CFR (draft), XXXX, XX: 1–15

Dissecting market expectations in the cross-section of book-to-market ratios

Thiago de Oliveira Souza

[1]*University of Southern Denmark, Department of Business and Economics, Campusvej 55, DK-5230 Odense M; tsouza@sam.sdu.dk*

ABSTRACT

This paper starts by successfully replicating all the main results in Kelly and Pruitt (2013) for the return on the market—and providing some evidence of market premium predictability—based on their original empirical choices in the 1930-2010 sample. However, the evidence of market premium predictability, in particular, essentially disappears by making any one of the following changes: (i) Updating the sample to June 1926 – December 2019; (ii) not taking logs of the book-to-markets used as regressors; (iii) not dividing book-to-markets by their time-series standard deviations; or (iv) not taking one extra book-to-market lag (for monthly forecasts). In summary, I find no evidence that the procedure generates a valid forecasting model of market premiums with persistently positive out-of-sample R^2 in the full 1926-2019 sample, especially since the Oil Shock (or early 2000).

Keywords: Predictability, out-of-sample, equity premium, disaggregated book-to-markets

JEL Codes: G11, G12, G14

I thank Ivo Welch and the anonymous referee (among other things, for explaining the construction of monthly book-to-markets based on the data of Kenneth French). I also thank Seth Pruitt and Bryan Kelly for sharing their estimation code and data.

ISSN 2164-5744 ; DOI 10.1561/104.XXXXXXXX

Welch and Goyal (2008) argue that in-sample (IS) predictability is insufficient to validate forecasting models of the market premium. They propose a higher hurdle: In summary, apart from being significant in sample, the model should have out-of-sample (OOS) performance that is positive over the entire sample period, especially towards the end, since the Oil Shock. Conditioned on this hurdle and after analyzing the performance of all typical predictors in the literature, they essentially conclude that no valid forecasting model of the market premium exists. The present paper adds to this discussion by investigating if the partial least squares (PLS) forecasts based on disaggregated book-to-market (BM) ratios in Kelly and Pruitt (2013) (henceforth, KP) finally overcome the hurdle of Welch and Goyal (2008). The short answer, based on the most recent evidence, is no.

Before reaching this conclusion, I successfully replicate the main results of KP for the return on the market—and also find some evidence of market premium predictability—in their original 1930-2010 sample and based on their other empirical choices.[1] I also find some evidence of predictability based on different empirical choices. However, the evidence of market premium predictability, in particular, essentially disappears by making any one of the following changes: (i) Updating the sample to June 1926 – December 2019; (ii) failing to take logs of the BMs used as regressors; (iii) failing to divide (log) BMs by their time-series standard deviations, as I explain later; or (iv) failing to take one extra BM lag (for monthly forecasts). As I will show, the PLS procedure of KP fails to generate a valid forecasting model of the equity premium especially due to the poor performance of these forecasts since the Oil Shock and for the last 20 years in the full sample.

The paper proceeds as follows. Sections 1 and 2 briefly describe the PLS procedure and data, respectively. Section 3 replicates KP in their original 1930-2010 sample, revealing the importance of their explicit choices of using BMs in logs as regressors and market returns as the forecasting target, and their less explicit choice of dividing BMs by their time-series standard deviations prior to the estimation. Section 4 extends the sample to 1926-2019, reveals the importance of the sample period and OOS split date choices for all forecasts, and the importance of the extra BM lag for the

[1] In fact, Seth Pruitt kindly shared the Matlab code in KP, market returns, and monthly BMs of 6, 25, and 100 portfolios sorted by market capitalization (ME) and BM that they create. Hence, I can also obtain their results exactly by running their code on their data.

forecasts in monthly frequency. Section 5 concludes.

