6	−79.0

This table presents estimates of return predictability regressions of the form:

$$r_m = a + b \cdot \text{Dem}_m + \epsilon_m,$$

where r_m is the log dividend-inclusive excess return of the CRSP value-weighted index in month m, in annualized percentages. Dem_m is an indicator for whether the President of the United States is a Democrat at the end of month m. I estimate a and b using OLS and WLS-EV, detailed in Section 1, using RV $\hat{\sigma}^2_{m-1}$. Panel A estimates this regression across different subsamples partitioned by time and panel B by RV $\hat{\sigma}_{m-1}$. For each subsample and estimator, I present the Democrat mean ($\hat{a} + \hat{b}$), the Republican mean ($\hat{a}$), the Difference ($\hat{b}$), and p-values for $\hat{b}$ computed using using Newey and West (1987) (Asy) and simulation procedures described in Appendix Appendix B. without and with the Ferson, Sarkissian, and Simin (2003) effect (Sim and FSS Sim). Panel C presents RV $\hat{\sigma}_{m-1}$ and r_m in the top 1% of the sample by RV $\hat{\sigma}_{m-1}$, partitioned by whether the president was a Democrat or Republican. *$p < .1$; **$p < .05$; ***$p < .01$.

returns is insignificant using both OLS and WLS-EV.[12] Second, panel B shows that in time periods with normal volatility, defined as the bottom 80% of months by RV $\hat{\sigma}_{t-1}$, the difference in average returns is also insignificant. Of course, by arbitrarily excluding observations one can always strengthen or weaken an empirical result. However, the subsample analysis in Table 7 illustrates why WLS-EV estimates, which use the full sample but put less weight on observations with high ex ante volatility (and low signal-to-noise ratios), are so much smaller than OLS estimates of the presidential puzzle.

A potential statistical bias affecting the presidential puzzle evidence is the Ferson, Sarkissian, and Simin (2003) effect: if true expected returns are time varying and persistent in a way unrelated to the predictors being tested, these variations will be absorbed in the regression residuals, creating autocorrelation at long lags that is cannot be corrected for using the standard Newey and West (1987) approach. Instead, Ferson, Sarkissian, and Simin (2003) suggest using simulations in which returns are generated by combining an AR(1) process for expected returns with independent unexpected returns. I apply this suggestion, combined with the heteroscedastic unexpected returns detailed in Appendix Appendix B., to the presidential dummy predictor because it is extremely persistent and Powell et al. (2007) show that a related bias could be driving its significance.[13]

Table 7 shows that using Ferson, Sarkissian, and Simin (2003) simulations weakens the statistical significance of the presidential dummy, as evidenced by the "FSS Sim" p-values being larger than p-values from simulations with a constant mean. However, the Ferson, Sarkissian, and Simin (2003) effect is small relative to the WLS-EV effect. The impact of time-varying means is muted in my simulations because of the calibration I use (see Appendix Appendix B. for details). I allow true expected returns to be extremely persistent, which maximizes their impact, but restrict the extent of their variations so that the 90% confidence interval for annualized expected excess returns is $[-1.7\%, 14.3\%]$. This variation in expected returns is on the low end of the range studied in Ferson, Sarkissian, and Simin (2003), limiting its impact. More variable expected returns would have a bigger impact on statistical inference, but are difficult to reconcile with rational asset pricing theories.

4.2 Discussion

The insignificant WLS-EV and joint tests of predictability by the 10 variables in Novy-Marx (2014), including the party of the U.S. president, are consistent with the hypothesis that the predictive relations in Novy-Marx (2014) are

[12] I define the post-crash periods as starting in 1929:11 and 2008:11 and ending when annualized ex ante volatility RV $\hat{\sigma}_{t-1}$ next falls below 20% in 1934:11 and 2009:12.

[13] Powell et al. (2007) uses simulations where *realized* returns, instead of expected returns, follow an AR(1).

false positives driven by data mining targeting OLS significance. Because WLS-EV and joint significance were not used to select the predictors in Novy-Marx (2014) and therefore not the target of data mining, they naturally produce weaker predictability evidence than univariate OLS. Appendix Appendix B. illustrates this point by showing that, in samples simulated under the no-predictability null for which the OLS estimate is nevertheless significant, perhaps because of data mining, WLS-EV estimates are 40–50% closer to zero on average and insignificant in 55–65% of these false-positive samples. Of course, WLS-EV evidence could also be subject to data mining if predictors are selected from many candidates based on their WLS-EV significance, in which case I also show OLS estimates are often insignificant. However, any data mining in prior literature is likely to have targeted OLS, making WLS-EV useful in revisiting return predictability.

The presidential puzzle evidence supports and provides nuance to the false-positive hypothesis for the Novy-Marx (2014) results. Using OLS, a false positive is likely to arise when a predictor happens to historically portend large unexpected returns during periods of high ex ante volatility. I find this exact pattern in Table 7, with returns being highly positive for a few volatile observations under Democrats and negative under Republicans. The magnitudes of these realizations are difficult to reconcile with variations in risk premiums, which should be small and positive. Instead, they are consistent with unexpected returns driven by randomness in small samples.

5. Conclusion

I study time-series return predictability by developing and applying WLS-EV, an estimator that incorporates information about time-varying volatility into both point estimates and standard errors. The WLS-EV approach is convenient to use and 25%–35% more efficient than OLS in standard settings.

