Is the ex-ante equity risk premium always positive? Evidence from a new conditional expectations model

Hoang, Khoa; Faff, Robert

Published in:
Accounting and Finance

DOI:
10.1111/acfi.12557

IS THE EX-ANTE EQUITY RISK PREMIUM ALWAYS POSTIVE?

EVIDENCE FROM A NEW CONDITIONAL EXPECTATIONS MODEL

Khoa Hoang[1]

Robert Faff *****

ABSTRACT

We model the conditional risk premium by combining Principal Component Analysis and a statistical learning technique, known as Boosted Regression Trees. The method is validated through various out-of-sample tests. We apply the estimates to test the positivity restriction on the risk premium and find evidence that the risk premium is negative in periods of low corporate and government bond returns, high inflation, and downward sloping term structure. These periods are linked with changes in business cycles; the states when theories predict the existence of negative risk premium. Based on the evidence, we reject the Conditional Capital Asset Pricing model and raise a question over the practice of imposing the positive risk premium constraint in predictive models.

Keywords: Risk premium, principal component, regression tree, inequality restriction, conditional asset pricing model, multiple hypothesis testing.

JEL Classification: G12, G14, G17

---

[1] Both authors are from the UQ Business School, University of Queensland. Please address all correspondence to Khoa Hoang Email: t.hoang@business.uq.edu.au; Phone: +61 7 3346 0753; Fax: +61 7 3346 8166. We are grateful for valuable comments from Lorenzo Garlappi, Mamiza Haq, Allan Kleidon, Yong Li, Barbara Ostdiek, Terry Pan, Tom Smith, and Kathleen Walsh. We also benefit from suggestions by participants at 6th Financial Markets and Corporate Governance Conference, Perth Australia, April 2015; and Accounting and Finance Association of Australia and New Zealand Conference, Hobart Australia, July 2015.

1. Introduction

The risk premium serves as a central theme in asset pricing models. If marginal investors are strictly risk-averse and expected utility maximisers, they demand higher returns for investments that have higher risk. Consequently, the ex-ante market return should always exceed the risk-free rate, leading to the positivity of the ex-ante risk premium. The positivity restriction is embedded in the Conditional CAPM to ensure the mean-variance efficiency of the market portfolio. Merton (1980) argues that for equilibrium in capital markets, this condition should be explicitly incorporated as a necessary condition. However, general equilibrium asset pricing models do not impose the positivity restriction on the ex-ante risk premium since the market return is not necessary negatively correlated with the marginal rate of substitution in all states of the economy. Whitelaw (2000) shows that the market risk premium can be negative since the market can provide a desirable hedge to adverse shocks to the investment opportunity set, particularly in states associated with a high probability of regime shift.

As such, the positivity restriction on the ex-ante risk premium remains an open empirical question, of particular interest across asset pricing models. While the belief in a positive risk premium is so strong that the recent empirical literature has gone so far as to impose this constraint directly in the predictive stock returns models,[2] early empirical evidence from Boudoukh, Richardson and Smith (1993) and Ostdiek (1998) detects a negative risk premium in both US and international markets. Further, Huang, Jiang, Tu and Zhou (2015) suggest that optimally switching the positive risk premium constraints on and off, in predictive models can be a fruitful practice. In

[2] For example, see Campbell and Thompson (2008); and Pettenuzzo, Timmermann and Valkanov (2014).

the current paper, we re-investigate this research question with a new conditional expectations model. The key motivation for our study surrounds the criticisms levelled at existing proxies for the expected return employed in the empirical literature. We highlight the criticisms below.

Since conditional expectations are unobservable, testing reliable asset pricing models requires sound empirical proxies. A popular candidate is the ex-post realised return. However, information surprises in the realised return are either large by themselves or highly correlated so that the aggregate effect is large, which might have a major impact on the average realised return, even over a long period of time (Elton, 1999). Lundblad (2007) finds that the required data span could well exceed 100 years, if one wishes to uncover the positive risk-return relation, relying on the realised return. Worse still, structural breaks can undermine the use of a long period of data.