1 PLS estimation

This section briefly presents the PLS procedure of KP and highlights a few points that can be unclear in their paper. I refer readers to KP for other details. Essentially, the "axiomatic" theoretical claim in KP is that a single factor, F_t, drives (one-period) expected log returns, $\mu_{i,t}$, and also drives the expected return on equity (ROE), $g_{i,t}$, of all individual assets i,[2]

$$\mu_{i,t} = \mathrm{E}_t\left[r_{i,t+1}\right] = \gamma_{i,0} + \gamma_i F_t, \tag{1}$$

$$g_{i,t} = \mathrm{E}_t\left[roe_{i,t+1}\right] = \delta_{i,0} + \delta_i F_t + \epsilon_{i,t}. \tag{2}$$

Next, based on Vuolteenaho (2002), KP manipulate these equations into

$$bm_{i,t} = \Phi_{i,0} + \Phi_i F_t + \epsilon_{i,t}, \tag{3}$$

where $bm_{i,t}$ is the log BM ratio of asset i. Based on many assets, KP argue that the expected return on the market should be estimated via the three ordinary least squares regressions that follow:

1. Run a time-series regression of the (log) BMs of each asset i, $bm_{i,t}$, on the (future) return on the market (their forecasting target),

$$bm_{i,t} = \hat{\Phi}_{i,0} + \hat{\Phi}_i y_{t+h} + e_{i,t}, \tag{4}$$

   where y_{t+h} is the future market return at frequency h (1-month or 1-year).

2. Run a cross-sectional regression, for each period, t, of all BMs (at once) on the loadings estimated in the first stage, $\hat{\Phi}_i$, in Eq. (4),

$$bm_{i,t} = \hat{c}_t + \hat{F}_t \hat{\Phi}_i + \omega_{i,t}, \tag{5}$$

   which results in an estimate for the latent factor, F_t, for each period.

3. Run a standard predictive regression of returns on the factor estimated in the second stage, $\hat{F}_t$,

$$y_{t+h} = \beta_0 + \beta \hat{F}_t + u_{t+h}. \tag{6}$$

[2]In fact, KP start with the assumption of a $(K_F \times 1)$ vector of factors, $\boldsymbol{F}_t$. Hence, the model generalizes to this setting. They also first state the model in terms of expected **cash flow** growth (not ROE). But the accounting relation in Vuolteenaho (2002)—which KP use exactly to explain their choice of regressors—is based on ROE (not on cash flow growth).

BM standardization:
An important point not explicitly mentioned by KP is that they do not run the exact linear regressions above. Instead, KP divide each BM by its respective time-series standard deviation, σ_{vi}, before running the regressions.[3] The estimated counterparts of Eq. (4) and Eq. (5) are

$$\frac{bm_{i,t}}{\sigma_{vi}} = \hat{\Phi}_{si,0} + \hat{\Phi}_{si} y_{t+h} + e_{si,t}, \quad (7)$$

$$\frac{bm_{i,t}}{\sigma_{vi}} = \hat{c}_{s,t} + \hat{F}_{s,t} \hat{\Phi}_{si} + \omega_{si,t}, \quad (8)$$

while $\hat{F}_{s,t}$ substitutes $\hat{F}_t$ as the (standardized) estimate of the latent factor, F_t, in Eq. (6). For the recursive estimation, σ_{vi} becomes time-varying (because the samples change over time): The value is calculated recursively, over the sample that is inside the information set when the forecast is created.[4] From a strict asset pricing perspective, this adjustment implies that the empirical results of KP rely on a possible theoretical linear relation between the latent factor and *ratios between BMs and standard deviations*, not exactly BMs (which is what the theories mentioned by KP seem to imply).

2 Data

All data are available from June 1926 to December 2019. The market return (RM) is the monthly return on the Center for Research in Security Prices index portfolio. The market premium (MP) subtracts from this value the risk-free rate from Kenneth French data library. I transform both to continuous compounding to match KP.

The main forecasts in KP use monthly BMs of portfolios formed by market capitalization (ME) and BM as regressors. The book equity (BE)

[3] The procedure is not mentioned by KP, but it is implemented in their Matlab code and the regressors are assumed to be standardized in Kelly and Pruitt (2015).