My results indicate that the apparent failures of traditional variables, discussed in Goyal and Welch (2008), are often false negatives due to a lack of power in OLS estimates rather than a fundamental failure of return predictability. When combined with the restriction that out-of-sample forecasts must always be positive, seven predictors significantly predict next-month and next-year returns, both in- and out-of-sample, when using WLS-EV: dividend yield, Treasury-bill yield, long-term Treasury return, term spread, inflation, payout yield, and consumption-to-wealth ratio. None of these predictors work as well or as consistently when using OLS. Of the remaining predictors, three have WLS-EV evidence that is stronger than OLS but still inconsistent (dividend-to-price ratio, earnings-to-price (ep) ratio, and long-term Treasury yields) and the remaining six (dividend-to-earnings ratio, conditional variance, default spread, book-to-market ratio, cross-sectional beta

premium, and net equity issuance) are insignificant for both OLS and WLS-EV estimates.

On the other hand, I show that the evidence of a predictive relation between the variance risk premium and market returns depends critically on a few observations with high ex ante volatility, indicating the need for a longer sample period or further nonlinear analysis to reach a conclusion on the true empirical relation. My results are also consistent with the false-positive interpretation of the predictive variables presented in Novy-Marx (2014), including the party of the U.S. president, the weather in New York City, and the angle between Mars and Saturn, illustrating that WLS-EV is useful for reassessing predictors that may be spurious or have been selected via data mining based on OLS significance.

Overall, my more-efficient estimator affirms the predictability afforded by predictors with long sample periods and a connection to the equity risk premium grounded in theory and calls into question predictors with shorter sample periods or weak grounding in theory.

Appendix A. Technical Details

A.1. Transforming Overlapping Regressions into Equivalent Nonoverlapping Regressions

For regressions with overlapping return observations, I first map these regressions into equivalent nonoverlapping return regressions following a trick suggested in Hodrick (1992). Consider a regression of next h-period log returns returns on X_t:

$$r_{t+1,t+h} = X_t \cdot \beta + \epsilon_{t+1,t+h}. \tag{A1}$$

Two potential estimates of β are the standard overlapping OLS estimates $\hat{\beta}_{\text{OLS}}^{\text{overlap}}$ and the Hodrick (1992) estimates $\hat{\beta}_{\text{OLS}}^{\text{hodrick}}$, defined as follows:

$$\hat{\beta}_{\text{OLS}}^{\text{overlap}} = \mathbb{E}_T(X_t' X_t)^{-1} \mathbb{E}_T(X_t' r_{t+1,t+h}), \tag{A2}$$

$$\hat{\beta}_{\text{OLS}}^{\text{hodrick}} = \mathbb{E}_T(X_t' X_t)^{-1} \mathbb{E}_T(\bar{X}_t' \bar{X}_t) \underbrace{\mathbb{E}_T(\bar{X}_t' \bar{X}_t)^{-1} \mathbb{E}_T(\bar{X}_t' r_{t+1})}_{\equiv \hat{\beta}_{\text{OLS}}^{\text{roll}}}, \tag{A3}$$

where $\mathbb{E}_T$ represents the sample average and $\bar{X}_t \equiv \sum_{s=0}^{h-1} X_{t-s}$ is the rolling sum of past X_t. $\hat{\beta}_{\text{OLS}}^{\text{roll}}$ is the OLS coefficient estimate in a nonoverlapping regression of r_{t+1} on $\bar{X}_t$.

These two estimates both converge to the true β asymptotically, and they are approximately equal to each other in finite samples.

Theorem 1. Under the identifying assumption $E[X_t \cdot \epsilon_{t+1,t+h}] = 0$, both $\hat{\beta}_{\text{OLS}}^{\text{overlap}}$ and $\hat{\beta}_{\text{OLS}}^{\text{hodrick}}$ are asymptotically consistent estimates of β:

$$\beta_{\text{OLS}}^{\text{overlap}} \equiv \text{plim}_{T\to\infty} \hat{\beta}_{\text{OLS}}^{\text{overlap}} = \beta, \tag{A4}$$

$$\beta_{\text{OLS}}^{\text{hodrick}} \equiv \text{plim}_{T\to\infty} \hat{\beta}_{\text{OLS}}^{\text{hodrick}} = \beta. \tag{A5}$$

Proof. Standard econometric results show that $\hat{\beta}_{\text{OLS}}^{\text{overlap}}$ converges in probability to

$$\beta_{\text{OLS}} = \mathbb{E}(X_t'X_t)^{-1}\mathbb{E}(X_t'r_{t+1,t+h}) = \beta. \tag{A6}$$

Substituting in $r_{t+1,t+h} = \sum_{s=1}^{h} r_{t+s}$, we have:

$$\beta_{\text{OLS}} = \mathbb{E}(X_t'X_t)^{-1}\mathbb{E}\left(\sum_{s=1}^{h} X_t'r_{t+s}\right) \tag{A7}$$

$$= \mathbb{E}(X_t'X_t)^{-1}\mathbb{E}(\bar{X}_t'\bar{X}_t)\mathbb{E}(\bar{X}_t'\bar{X}_t)^{-1}\mathbb{E}\left[\left(\sum_{s=0}^{h-1} X_{t-s}'\right)r_{t+1}\right] \tag{A8}$$

$$= \beta_{\text{OLS}}^{\text{hodrick}}. \tag{A9}$$

Combined, we have $\beta_{\text{OLS}}^{\text{hodrick}} = \beta_{\text{OLS}}^{\text{overlap}} = \beta$. □