An alternative approach is to project realised returns on a small set of predetermined variables. However, since the identity of the investors' information set is unknown, omitted variable bias surrounding the linear regression is likely to yield misleading inferences (Ang and Bekaert, 2007). The inclusion of as many variables as possible to span the true information set creates another problem: the degrees of freedom quickly exhaust when the number of predictors approaches the number of observations (Ludvigson and Ng, 2007; and Kelly and Pruit, 2013). Adding to this complication, data snooping bias is prevalent and the statistical results are sensitive to the choice of conditioning variables (Foster, Smith and Whaley, 1997; and Harvey, Liu and Zhu, 2016). After a careful assessment of the literature, Goyal and Welch (2008) reach a startling conclusion that most existing economic variables do not outperform a naïve historical mean in

predicting future stock returns. If returns are not predictable, testing conditional asset pricing models becomes an elusive task (Chen and Zhao, 2009).[3]

To address the above critiques, we propose a two-stage procedure in modelling the conditional risk premium, including Principal Component Analysis and a state-of-the-art supervised learning technique, known as Boosted Regression Trees (BRT). Accordingly, in essence, the core novelty of our study is in how we combine two powerful techniques to provide an eminently implementable and meaningful solution to a devilishly complex problem, plagued by high dimensionality and extreme non-linearity.

Specifically, in the first stage, the Principal Component Analysis (PCA) is performed to capture the common information underlying 156 financial variables from Ludvigson and Ng (2007) and Goyal and Welch (2008). PCA is a popular method in summarising information in a large set of variables (Stock and Watson, 2002). Furthermore, Kozak, Nagel, and Santosh (2018) utilise strong commonality in asset returns and use the PCA to model the stochastic discount factor. They show that strong cross-sectional predictive power of the principle component factors. With this step, we aim to span the identity of the investors' information set, without relying on a small

[3] Another strand of literature relies on forward-looking information to infer the first moment of returns. Brav, Lehavy, and Michaely (2005) and Bali, Hu, and Murray (2019), among others, use analysts price targets, while Pastor, Sinha, and Swaminathan (2008) back out the implied cost of capital from price and analysts earnings forecasts, to proxy for expected returns. Recently, Martin (2017), using information from the options market, links volatility index data to expected return and derives a lower bound on the market risk premium. Martin and Wagner (2019) leverage the idea from Martin (2017) and estimate the expected return at the individual stock level.

subset of arbitrary conditioning variables, by compressing a much richer source of information into a small set of factors.

Our new set of predictors, including a small number of principal components and the well-known predicting variables in Goyal and Welch (2008), enters the second stage. In this stage, to address the inherent non-linear relations between the expected excess return and the information set, we use the BRT technique developed in the statistical learning literature. Instead of assuming a linear functional form or impose strong modelling assumptions, the regression trees is a non-parametric method that approximates the unknown function by recursively partitioning the predictor space X into disjoint sub-regions; simple constant models are fitted into these regions, ideally, until the information is "tamed".

Figure 1 provides a visual explanation of how the regression trees method works. The variables and their associated values are the splitting parameters. Numbers within circles are constant values within each partitioned region, which are used to model the expected risk premium. For example, corporate bond return is first chosen to split the sample at -2.5%. In the region where corporate bond return ($corpr$) is less than -2.5%, the 8^{th} principal component ($PC8$) is a subsequent split at the value of 0.45. If the $PC8 \geq 0.45$, the constant -7.1% is a fitted value of the equity premium, whereas a further split is needed for the region in which $PC8 < 0.45$. The procedure continues until some stopping criteria are reached. The model can then form the prediction by summing up constant values in corresponding bins.

We employ an ensemble method, known as *boosting*, to improve the model's fit. The idea of this scheme, similar to those discussed in the forecast combination literature (Rapach, Strauss and Zhou, 2010) is to aggregate the predictions from models which do not perform well individually into one with considerably improved properties. In this context, the simple individual

models are the trees. The algorithm builds a sequence of small trees (typically after 1 split), in which subsequent trees seek to minimise the residuals, weighted by previous trees' errors. As a result, the final model is the sum of the forecasts from all of the small trees. The BRT methodology, perhaps one of the most powerful ideas recently developed in the statistical learning literature, not only helps extract unknown functional forms, but also has the ability to handle high dimensional data (Hastie, Tibshirani and Friedman, 2009).