[4] In any training sample, each BM is scaled by its sample volatility in that training sample. The scaling variable for each BM series is constant in any given sample and only changes every period because the sample changes every period (yielding a new σ_{vi}). The fact that σ_{vi} updates over time is a peculiar feature of the OOS procedure. In addition, I include BMs from months $t-11$ to $t-1$ in this calculation for annual forecasts, although these BMs are never used in the estimation and it is not obvious that they should be included in the calculation.

and ME data necessary to calculate these BMs are indirectly available from Kenneth French's website:[5] Kenneth French reports average ME and the number of stocks in each double sorted portfolio by ME and BM in monthly frequency. Their product is the (total) ME of the ME/BM portfolio. However, the data follow a beginning-of-period convention: ME in July is actually for (the end of) June. There are also two versions of monthly value-weighted BMs in the same spreadsheet (with the same timing convention). One version scales the portfolio BE by its ME in June. Hence, every July (and only in July), this series coincides with the BM of the portfolio (and given the timing convention, this is the BM for June). Therefore, I reverse the calculation by multiplying the calculated ME by the BM for June (both reported as July) to obtain the total BE of the portfolio for June, which is fixed for a year. Next, the procedure is the one described by KP: Divide BE (which changes yearly) by the ME (changing monthly) for the remaining months to obtain monthly BMs.

BMs with one extra lag:
In Section 4, I show that taking one extra lag of the BMs improves the performance of the monthly estimates, while having negligible impact at the annual frequency.[6] Hence, I use BMs with one extra lag in the entire paper, unless stated otherwise. This is equivalent to assume that the BM data follow an end-of-period timing convention, instead of the one that I describe above.

3 PLS and some empirical choices in the original sample

In addition to replicating KP in their original 1930-2010 sample and with all their empirical choices, this section shows that three of these choices improve the PLS performance: (i) The choice of forecasting the return on the market, as opposed to the market premium; (ii) the choice of using standardized instead of regular BMs; and (iii) the choice of using BMs in logs and not in levels. In the entire paper, I also ignore certain series of

[5] The Python code on my website that generates all results in the paper also implements this calculation.

[6] Although unreported in this version, the extra lag also makes the performance more similar to the one in KP, based on their original empirical choices.

BMs with more than 20% missing data in a given subsample.[7] Table 1 summarizes the performances of 6 groups of models based on these choices: It starts with the choices in KP for market returns at the top of the table, relaxes (ii) and (iii) (one at a time) in the next two groups, and repeats these choices for the market premium in the last three (MP) groups.

Table 1 has the same fields as the main table (Table I) in KP, but without the extra p-values. It reports in-sample R^2 (IS_y and IS_m, for yearly and monthly returns) and OOS R^2 (OOS_y and OOS_m) for the third-step PLS regression in Eq. (6); the respective in-sample p-values of the β coefficients with Newey and West (1987) standard errors (12 lags or one lag, for annual or monthly returns, respectively); and—if the OOS R^2 is positive—the respective p-values of the ENC-NEW forecast encompassing test of Clark and McCracken (2001) with Newey-West standard errors with 12 lags for annual returns.[8] The main restriction in interpreting the values in this table is that the OOS split is always January 1980, and the sample ending is always 2010. Section 4 investigates other split and ending dates in 1926-2019.

The first column in Table 1 identifies the forecasting target (RM or MP), the third column shows the number of ME/BM portfolios used (6, 25, or 100), and the second column identifies the BMs used as regressors ($bm_{\sigma,t-1}$, bm_{t-1}, or $BM_{\sigma,t-1}$): In this table, the BM regressors always have one extra lag (with subscript $t-1$). The BMs in logs (in lowercase) are either divided by their time-series standard deviations ($bm_{\sigma,t-1}$) or not (bm_{t-1}). But the standardized BMs (indicated by the subscript σ) can also be in levels ($BM_{\sigma,t-1}$).

3.1 Market return forecasts

The six [RM, $bm_{\sigma,t-1}$] models in the first three rows (in monthly and yearly frequency) closely replicate KP based on their exact empirical choices: All of them deliver positive and significant performances in this case. Four

[7]This restriction is not mentioned by KP, but it is included in their Matlab code. It tends to systematically remove some portfolios with more diverging risk proxies from the regressors, such as extreme deciles of "small-growth" and "big-value", and mostly benefits the recursive performance at the beginning of the sample (there is no sizable effect in the full sample). In addition, the restriction is only biding for 100 ME/BM portfolios (not 25 or 6).