Lemma 2. In settings in which X_t includes a constant and a univariate predictor x_t, Theorem 1 implies $b_{\text{OLS}}^{\text{overlap}} = b_{\text{OLS}}^{\text{hodrick}}$, where

$$b_{\text{OLS}}^{\text{overlap}} \equiv \frac{\text{Cov}(r_{t+1,t+h}, x_t)}{\text{Var}(x_t)} \tag{A10}$$

$$b_{\text{OLS}}^{\text{hodrick}} \equiv \frac{\text{Var}(\bar{x}_t)}{\text{Var}(x_t)} \underbrace{\frac{\text{Cov}(r_{t+1}, \bar{x}_t)}{\text{Var}(\bar{x}_t)}}_{b_{\text{OLS}}^{\text{roll}}}. \tag{A11}$$

Equation (A11) is the basis for the approach I use throughout the paper in settings with overlapping observations.

In addition to converging to the same value asymptotically, $\hat{\beta}_{\text{OLS}}^{\text{overlap}}$ and $\hat{\beta}_{\text{OLS}}^{\text{hodrick}}$ are approximately equal in small samples. We have

$$\hat{\beta}_{\text{OLS}}^{\text{overlap}} = \left(\frac{1}{T-h}\sum_{t=1}^{T-h} X_t'X_t\right)^{-1} \frac{1}{T-h}\sum_{t=1}^{T-h}\sum_{s=1}^{h} X_t'r_{t+s}, \tag{A12}$$

$$= \left(\frac{1}{T-h}\sum_{t=1}^{T-h} X_t'X_t\right)^{-1} \frac{1}{T-h}\sum_{\substack{t\in[1,T-h] \\ (u-t)\in[1,h]}}\sum X_t'r_u, \tag{A13}$$

$$\hat{\beta}_{\text{OLS}}^{\text{hodrick}} = \left(\frac{1}{T}\sum_{t=1}^{T} X_t'X_t\right)^{-1} \frac{1}{T-h}\sum_{t=h+1}^{T}\sum_{s=1}^{h} X_{t-s}'r_t, \tag{A14}$$

$$= \left(\frac{1}{T}\sum_{t=1}^{T} X_t'X_t\right)^{-1} \frac{1}{T-h}\sum_{\substack{u\in[h+1,T] \\ (u-t)\in[1,h]}}\sum X_t'r_u. \tag{A15}$$

Equations (A13) and (A15) show that the two estimators both sum $X_t'r_u$ for all observations with $u - t \in [1, h]$, and premultiply by the inverse of an estimated covariance matrix for X_t. The only difference is the set of observations that are available at the very end and beginning of the sample. The overlapping estimator skips observations with $t > T - h$ because the full next-h period returns are not available. The Hodrick (1992) estimator skips observations with $u < h + 1$ because the full last-h period sum of X_t is not available. As a result, Equations (A13) and (A15) each

Table A1
Size, power, and data mining in simulations

A. No predictability null (b = 0)

	dp		VRP	
	OLS	WLS-EV	OLS	WLS-EV
Mean $\hat{b}$	0.000	0.000	0.001	0.001
Median $\hat{b}$	0.001	0.000	0.001	0.001
Standard dev $\hat{b}$	0.423	0.289	0.246	0.162
Mean Asy SE $\hat{b}$	0.403	0.283	0.236	0.161
Prob(Asy p-val < 10%) (%)	11.5	11.1	11.8	10.4
Prob(Asy p-val < 5%) (%)	5.7	5.9	6.3	5.3
Prob(Asy p-val < 1%) (%)	1.2	1.4	1.4	1.1
90th percentile $\lvert\hat{b}\rvert$	0.697	0.476	0.404	0.266
95th percentile $\lvert\hat{b}\rvert$	0.829	0.566	0.480	0.317
99th percentile $\lvert\hat{b}\rvert$	1.088	0.745	0.634	0.418
Mean OOS R^2 (%)	–0.70	–0.31		
Standard dev OOS R^2 (%)	0.87	0.30		
Prob(OOS $R^2 > 0$) (%)	9.49	10.12		
90th percentile OOS R^2 (%)	–0.02	0.00		
95th percentile OOS R^2 (%)	0.20	0.17		
99th percentile OOS R^2 (%)	0.72	0.56		

B. Predictability null (b > 0)

	dp (b = 1)		VRP (b = 0.4)	
	OLS	WLS-EV	OLS	WLS-EV
Mean $\hat{b}$	0.999	0.999	0.401	0.401
Median $\hat{b}$	0.998	0.999	0.401	0.401
Standard dev $\hat{b}$	0.423	0.290	0.246	0.162
Mean Asy SE $\hat{b}$	0.403	0.283	0.236	0.161
Prob(Sim p-val < 10%) (%)	76.2	96.4	49.2	79.6
Prob(Sim p-val < 5%) (%)	65.6	93.2	37.2	69.7
Prob(Sim p-val < 1%) (%)	41.2	81.2	17.3	45.5
Mean OOS R^2	0.66	0.89		
Prob(OOS R^2 sim p-val < 10%) (%)	65.1	77.4		
Prob(OOS R^2 sim p-val < 5%) (%)	57.1	70.2		
Prob(OOS R^2 sim p-val < 1%) (%)	38.1	51.6		