Recently, the BRT method has found interesting applications in the economics and finance literature. Ng (2014) employs the boosting tree method in a classification setting to predict recessions. Kleinberg, Lakkaraju, Leskovec, Ludwig and Mullainathan (2018) investigate whether the gradient boosted decision trees can help judges make better bail decisions. The closest paper to ours is Rossi and Timmermann (2015) who use the BRT to model conditional covariance between stock market returns and the economic activity index. They then link the market expected return with the estimated covariance to test the linear risk-return restriction from the Merton's Intertemporal Capital Asset Pricing Model (ICAPM). Different from their approach using the BRT to fit second moment of returns, we use the BRT to model the market expected return directly from the large set of information variables without imposing the linear functional form. Also, while they are interested in investigating the risk-return trade-off implied by the ICAPM, we assess the notion of non-negativity restriction imposed by the CAPM.

[Insert Figure 1 here]

Our findings can be summarised as follows. First, we employ the out-of-sample R^2 to verify the validity of our two-stage method in forming the conditional risk premium. The out-of-sample R^2 is perhaps the strictest and most common test in the return predictability literature (Goyal and Welch, 2008). The PCA+BRT consistently delivers positive out-of-sample R^2 across

model specifications. The predictive power is economically significant assessed via the hypothetical utility gain experienced by a mean-variance investor. To illustrate, an hypothetical investor with a relative risk aversion coefficient of 3 is willing to pay 1.21% p.a. management fee to access the predictive signal from the PCA+BRT model.

Second, the encompassing test Harvey, Leybourne, and Newbold (1998) suggests that the predictive power of future returns consistently beats those of their linear methods (kitchen-sink OLS, mean combination, and Least Angle Regression of Efron, Hastie, Johnstone and Tibshirani, 2004). We attribute this superior performance to the ability of the technique capturing the non-linear nature of the information set and expectations.

Third, we find evidence to support the existence of negative risk premium in some states of the world, which are related to periods in which corporate bond returns and long-term government bond rates of return are low, along with the preceding period experiencing a negative risk premium, and a downward-sloping term structure of interest rates. We, therefore, are able to reject the Conditional CAPM as the model's necessary condition (i.e. the positivity of the risk premium) is violated. In contrast, we do not detect a negative risk premium when realised return is used as a proxy for expected returns. The simulation result suggests that the lack of power is due to a well-known disadvantage of realised returns i.e. noise (Merton, 1980; and Elton 1999).

Fourth, our simulation rules out the spurious significance of empirical results driven by statistical biases. Collectively, our findings further raise questions over the recent practice of imposing the positive risk premium constraint in predictive models (see, e.g., Pettenuzzo, Timmermann and Valkanov, 2014).

We engage with and contribute to several strands of literature. Empirically, the positive risk premium restriction remains an under-researched question despite its importance in theoretical

modelling. We are among a few who have conducted a direct empirical test of this restriction (e.g., Boudoukh, Richardson and Smith, 1993; Ostdiek, 1998; and Walsh, 2015). Unlike previous studies using the realised return as a proxy for the expected return, we explicitly model conditional expected return by adopting the superior conditional expectations model. We join an emerging band of researchers to promote the statistical or machine learning approach in finance applications (for example, see Chinco, Clark-Joseph and Ye (2018) on LASSO regression; Rossi and Timmermann (2015) on the BRT; Kozak, Nagel and Santosh (2017) on LAR).[4] Finally, our PCA is in a similar spirit to that of Ludvigson and Ng (2007), and Jurado, Ludvigson and Ng (2015) in addressing the dimensionality problem in large datasets.

The remainder of our paper is structured as follows. Section 2 presents the main two-stage method, followed by a brief data description in Section 3. Empirical results are presented and discussed in Section 4. A conclusion is offered in Section 5.

2. Research Method

2.1 Regression Trees

We employ a two-stage empirical strategy. The first stage involves Principal Component Analysis (PCA). PCA circumvents the issue of having a "degrees of freedom" problem (having such a vast number of potential predictors relative to the time series sample size) by linearly combining the available predictors into several orthogonalised common factors that collectively

[4] LASSO stands for "least absolute shrinkage and selection operator", while LAR stands for "Least Angle Regression". They belong to the shrinkage linear regression class which impose statistical constraints on estimating parameters.

help to explain most variation among the original set.[5] The second stage involves regression trees. We provide the intuition how tree-based regression works and relegate some more technical details to the Online Appendix A.1.