[8]The ENC-NEW is **not** a significance test for the OOS R^2. Hence, even if the OOS R^2 is negative, the statistic is often significant.

Table 1: **Predictability of the 1-year or 1-month return on the market (RM) or market premium (MP) from January 1930 to December 2010.**

			IS_y R^2	p	OOS_y R^2	p	IS_m R^2	p	OOS_m R^2	p
RM	$bm_{\sigma,t-1}$	6	8.53	0.00	6.77	<0.01	0.58	0.03	0.67	<0.05
		25	12.31	0.00	8.06	<0.01	0.89	0.01	0.73	<0.01
		100	19.83	0.00	14.07	<0.01	2.53	0.01	0.63	<0.05
	bm_{t-1}	6	3.57	0.14	5.06	<0.05	0.21	0.39	-0.05	-
		25	5.63	0.09	5.90	<0.05	0.37	0.20	-0.96	-
		100	14.31	0.01	5.69	<0.05	2.25	0.05	-0.61	-
	$BM_{\sigma,t-1}$	6	6.77	0.01	5.70	<0.01	0.27	0.19	0.19	-
		25	19.51	0.00	0.40	<0.01	2.60	0.06	-0.07	-
		100	26.42	0.00	-2.86	-	3.24	0.04	-0.29	-
MP	$bm_{\sigma,t-1}$	6	5.92	0.02	0.73	<0.05	0.46	0.06	0.31	<0.05
		25	9.83	0.01	2.16	<0.01	0.74	0.02	0.09	<0.05
		100	18.14	0.00	9.26	<0.01	2.41	0.03	-0.19	-
	bm_{t-1}	6	6.95	0.04	-10.60	-	0.76	0.01	-0.96	-
		25	14.40	0.00	-20.05	-	0.91	0.00	-2.55	-
		100	19.18	0.00	-7.60	-	2.34	0.04	-1.63	-
	$BM_{\sigma,t-1}$	6	5.52	0.03	1.29	<0.05	0.81	0.05	-1.00	-
		25	19.71	0.00	-5.07	-	2.86	0.09	-0.42	-
		100	27.93	0.00	-7.25	-	3.02	0.06	-0.55	-

Note: The table reports IS and OOS R^2 for yearly (IS_y, OOS_y) or monthly forecasts (IS_m, OOS_m); in-sample p-values of the slope coefficients of predictive regressions with Newey and West (1987) standard errors (12 lags or one lag, for annual or monthly returns, respectively); and p-values of the OOS forecast encompassing test of Clark and McCracken (2001) with Newey-West standard errors with 12 lags for annual returns (unreported, as "-", if the p-value is larger than 0.1 or if the OOS R^2 is negative). The OOS split is January 1980. The regressors are the BMs of the 6, 25, or 100 ME/BM portfolios in lowercase if the BMs are in logs, and with subscript σ if they are divided by their time-series standard deviations. Subscript $t-1$ indicates that BMs have one extra lag.

Interpretation: The first rows [RM, $bm_{\sigma,t-1}$] replicate KP and confirm that the performance is better exactly for (i) the return on the market, (ii) with BMs in logs, (iii) with standardized BMs, and (iv) in annual frequency, as in KP. The lower half of the table shows that market premium predictability essentially disappears if we relax any one of the empirical choices in KP.

models even have slightly larger OOS R^2 in Table 1 than in KP. The other groups of market return models relax one of the empirical choices in KP at a time: The six [RM, bm_{t-1}] models relax the BMs standardization and the six [RM, $BM_{\sigma,t-1}$] models relax the BM log transformation.