C. Data mining under no predictability null (b = 0)

	dp		VRP	
	OLS	WLS-EV	OLS	WLS-EV
All Samples				
Mean $\lvert\hat{b}\rvert$	0.338	0.231	0.196	0.129
Prob(Asy p-value < 5%) (%)	5.7	5.9	6.3	5.3
Samples with OLS Asy p-value < 5%				
Mean $\lvert\hat{b}\rvert$	0.904	0.520	0.532	0.270
Prob(Asy p-value < 5%) (%)	100.0	42.9	100.0	34.7

(continued)

Table A1
Continued

C. Data mining under no predictability null (b = 0)

	dp		VRP	
	OLS	WLS-EV	OLS	WLS-EV
Samples with WLS-EV Asy p-value < 5%				
Mean $\lvert\hat{b}\rvert$	0.686	0.649	0.402	0.369
Prob(Asy p-value < 5%) (%)	42.0	100.0	40.8	100.0

This table presents results of simulations in which I generate returns according to

$$r_{t+1}^{\text{sim}} = \mu_r + b \cdot x_t^{\text{data}} + \sqrt{RV_{t+1}^{\text{data}}}\psi_{t+1}^{\text{resampled}},$$

where μ_r is the in-sample mean of r_t; x_t^{data} is a return predictor from the data; RV_t^{data} is the intraperiod realized variance; and $\psi_{t+1}^{\text{resampled}}$ is randomly resampled in each simulation. In panel A, I simulate under the no-predictability null that $b = 0$. In panel B, I set $b = 1$ or $b = 0.4$. For dividend-to-price (dp) ratio simulations, I express returns as percentages and use 1,062 monthly observations from 1927 to 2015 of $x_t = dp_t$, the log dividend-to-price ratio of the market portfolio, along with RV $\hat{\sigma}_t^2$ for WLS-EV. For the variance risk premium (VRP) simulations, I express returns as annualized percentages and use 6,552 daily observations from 1990 to 2015 of $x_t = \widehat{\text{VRP}}_t$, Drechsler and Yaron (2011) variance risk premium detailed in Section 3, along with VIXF $\hat{\sigma}_t^2$ for WLS-EV. For each simulated return series, I compute point estimates ($\hat{b}$), asymptotic standard errors (Asy SE $\hat{b}$) and p-values (Asy p-val) using Newey and West (1987), simulated p-values (Sim p-val), and out-of-sample R^2 (OOS R^2) using both OLS and WLS-EV as detailed in Section 1. Panels A and B describe the distribution of these statistics all simulated samples, and panel C focuses on simulated samples with OLS or WLS-EV false positives only.

sum a total of $h(T-h)$ terms, with all but $\frac{h(h-1)}{2}$ terms shared with the other approach ($\frac{h-1}{2(T-h)}$ of the total). When $T >> h$, as is the case in the settings I study, this difference is small and $\hat{\beta}_{\text{OLS}}^{\text{hodrick}}$ is approximately equal to $\hat{\beta}_{\text{OLS}}^{\text{overlap}}$.

In the Online Appendix, I show empirically that $\hat{\beta}_{\text{OLS}}^{\text{hodrick}}$ is approximately equal to $\hat{\beta}_{\text{OLS}}^{\text{overlap}}$ in the settings I study, with differences of at most 8%. Using simulations, I also show that the Hodrick (1992) approach results in more accurate standard errors for both OLS and WLS-EV and makes WLS-EV substantially more efficient. Therefore, I use the Hodrick (1992) approach to address overlapping observations throughout the paper.

A.2. When Is WLS-EV the Most Efficient Linear Unbiased Estimator?

The estimator in Equation (2) is the asymptotically most efficient linear unbiased estimator (GLS) if and only if

$$\text{Var}(\epsilon_{t+1}) = \hat{\sigma}_t^2, \text{ and} \tag{A16}$$

$$\text{Cov}(\epsilon_s, \epsilon_t) = 0 \quad \forall s \neq t. \tag{A17}$$

The condition in Equation (A17) requires that any autocorrelation in returns arises through the X_t variables, making unexpected returns uncorrelated at any lag. My Hodrick-style (1992)

approach to addressing overlapping observations only estimates WLS-EV in a nonoverlapping regression, as described above, assuring there is no overlap-driven autocorrelation in ϵ_t. Rational asset pricing models predict that, given the right vector of X_t variables, nonoverlapping returns satisfy (A17). However, because most of my regressions are univariate, I expect some autocorrelation in returns. Rather than making the additional assumptions needed to incorporate this autocorrelation into a complete covariance matrix of errors and using GLS, I ignore autocorrelation in calculating point estimates and use WLS-EV, but address it when making inferences by using Newey and West (1987) standard errors.