The regression trees method seeks to estimate the unknown functional form by dividing the predictor space into non-overlapping regions and simply models the response variable as a constant within each region. In the first step, the tree algorithm searches through the entire predictor space, and chooses an independent variable ′j′ and a splitting point ′s′. The subsequent optimal split (a splitting point and a predictor) is decided by minimising the sum of squared residuals in the resulting regions from previous splits, instead of searching on the entire sample space. The process continues until some stopping criteria are satisfied.

2.2 Boosting

Instead of fitting the data into a single large tree with a number of splits, boosting builds a number of simple trees sequentially and seeks an efficient process to combine them. This is a type of ensemble method. It exploits the idea that a series of individual models that are poor in isolation, can have considerably improved properties by being treated as a "team" - collectively in a type of "portfolio".

[5] The first component is the linear combination of predictors that capture the largest variation in in the predictor space. The second component repeats the maximisation with one additional constraint, namely, that it is also orthogonal to the first component. The N principal components can be extracted by calculating the N largest eigenvalues and eigenvectors from the covariance matrix of the original predictor space.

The aim is to build, say B simple trees, in B iterations. There are three steps in each iteration. First, a small tree is built, typically after a small number of splits by fitting the residuals rather than the original response variable in each region.[6] The reason behind fitting residuals is that it aims to grow a new tree in the region where the current fitted model does not perform well, which results in a high sum of squared residuals. That is, the essential idea of a supervised learning technique, is that it can iteratively "learn" the true functional form in response to differences between the original and generated output. Second, the additional information of the new tree is "slowly" updated to the current predicting model. Third, a new set of residuals is calculated with the new model to prepare for the next iteration. The process continues after B iterations and the final model is just the sum of B fitted trees.

The superior forecasting performance of boosting can be attributed to three factors. First, building small trees can reduce the risk of overfitting in a large regression tree with fewer observations in each region. Second, and more critical to the boosting algorithm, is the use of a small λ shrinkage parameter that controls for the learning rate of new information in the current model. Empirically, a small λ, i.e. a slow updating rate, tends to produce better out-of-sample forecast accuracy (Friedman, 2002). Finally, boosting provides a type of model averaging by summing up all of the small trees, which enhances the stability of the forecasts (Rapach, Strauss and Zhou, 2010).

Three main intertwined parameters are required for the empirical implementation: 1) the number of splits to use in building a tree in each iteration, 2) the number of trees to be built, i.e.

[6]We use one split, which effectively focuses on the main effect of individual trees, on approximation of the functional form.

the number of boosting iterations, 3) the value of the learning rate λ. We follow common practice in the statistical learning literature and adopt: 1) one split, 2) 1,000 boosting iterations, and 3) $\lambda = 0.001$ (see, for example, Hastie, Tibshirani and Friedman, 2009).[7]

To assess the model validity, we rely on metrics used in prior literature. First, we employ the relative influence measure by Breiman, Friedman, Stone and Olshen (1984) to evaluate the contribution of the predictors in a single regression tree. Second, partial dependence plots inform on the marginal effect of individual predictors on the conditional expected return. Finally, we employ out-of-sample R^2, its economic significance, the Mincer and Zarnowitz (1969) unbiased test, and a forecast encompassing test to assess the predictive power of the PCA-BRT model, following Campbell and Thompson (2008) and Harvey, Leybourne and Newbold (1998). The details of these statistics are discussed in the Online Appendix A.2. In particular, the Online Appendix A.2.1 discusses relative influence measures and partial dependence plots, A.2.2 presents the construction details of the R^2_{OOS}, A.2.3 illustrates Harvey, Laybourne, and Newbold (1998) forecast encompassing test.

2.3 Inequality Constraint Method

The restriction that the conditional expected return exceeds the conditional risk free rate can be represented as:

$$\mathrm{E_t}\left(\mathrm{R_{m,t+1}}\right) - \mathrm{R_{ft}} \geq 0 \qquad (1)$$

where $\mathrm{R_{m,t+1}}$ is the market expected return from time t to $\mathrm{t}+1$ and $R_{f,t}$ is the risk free rate from t to $\mathrm{t}+1$.

[7]We use R package gbm to perform gradient boosting regression trees.