As I explain in Section 4, the 1930-2010 sample and 1980 split date are particularly favorable for the PLS forecasts: Although the IS performance is not always significant, even the three annual [RM, bm_{t-1}] models based on BMs in logs but not standardized deliver positive OOS performance in this case. And two of the annual [RM, $BM_{\sigma,t-1}$] models in the subsequent three rows (for 6 and 25 portfolios), based on standardized BMs in levels, also deliver positive OOS performance. On the other hand, none of the monthly models in these other two groups deliver positive OOS performance.

In summary, standardization and especially the log transformation of BMs improve the PLS forecasts of market returns **in this sample**.[9] Indeed, the main conclusion of persistently positive OOS R^2 in KP (based on 100 portfolios) only holds if the BMs are in logs. Otherwise, both monthly and yearly [RM, $BM_{\sigma,t-1}$] models deliver negative OOS R^2.

3.2 Market premium forecasts

The only group of market premium models with mostly positive OOS R^2 is the [MP, $bm_{\sigma,t-1}$] in the first three MP rows. These are the equivalent of the standard models in KP for market returns. However, even in this group, the monthly forecasts based on 100 portfolios have negative OOS R^2, while the other models have positive but often small OOS R^2. With respect to all remaining models in the other groups, none delivers positive OOS R^2, except for the annual [MP, $BM_{\sigma,t-1}$] with 6 portfolios: 1.29% OOS R^2 (significant at 5%).

In summary, if we relax either of the two BM transformations in this table (standardization or logs), the evidence of market premium predictability essentially disappears. In addition, the OOS R^2 of the market premium models in Table 1 are proportional to, but substantially lower than the OOS R^2 of the market return models, for the same empirical choices. This (together with the analysis in Section 4) suggests that the other important empirical choice to obtain the results in KP is to use market returns instead of market premiums as forecasting targets.

[9] A previous version of this paper shows that non-standardized BMs deliver less problematic forecasts after 2010.

4 Extended sample and other empirical choices

This section extends the sample to June 1926 – December 2019 and investigates the importance of three other empirical choices: (i) The sample period, (ii) the OOS split date, and (iii) the use of BMs with one extra lag. I start by reporting the results for the 1926-2019 sample in Table 2, and comparing them with Table 1 for the baseline empirical choices in KP, [RM, $bm_{\sigma,t-1}$].

As I will show, one of the most curious conclusions from Table 2 is that PLS delivers better monthly forecasts when BMs contain one extra lag. The explanation is two-fold: First, monthly innovations in BMs are essentially negative of the previous month return, because BE is fixed 11 months per year. Second, market returns have small positive autocorrelation at one month lag (de Oliveira Souza, 2020). Hence, skipping one month removes the confounding effect of negative and positive prediction. But for returns in annual frequency, this confounding effect does not exist and there is little impact on the forecasts.

Market returns:
As before, the six [RM, $bm_{\sigma,t-1}$] models in the first three rows of Table 2 reflect the baseline empirical choices of KP. All these models have positive OOS R^2, based on the 1980 OOS split of the 1926-2019 sample: The results are weaker, but similar to the ones in Table 1. The [RM, $bm_{\sigma,t}$] models in the next three rows are equivalent to the baseline models, except that the BM regressors do not contain the extra lag. In line with the explanation in the previous paragraph, removing the extra lag has almost no effect on the annual forecasts (which have slightly higher OOS R^2 in this case), while the monthly forecasts have lower OOS R^2. In fact, the only monthly model with non-negative OOS R^2 is based on 6 portfolios in this case.

Market premiums:
The other models at the bottom of Table 2 are the market premium equivalents of the models of market return (at the top). The six [MP, $bm_{\sigma,t-1}$] models reveal that, with respect to the market premium, the OOS evidence supporting the baseline PLS forecasts of KP in Table 1 largely disappears by simply updating the sample: The only model with positive OOS R^2 in the extended sample (which is also challenged in Section 4.1) is based on 100 BMs in annual frequency. Finally, the effect of removing the extra BM

Table 2: **Predictability of the 1-year or 1-month return on the market (RM) or market premium (MP) from June 1926 to December 2019.**