The condition in Equation (A16) requires that the $\hat{\sigma}_t^2$ used empirically are the true variances for future unexpected returns. Since the true $\text{Var}(\epsilon_{t+1})$ are unobservable, I strive to find ex ante proxies that are as accurate as possible. Because these proxies are imperfect measures of the true conditional return variance, it is unlikely that WLS-EV is the perfectly efficient GLS estimator in practice. However, simulation evidence in my setting and in other weighted least-squares settings (e.g., Romano and Wolf 2017) show WLS nevertheless generates substantial efficiency gains relative to OLS. I also use heteroscedasticity-consistent standard errors to address any heteroscedasticity remaining due to imperfect ex ante variance proxies.

A.3. Weighting by Ex Post Measures of Variance

A natural alternative to weighting be ex ante variance predictors is weighting by ex post variance estimates. One example is "robust least squares" (RLS) estimates (e.g., Drechsler and Yaron 2011), which weight observations using a function of estimated $|\epsilon_{t+1}|$. Observations with larger $|\epsilon_{t+1}|$ presumably also have more volatile ϵ_{t+1}, on average, and therefore receive smaller weights. These weights use information from the period returns are realized, $t+1$, rather than the ex ante measures available at time t I use in WLS-EV. The advantage of using time $t+1$ information is that it can provide more accurate estimates of $\text{Var}(\epsilon_{t+1})$.

However, using time $t+1$ information comes with a critical disadvantage: the strong correlation between realized variance and the directional realization of ϵ_{t+1} biases the coefficient estimates. Because negative returns are more volatile than positive returns, negative ϵ have larger variance and smaller weights than positive ϵ. As a result, when the predictor X_t is positively (negatively) correlated with return variance, the coefficient estimated with RLS or any ex post weighting scheme will be biased upward (downward). Therefore, it is unsurprising, given the variance risk premium is positively correlated with return variance, that Drechsler and Yaron (2011) finds RLS coefficients are more positive than OLS coefficients. The WLS-EV approach avoids this mechanical connection between weights and ϵ_{t+1} by using ex ante variance weights.

Appendix B. Simulations

In this appendix, I describe the heteroscedastic small-sample simulations I use to compute standard errors and p-values throughout the paper. These simulations also provide further insight into OLS and WLS-EV's size, power, and potential to detect data mining.

B.1. Simulation Procedure

The efficiency and small-sample biases of each estimation procedure depend critically on variability of return variance, the asymmetry in the return distribution, the time-series distribution of the predictor, and the correlations among these variables. Rather than attempting to model these distributions, I use a variation of the "wild bootstrapping" technique often employed to generate heteroscedastic samples. (See Liu (1988), Mammen (1993), and Davidson and Flachaire (2008) for details.)

My simulation procedure begins by taking observed excess returns r_t^{data} and computing the standardized next-period return for each observation:

$$\psi_t^{\text{data}} \equiv \frac{r_t^{\text{data}} - \mu_r}{\sqrt{RV_t^{\text{data}}}}, \tag{B1}$$

where RV_t^{data} is the observed realized intra-period variance, and μ_r is chosen so that $\mathbb{E}(\psi_{t+1}^{\text{data}}) = 0$. I then create 100,000 simulated samples by re-sampling the ψ_t^{data} (with replacement) and computing the next-period returns as follows:

$$r_{t+1}^{\text{sim}} = \mu_r + b \cdot x_t^{\text{data}} + \sqrt{RV_t^{\text{data}}}\psi_{t+1}^{\text{re-sampled}}, \tag{B2}$$

where x_t^{data} are the observed values of a predictor variable, and I specify the population prediction coefficient b.[14] These simulated returns inherit the skewness, any heteroscedasticity not captured by RV_t^{data}, and other properties of the observed return distribution while still having conditional mean $\mu_r + b \cdot x_t^{\text{data}}$ and variance RV_t^{data}.

For each simulated return sample, I regress the redrawn returns r_{t+1}^{sim} on x_t^{data} and a constant using both OLS and WLS-EV, where WLS-EV weights are as described in Section 1, and record the resultant coefficients ($\hat{b}^{\text{sim}}$) and HAC standard errors (SE $\hat{b}^{\text{sim}}$). Note that, in this simulation procedure, the weights I use for WLS-EV do not equal the true variances of returns, and instead only forecast $\text{RV}_{t+1}^{\text{data}}$ as well as they do in observed data. Therefore, WLS-EV only results in more efficient estimates to the extent that the ex ante weights I use predict future realized variance and do not introduce additional noise or bias.

B.2. Simulated Standard Errors and *p*-Values

The simulated standard errors and *p*-values I use are based on the simulations described above. Simulated standard errors for in-sample tests equal the standard deviation across simulated samples of the point estimates $\hat{b}^{\text{sim}}$. Simulated *p*-values for in-sample tests are two-sided and equal the fraction of simulated samples with $|\hat{b}^{\text{sim}}| > |\hat{b}^{\text{data}}|$.

I also use these simulations to assess the statistical significance of out-of-sample R^2 in Table 3. To do so, for each simulated sample I repeat the out-of-sample forecasting exercise described in Section 2.2, using past data only to estimate predictability coefficients, and applying these past-only coefficients to the present values of x to compute an out-of-sample forecast for each time period. Following this procedure yields an OOS R^2 for each simulated sample, which I use to compute a one-sided *p*-value for the out-of-sample tests as the fraction of simulated samples with Sim OOS $R^2 >$ Data OOS R^2.