We employ the multiple inequality constraints testing framework from Boudoukh, Richardson and Smith (1993) to test restriction (1). Define μ_t as the ex-ante risk premium:

$$E_t(R_{mt+1} - R_{ft}) = \mu_t \geq 0 \tag{2}$$

As μ_t is unobservable, we replace μ_t in the inequality **Error! Reference source not found.**) with its proxy, $\hat{f}(x_t)$, estimated by the two-stage procedure. Suppose $z_{i,t}^+$ is the positive information set i.e. greater than or equal 0, that is available to the econometrician, applying the law of iterated expectation, the above inequality (2)**Error! Reference source not found.** implies multiple restrictions

$$E\left(\hat{f}(x_t) \otimes z_t^+ - \theta_{\mu z^+}\right) = 0 \tag{3}$$

where $\theta_{\mu z^+} = E(\mu_t \otimes z_t) \geq 0$.

First, the sample means of the product of the observable variables are estimated. In particular,

$$\hat{\theta}_{\mu z_i}^+ = \frac{1}{T}\sum_{t=1}^{T}\left[\hat{f}(x_t) z_{i,t}^+\right] \ \forall i = 1 \dots N \tag{4}$$

$\hat{\theta}_{\mu z_i}^+$ are referred to as the unconstrained estimates because there is no sign restriction imposed on the parameters. Next, the sample means are calculated under the inequality restriction in the null $\hat{\theta}_{\mu Z_i}^R$ by minimising deviations from the unrestricted model under the quadratic form:

$$\min\left(\hat{\theta}_{\mu z^+}^+ - \theta_{\mu z^+}\right)' \hat{\Omega}^{-1}\left(\hat{\theta}_{\mu z^+}^+ - \theta_{\mu z^+}\right) \tag{5}$$

subject to $\theta_{\mu z^+} \geq 0$.[8]

[8]We use R package nloptr with the algorithm NLOPT_LN_BOBYQA which performs derivative-free optimisation using quadratic approximation for the objective function.

where $\widehat{\Omega}$ is the consistent variance-covariance matrix of the moments. To estimate $\widehat{\Omega}$, we use the Quadratic Spectral kernel with bandwidth parameter estimated by fitting a univariate AR(1) model to each element of $\theta_{\mu z^+}$ (Andrews, 1991).

The test statistic is:

$$W \equiv T\left(\hat{\theta}_{\mu z}^{R} - \hat{\theta}_{\mu z}^{+}\right)' \widehat{\Omega}^{-1} \left(\hat{\theta}_{\mu z}^{R} - \hat{\theta}_{\mu z}^{+}\right) \quad (6)$$

The idea of the test statistic *W* is to measure how close the parameters of the restricted model $\hat{\theta}_{\mu z}^{R}$ are to those of the unrestricted model $\hat{\theta}_{\mu z}^{+}$. Under the null, the difference should be small. Wolak (1989) shows that W is distributed as a weighted sum of χ^2 with different degrees of freedom $\sum_{k=0}^{N} \Pr[\chi_k^2 \geq c] w(N, N-k, \frac{\widehat{\Omega}}{T})$ where c is the critical value for a given size test, and the weighting function $w(N, N-k, \frac{\widehat{\Omega}}{T})$ has exactly $N-K$ positive elements. To calculate the weights, we use a Monte Carlo simulation based on the multivariate normal distribution with zero mean and covariance matrix $\frac{\widehat{\Omega}}{T}$. We generate 10,000 simulations. For each simulation, we perform the optimisation (6) and obtain $\hat{\theta}_{\mu z}^{R}$. The weights are estimated as the fraction of replications in which $\hat{\theta}_{\mu z}^{R}$ has exactly N-K elements are positive.

Although out-of-sample statistics are less prone to the well-known small-sample bias (Huang et al., 2015; and Goyal and Welch, 2008), it is unclear to what extent that our approach suffers from this bias. Further, the distribution of the asymptotic W statistic depends on the values of the estimated parameters $\theta_{\mu z}$ (Wolak, 1991). It is, therefore, problematic to derive a globally exact sized test.[9] We design a simulation experiment to assess these issues. In a nutshell, we

[9] This problem is highlighted in footnotes (6) and (9) in Boudoukh et al. (1993). If the existence of dependence is in fact true and if it results in the vector of zeros not being the least favourable

simulate our samples under the null that returns are always positive and unpredictable i.e. constant positive risk premium. We subsequently run the empirical procedure (PCA + BRT + inequality testing) to obtain our simulated distributions for R^2_{OOS} and W statistics. We then compare our empirical statistics with the simulated distributions to assess the statistical significance. Details of the simulation procedure detail are described in Online Appendix A.2.4.