			IS_y R^2	p	OOS_y R^2	p	IS_m R^2	p	OOS_m R^2	p
RM	$bm_{\sigma,t-1}$	6	7.65	0.00	4.05	<0.01	0.40	0.06	0.34	<0.05
		25	9.59	0.00	2.05	<0.01	0.57	0.02	0.15	<0.05
		100	17.45	0.00	8.21	<0.01	2.61	0.01	0.34	<0.05
	$bm_{\sigma,t}$	6	7.17	0.01	4.11	<0.01	0.43	0.05	0.14	<0.05
		25	9.08	0.00	2.24	<0.01	1.11	0.00	-0.50	-
		100	18.39	0.00	8.66	<0.01	3.90	0.00	-0.31	-
MP	$bm_{\sigma,t-1}$	6	5.28	0.03	-2.66	-	0.28	0.14	-0.10	-
		25	7.21	0.02	-5.37	-	0.43	0.06	-0.50	-
		100	16.33	0.00	2.76	<0.05	2.53	0.03	-0.35	-
	$bm_{\sigma,t}$	6	5.02	0.03	-2.34	-	0.36	0.11	-0.42	-
		25	6.94	0.02	-5.03	-	1.39	0.00	-1.72	-
		100	16.44	0.00	3.09	<0.05	3.71	0.00	-0.92	-

Note: The table reports IS and OOS R^2 for yearly (IS_y, OOS_y) or monthly forecasts (IS_m, OOS_m); in-sample p-values of the slope coefficients of predictive regressions with Newey and West (1987) standard errors (12 lags or one lag, for annual or monthly returns, respectively); and p-values of the OOS forecast encompassing test of Clark and McCracken (2001) with Newey-West standard errors with 12 lags for annual returns (unreported, as "-", if the p-value is larger than 0.1 or if the OOS R^2 is negative). The OOS split is January 1980. The regressors are the BMs of the 6, 25, or 100 ME/BM portfolios, all in lowercase and with subscript σ because they are divided by their time-series standard deviations and in logs. The subscripts $t-1$ or t, respectively, indicate BMs with one extra lag or not.

Interpretation: PLS with the empirical choices in KP, [RM, $bm_{\sigma,t-1}$], still delivers positive performances in the full sample. But the market premium predictability essentially disappears, even with these choices. The extra BM lag improves the monthly forecasts, but slightly impairs the yearly ones.

lag in the [MP, $bm_{\sigma,t}$] models is again to obtain slightly higher OOS R^2 for the annual forecasts, but even more negative OOS R^2 for the monthly forecasts.

4.1 Sample period and OOS split choices

This section analyzes time variation in predictability, considering the baseline forecasts of KP (with standardized BMs of 100 portfolios, in logs, and with one extra lag as regressors). These are the models (in annual frequency) that appear to have positive OOS R^2 for the market premium in Table 2. However, as I will show, these forecasts have underperformed the historical mean for about two decades. The (unreported) reason is that these forecasts are frequently below the historical average since the start of this century, while realized returns are often above average: The procedure in KP correctly predicts the negative returns of the 2002 downturn and the financial crisis, but it also keeps mistakenly predicting a series of low returns that do not materialize.

The left-hand panel of Fig. 1 replicates the main graphs in Welch and Goyal (2008), starting in 1945. They display the cumulative squared prediction errors of the prevailing mean minus the cumulative squared prediction error of the standard PLS model with 100 portfolios (cumulative ΔSSE). As in Welch and Goyal (2008), the units are not intuitive (and unreported here), but the pattern is: Increases in the cumulative ΔSSE graph imply that the PLS predicts returns better in that period, otherwise, the mean is a better prediction. Hence, it is possible to adjust starting and ending dates to check the performance in different subperiods.

For example, consider the bottom graph (market premium in annual frequency). If we start at the peak, around 1975 corresponding to the Oil Shock, we would have to wait until around 2010 (the final date in KP's sample) for the forecasts to outperform the mean. And this positive performance would only hold for the years around 2010. In contrast, the cumulative ΔSSE is very low around 1980. Hence, the split date of KP provides the opposite example: Starting in 1980, almost any ending date results in positive OOS performance.