B.3. Test Size and Efficiency

To assess the size and efficiency of OLS and WLS-EV tests, I first implement my simulation procedure on a monthly sample from 1927 to 2015 using the log dividend-to-price (dp) ratio as the predictor and RV $\hat{\sigma}_m^2$ for the WLS-EV estimates. Panel A of Table A1 shows summary statistics for these simulations under the no-predictability null $b = 0$. Using WLS-EV rather than OLS results in large efficiency gains, reducing the standard deviation of $\hat{b}$ from 0.423 to 0.289, a 32% decrease. Both estimators are unbiased in these simulations, with mean values of 0. Here, I do not find a Stambaugh (1999) bias, because the redrawn standardized returns are uncorrelated with innovations in the dp ratio. I correct for the Stambaugh (1999) bias throughout the paper using the procedure described in Section B.5.

[14] For bootstrapped standard errors, I specify the null $b = 0$. In this appendix, I also study simulations with $b > 0$.

Given the true b is zero, an effective estimator rejects the $b = 0$ null (a "false positive") as infrequently as possible. An estimator would reject with a C% critical value in more than C% of simulations for two reasons: a directional bias in the average $\hat{b}$ or a downward bias in asymptotic standard errors. In addition to having unbiased $\hat{b}$, both OLS and WLS-EV have average standard errors only slightly smaller than the standard deviation of $\hat{b}$, indicating that the asymptotic HAC standard errors are quite accurate for the dp ratio in this sample and can account for the remaining heteroscedasticity caused by imperfect variance weights (as suggested by Romano and Wolf 2017). As a result, OLS and WLS-EV t-tests reject the null at the 5% level in 5.7% and 5.9% of simulations, respectively.

Of course, even with a relatively small false-positive rate, it is important to use simulated standard errors that, by construction, reject the null in exactly 5% of simulated samples. As described above, I do so using two-sided tests comparing $|\hat{b}|$ in the data to the distribution of $|\hat{b}|$ in simulations. I summarize this distribution in panel A by showing the 90th, 95th, and 99th percentiles of $|\hat{b}|$, which serve as the critical values for my tests. Because WLS-EV is more efficient, its critical values for rejecting the no-predictability null are smaller, allowing WLS-EV to reject in many cases that OLS fails to reject.

An alternative test statistic for return predictability I consider in Section 2 is OOS R^2. I illustrate the properties of this statistic using the same simulation procedure. As discussed in Goyal and Welch (2008) and elsewhere, under the no-predictability null OOS R^2 will be negative, on average, due to overfitting associated with estimation error. Panel A of Table A1 confirms this is the case in my simulations, with negative mean OOS R^2 for both OLS and WLS-EV. The more-efficient WLS-EV limits estimation error, resulting in a less-negative mean OOS R^2. Estimation error affects the variance across simulations of OOS R^2 as well as its mean, with OLS OOS R^2 being substantially more volatile. These two effects offset, making OLS and WLS-EV OOS R^2 about equally likely to be positive, and giving them similar 90th and 95th percentiles under the null. These percentiles serve as critical values for the statistical tests I use in Table 3.

Given the true b is nonzero, an effective estimator fails to reject the $b = 0$ null (a "false negative") as little as possible. To assess the frequency of false negatives, I repeat the simulation exercise assuming $b = 1$. For each simulated sample, I compute the small-sample p-value based on the distribution of $\hat{b}$ for each estimator under the no-predictability null. A false negative is defined by a failure to reject the no-predictability null.

The main result in panel B is that false negatives are much less likely for WLS-EV than OLS both in- and out-of-sample, illustrating one benefit of more efficient point estimates. For OLS, in-sample p-values are less than 5% in 65.6% of simulations, meaning the false negative rate is 34.4%, much higher than the false negative rate for WLS-EV of 6.8%. Out-of-sample p-values tell a similar story, with OLS having a false negative rate of 42.9% while WLS-EV's rate is 29.8%.[15] False negatives occur less frequently for WLS-EV because the added efficiency allows for smaller standard errors, resulting in sharper inferences.

To assess WLS-EV in shorter samples and using a predictor more directly related to ex ante variance, I also implement this simulation procedure on an overlapping daily sample from 1990 to 2015, using the variance risk premium (VRP) proxy defined in Drechsler and Yaron (2011) as to predict next-month returns. The results are in the second column of Table A1. The conclusions are largely the same as for the dp ratio, but the effects are stronger because VRP are more strongly correlated with ex ante volatility than the dp ratio. As a result, WLS-EV estimates are 34% more efficient than OLS estimates. WLS-EV estimates are also less susceptible to false positives, with asymptotic t-stats rejecting the null at the 5% level in 5.3% of samples, compared to 6.3% of samples for OLS. Finally, I assess the false negative rate of the three estimators in the VRP setting with $b = 0.4$. Mirroring the results in panel B for the dp ratio, WLS-EV has false negatives in

[15] As emphasized in Cochrane (2008), out-of-sample tests are less powerful than in-sample tests, resulting in more false negatives for a given estimator.

around 30.3% of simulations, compared to 62.8% for OLS. Given the shorter sample period, I do not use out-of-sample tests for VRP.

B.4. Data Mining

Imagine a researcher tried different predictors until they found one that had a sufficiently low *p*-value. If predictors were selected via this form of data mining but the true correlation with future returns was zero, another estimator might be useful as a partially independent test of the same null hypothesis that may fail to reject using the same data.

To illustrate this point, I use the same simulation procedure described above with the $b = 0$ null hypothesis but focus only on samples where the asymptotic OLS or WLS-EV *p*-value is below 5%. For this false-positive sample, panel C of Table A1 shows the average absolute point estimates $|\hat{b}|$ and the probability the asymptotic *p*-value is less than 5% for both OLS and WLS-EV.

I find that for samples selected for their OLS significance, WLS-EV point estimates are 42% and 49% closer to zero, on average, than OLS estimates for the dp and VRP samples, respectively. This is consistent with the magnitudes of the decline in point estimates I find for both the variance risk premium and the Novy-Marx (2014) predictors. Furthermore, only 43% and 35% of OLS false-positive samples have statistically significant WLS-EV estimates for the dp ratio and VRP, respectively, indicating WLS-EV will often fail to reject when there is no true return predictability and the OLS evidence is due to a false positive.

These results do not make WLS-EV immune to data mining. If, instead of OLS-based data mining, samples were selected based on WLS-EV significance, I find that only 42% and 41% of false-positive samples have statistically significant OLS estimates for the dp ratio and VRP, respectively. This pattern illustrates how *any* partially independent estimator is a useful diagnostic when one suspects a predictor may have been data mined for significance using a different estimator. In practice, though, any data mining used to select the body of predictors in the literature was based on OLS rather than WLS-EV, making WLS-EV useful in revisiting return predictability evidence.

B.5. Stambaugh (1999) Bias Correction

I account for the small-sample bias described in Stambaugh (1999) by simulating both the x_t and subsequent returns r_{t+1} under the no-predictability null, as suggested in Goyal and Welch (2008). I generate r_{t+1} and x_t using the following processes:

$$r_{t+1}^{\text{sim}} = \mu_r + \epsilon_{t+1}^{\text{re-sampled}} \tag{B3}$$

$$x_{t+1}^{\text{sim}} - \mu_x = \rho_x (x_t^{\text{sim}} - \mu_x) + \delta_{t+1}^{\text{re-sampled}} \tag{B4}$$

where μ_r, μ_x, and ρ_x are estimated from the data for the predictor in question, and x_0 is chosen from a random observation. To preserve the correlation between innovations in r and x, I jointly re-sample (with replacement) the vector $[\,\epsilon_{t+1} \quad \delta_{t+1}\,]'$ from the innovations estimated in the data. For each simulated sample, I estimate the $\hat{b}$ using OLS and estimate the Stambaugh (1999) bias estimate as the cross-simulation average value of $\hat{b}$. I subtract this bias estimate from both OLS and WLS-EV point estimates throughout the paper to compute bias-corrected "Stambaugh $\hat{b}$."

B.6. Ferson, Sarkissian, and Simin (2003) Simulations

When studying the presidential puzzle in Section 4.1, I consider alternative standard errors reflecting the bias in predictability regressions posited in Ferson, Sarkissian, and Simin (2003). This bias occurs when expected returns are time-varying, persistent, and unrelated to the predictor

in question. I incorporate this possibility into the standard simulation framework described above by generating returns using:

$$r_{t+1}^{\text{sim}} = \mu_{r,t}^{\text{sim}} + b \cdot x_t^{\text{data}} + \sqrt{RV_t^{\text{data}}}\psi_{t+1}^{\text{re-sampled}}. \tag{B5}$$

The only difference between these simulations and the approach summarized by Equation (B2) is the time-varying conditional mean $\mu_{r,t}^{\text{sim}}$.

Following Ferson, Sarkissian, and Simin (2003), I simulate $\mu_{r,t}^{\text{sim}}$ according to the following AR(1) process:

$$\mu_{r,t}^{\text{sim}} = \mu_r + \rho_\mu(\mu_{r,t-1}^{\text{sim}} - \mu_r) + \gamma_t, \tag{B6}$$

where $\gamma_t \sim N(0, \sigma_\gamma^2)$. I use the sample average return for μ_r, the sample autocorrelation of the presidential dummy (0.983) for ρ_μ, and the value of σ_γ that makes $\mu_{r,t}$ negative in 10% of observations (0.07%). For this value of σ_γ, the interquartile range of monthly $\mu_{r,t}$ is [0.2%, 0.7%] ([2.8%, 9.3%] annualized) and the 90% confidence interval is $[-0.1\%, 1.1\%]$ ($[-1.7\%, 14.3\%]$ annualized). I view this as an upper bound on reasonable variations in equilibrium expected returns, and therefore view my simulations as an upper bound on the plausible magnitude of the Ferson, Sarkissian, and Simin (2003) effect in this setting.

References

Ang, A., and G. Bekaert. 2006. Stock return predictability: Is it there? *Review of Financial Studies* 20:651–707.

Bollerslev, T., J. Marrone, L. Xu, and H. Zhou. 2014. Stock return predictability and variance risk premia. Statistical inference and international evidence. *Journal of Financial and Quantitative Analysis* 49:633–61.

Bollerslev, T., G. Tauchen, and H. Zhou. 2009. Expected stock returns and variance risk premia. *Review of Financial Studies* 22:4463–92.

Boudoukh, J., R. Michaely, M. Richardson, and M. R. Roberts. 2007. On the importance of measuring payout yield: Implications for empirical asset pricing. *Journal of Finance* 62:877–915.