For comparison, the right-hand panel of Fig. 1 also displays the corresponding graphs reported by Kelly and Pruitt (2013), for the OOS R^2 by sample split date. In all graphs, the (dark) navy lines are for the full sample.

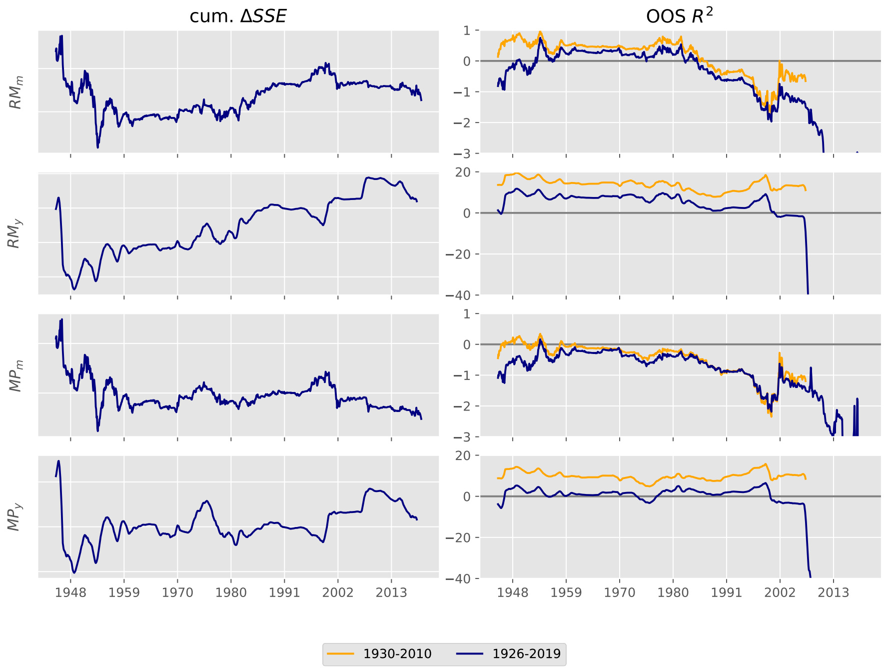

Figure 1: Out-of-sample statistics by sample split date for recursive forecasts of 1-year or 1-month return on the market or market premium based on BMs of 100 portfolios.

Description: The left-hand side graphs (cum ΔSSE) show the cumulative sum of squared forecasting errors of the historical mean minus the one from the model since 1945. An increase in the line indicates that the model outperforms the historical mean in that period. The right-hand side graphs show the OOS R^2 by sample split date. The graphs are for the market return in monthly, RM_m, or yearly frequency, RM_t, and for the market premium forecasts (MP_y and MP_m). The (dark) navy lines in each graph are for the full 1926-2019 sample. The (light) orange lines are their equivalent for the 1930-2010 subsample of KP. There are more extreme values outside the bounds in each graph (unreported for scaling reasons).

Interpretation: Both the OOS split and sample ending choices are consequential. For the 1980 OOS split in KP, the OOS performance of market premium forecasts in annual frequency is mildly positive if the sample ends in 2019 (OOS R^2 graph), and for most other ending dates (cum ΔSSE graph). But for the 1975 Oil Shock split date suggested by Welch and Goyal (2008), the annual market premium forecasts only outperform the mean for samples ending around 2010 (the ending in KP).

The (light) orange lines are for the 1930-2010 sample of KP.[10] The graphs display results for market return forecasts in yearly or monthly frequency (respectively RM_y or RM_m), and equivalently for market premium forecasts (MP_y or MP_m).

Market return:
Based on the top (RM) graphs on the right-hand panel of Fig. 1, the performance of the market return forecasts in the full sample (the navy lines) is less stable than the results based on the sample in KP (the orange lines). The monthly forecasts have negative OOS R^2 for almost all recent split dates after around 1985. In addition, the OOS R^2 of the annual forecasts plummets around 2010. However, the cumulative ΔSSE graphs still tend to have positive drifts since around 1950, especially in annual frequency. This indicates that PLS could still be useful to forecast the annual return on the market, despite its performance in the last 20 years.