Brenner, R. J., R. H. Harjes, and K. F. Kroner. 1996. Another look at models of the short-term interest rate. *Journal of Financial and Quantitative Analysis* 31:85–107.

Campbell, J. Y. 1987. Stock returns and the term structure. *Journal of Financial Economics* 18:373–99.

Campbell, J. Y., S. Giglio, C. Polk, and R. Turley. 2018. An intertemporal capm with stochastic volatility. *Journal of Financial Economics* 128:207–33.

Campbell, J. Y., and S. B. Thompson. 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies* 21:1509–31.

Cochrane, J. H. 2008. The dog that did not bark: A defense of return predictability. *Review of Financial Studies* 21:1533–75.

Davidson, R., and E. Flachaire. 2008. The wild bootstrap, tamed at last. *Journal of Econometrics* 146:162–9.

Drechsler, I., and A. Yaron. 2011. What's vol got to do with it. *Review of Financial Studies* 24:1–45.

Engle, R. F., D. M. Lilien, and R. P. Robins. 1987. Estimating time varying risk premia in the term structure: The arch-m model. *Econometrica* 55:391–407.

Fama, E. F., and K. R. French. 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics* 25:23–49.

Fama, E. F., and G. W. Schwert. 1977. Asset returns and inflation. *Journal of Financial Economics* 5:115–46.

Ferson, W. E., S. Sarkissian, and T. T. Simin. 2003. Spurious regressions in financial economics? *Journal of Finance* 58:1393–413.

French, K. R., G. W. Schwert, and R. F. Stambaugh. 1987. Expected stock returns and volatility. *Journal of Financial Economics* 19:3–29.

Ghysels, E., P. Santa-Clara, and R. Valkanov. 2005. There is a risk-return trade-off after all. *Journal of Financial Economics* 76:509–48.

Glosten, L. R., R. Jagannathan, and D. E. Runkle. 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* 48:1779–801.

Goyal, A., and I. Welch. 2008. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21:1455–508.

Hodrick, R. J. 1992. Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *Review of Financial Studies* 5:357–86.

Johannes, M., A. Korteweg, and N. Polson. 2014. Sequential learning, predictability, and optimal portfolio returns. *Journal of Finance* 3:161–74.

Keim, D. B., and R. F. Stambaugh. 1986. Predicting returns in the stock and bond markets. *Journal of Financial Economics* 17:357–90.

Kothari, S. P., and J. Shanken. 1997. Book-to-market, dividend yield, and expected market returns: A time-series analysis. *Journal of Financial Economics* 44:169–203.

Lamont, O. 1998. Earnings and expected returns. *Journal of Finance* 53:1563–87.

Lettau, M., and S. Ludvigson. 2001. Consumption, aggregate wealth, and expected stock returns. *Journal of Finance* 56:815–49.

Lettau, M., and S. Van Nieuwerburgh. 2008. Reconciling the return predictability evidence. *Review of Financial Studies* 21:1607–52.

Lintner, J. 1975. Inflation and security returns. *Journal of Finance* 30:259–80.

Liu, R. Y. 1988. Bootstrap procedures under some non-iid models. *Annals of Statistics* 16:1696–708.

Mammen, E. 1993. Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics* 21:255–85.

Newey, W. K., and K. D. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55:703–8.

Novy-Marx, R. 2014. Predicting anomaly performance with politics, the weather, global warming, sunspots, and the stars. *Journal of Financial Economics* 112:137–46.

Pástor, L., and P. Veronesi. 2017. Political cycles and stock returns. Working paper.

Pettenuzzo, D., A. Timmermann, and R. Valkanov. 2014. Forecasting stock returns under economic constraints. *Journal of Financial Economics* 114:517–53.

Polk, C., S. Thompson, and T. Vuolteenaho. 2006. Cross-sectional forecasts of the equity premium. *Journal of Financial Economics* 81:101–41.

Pontiff, J., and L. D. Schall. 1998. Book-to-market ratios as predictors of market returns. *Journal of Financial Economics* 49:141–60.

Powell, J. G., J. Shi, T. Smith, and R. E. Whaley. 2007. The persistent presidential dummy. *Journal of Portfolio Management* 33:133–43.

Romano, J. P., and M. Wolf. 2017. Resurrecting weighted least squares. *Journal of Econometrics* 197:1–19.

Rozeff, M. S. 1984. Dividend yields are equity risk premiums. *Journal of Portfolio Management* 11:68–75.

Santa-Clara, P., and R. Valkanov. 2003. The presidential puzzle: Political cycles and the stock market. *Journal of Finance* 58:1841–72.

Shiller, R. J., S. Fischer, and B. M. Friedman. 1984. Stock prices and social dynamics. *Brookings Papers on Economic Activity* 1984:457–510.

Singleton, K. J. 2006. *Empirical dynamic asset pricing: model specification and econometric assessment*. Princeton, NJ: Princeton University Press.

Stambaugh, R. F. 1999. Predictive regressions. *Journal of Financial Economics* 54:375–421.

Westerlund, J., and P. Narayan. 2014. Testing for predictability in conditionally heteroskedastic stock returns. *Journal of Financial Econometrics* 13:342–75.

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48:817–38.

Wooldridge, J. M. 2010. *Econometric analysis of cross section and panel data*. Cambridge: MIT Press.