Market premium:
The PLS forecasts of the market premium are substantially more problematic: The OOS R^2 on the right-hand side graphs is either negative for monthly forecasts or essentially zero for yearly forecasts based on most OOS split dates in the full sample. The increases in cumulative ΔSSE for other ending dates are often followed by decreases of similar magnitudes and the overall drifts are close to zero, especially in monthly frequency. From this perspective, the PLS procedure implemented by KP provides, at most, a marginally consistent model of market premium forecasting, only in annual frequency, and negative since the Oil Shock, since 2002, and strongly negative since the financial crisis. In summary, the PLS implementation of KP cannot generate a valid market premium model, according to the hurdle in Welch and Goyal (2008).

In fact, both cumulative ΔSSE graphs in annual frequency show sharp increases around the financial crisis, right before 2010, which explain the extraordinary results in KP, and sharp decreases until today, which explains the dismal results that I report. There are similar increases and decreases in performance in many periods and this time variation in predictability

[10]There are more extreme values outside the bounds in each graph (unreported for scaling reasons). The graphs start in 1945 to avoid very negative values for some models, so that all graphs have the same scale (per frequency).

seems to be an intrinsic feature of the estimates. This is one of the reasons why the PLS fails the stability test of Welch and Goyal (2008).

4.1.1 Why PLS forecasts market returns but not the market premium

The identity in Vuolteenaho (2002)—which KP use exactly for returns—relates BMs to **excess** returns (not to returns). According to this theory, BM ratios contain information about excess returns (and the market premium as a consequence). Hence, it seems surprising that the model fails to predict exactly the market premium (while sometimes predicting the return on the market). However, this seems to be an issue with the asset pricing theory in Vuolteenaho (2002), not with the PLS procedure. According to the theory in Berk (1995), in which BE is simply a proxy for expected cash flows, BMs are, indeed, more closely related to expected **returns** (not only excess returns, as in Vuolteenaho, 2002). This helps to explain the evidence in the present paper.

5 Summary

With respect to the main question in the present paper, we learn that the PLS method of KP does not deliver a forecasting model of the market premium that overcomes the hurdle of Welch and Goyal (2008): None of the forecasts have positive OOS R^2 since the Oil Shock, even if some models have positive performances during some periods.

We also learn that at least four empirical choices are crucial to obtain the results reported by KP: (i) Forecasting the market return instead of the market premium; (ii) the sample ending in 2010, right after the PLS strongly outperforms the historical mean (in annual frequency), but before it starts to underperform; (iii) using BMs divided by their standard deviations (instead of untreated BMs) as regressors in that particular sample period; and (iv) using BMs in logs. Finally, another crucial choice for monthly forecasts in particular is (v) to use BMs with one extra lag as regressors.

References

Berk, J. B. 1995. “A critique of size-related anomalies”. *Review of Financial Studies*. 8(2): 275–286.

Clark, T. and M. McCracken. 2001. "Tests of equal forecast accuracy and encompassing for nested models". *Journal of Econometrics*. 105(1): 85–110.

de Oliveira Souza, T. 2020. "Two out-of-sample forecasting models of the equity premium". *Unpublished working paper. University of Southern Denmark*.

Kelly, B. and S. Pruitt. 2013. "Market expectations in the cross-section of present values". *Journal of Finance*. 68(5): 1721–1756.

Kelly, B. and S. Pruitt. 2015. "The three-pass regression filter: A new approach to forecasting using many predictors". *Journal of Econometrics*. 186(2): 294–316.

Newey, W. K. and K. D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix". *Econometrica*. 55(3): 703–708. ISSN: 00129682, 14680262.

Vuolteenaho, T. 2002. "What Drives Firm-Level Stock Returns?" *Journal of Finance*. 57(1): 233–264. ISSN: 00221082, 15406261.

Welch, I. and A. Goyal. 2008. "A comprehensive look at the empirical performance of equity premium prediction". *Review of Financial Studies*. 21(4): 1455–1508.