3. Data Description

Employing an extensive sample of US data, we consider 156 financial variables that have been used in the return predictability literature. In particular, the information set includes 143 financial variables in Ludvigson and Ng (2007). The other 13 predictors are from Goyal and Welch (2008).[10] The stock market excess return (rp) is calculated as monthly Center for Research in

under the null, it is then difficult to derive global test statistics that are valid for all values of the parameters. In other words, the test can only be thought as a localized inequality test around the point $\theta_{\mu z} = 0$.

[10] These predictors are: Log dividend price ratio (*dp*); Earning price ratio (*ep*); Default yield spread calculated as the difference between BAA and AAA corporate bond yields (*def*); Term structure (*tms*); Long term government bond yield (*lty*); Long term bond return (*ltr*); Stock variance measured as squared daily returns (*svar*); Three-month T-bill rate (*tbl*); Inflation rate (*infl*); Lag excess returns (*lret*); Net equity expansion is measured as the sum of 12 months net issue for NYSE stocks divided by the market capitalisation of the stocks (*ntis*); Book to market ratio (*bm*), and Corporate bond return (*corpr*).

Security Prices (CRSP) market portfolio return including dividend minus the 1-month T-bill rate (Rfree) (continuously compounded).[11] The sample period spans from 1960 to 2016.

It is critical to build and test the model in two different datasets because the method tends to understate the error rate in-sample.[12] We adopt a recursive window, with the first training period being 120 months, to build the model; that is, we construct the components and initial regression trees using data up to 1969:12 and form the prediction in 1970:01. The second estimate in 1970:02 is calculated from the PCA and the regression trees fitted during the training period from 1960:01

[11] Later we see that the risk premium is negative in certain phases of the business cycle. An anonymous referee argues that this specific finding could be attributed to the selection of proxies. For example, during the post GFC period when the US government credit rating was downgraded, the 1-month T-bill rate may not be a good choice for the risk-free rate proxy. Our response to this critique has three elements argumentation. First, we need to find a set of powerful to detect negative risk premium. The instrumental variables are selected because theories suggest that they link to states where negative risk premium is likely to occur. Second, the T-bill is commonly used as the risk-free rate. It is arguably the best proxy for risk-free rate that we can find. Third, we also investigate our hypothesis in the pre-GFC period and our documented results remain consistent.

[12] Ideally, a sample is divided into 3 parts: namely, a training period, a validation period, and an evaluation period, to estimate and assess a model. The training period is used to fit the model; the validation period is used to select the best set of model parameters by minimising the prediction error criteria; once the final model with the best set of parameters is built, the test period is used to evaluate the true test errors of the resulting model. We omit the validation step because the data are insufficient (Hastie et al., page 241).

to 1970:01. As a result, the time series risk premium calculated out-of-sample includes 564 observations. These observations are fed into the multiple inequality constraints framework to test the null hypothesis of the positive risk premium.

4. Empirical Results

4.1 Summary Statistics

Panel A of Table 1 reports the summary statistics of the predictors from Goyal and Welch (2008) and the conditional risk premium generated by the two-stage method. The statistics for independent variables are similar to those that have been reported elsewhere (see, e.g., Pettenuzzo, Timmermann and Valkanov (2014)). The monthly mean risk premium $\hat{f}(x)$ is 0.33% with the associated standard deviation of 0.42%. There are negative observations in the risk premium prediction, suggesting that it is meaningful to detect whether the positivity of the equity risk premium states is violated, or whether the negative fitted values are just an artefact of sampling error.

For brevity, Panel B of Table 1 presents the statistics of the first ten principal components estimated in the first stage of the procedure. The first component ($PC1$) explains the largest proportion of variation (R^2), 63%, in the original set of 156 financial variables. Orthogonalised to the first component, the second component ($PC2$) explains 5% of the variation in the predictor space. The contribution explained by subsequent components is steadily declining, with the 10^{th}

component ($PC10$)[13] capturing only 1% of the variation in the information set. We show later in that the predictive performance of the model improves significantly once the first component is introduced, and remains stable as the number of components increases.

[Insert Table 1 here]

4.2 Relative Influence Measure

Panel C of Table 1 presents the relative influence statistics which measure the contribution of individual predictors in estimating the equity premium. We detail the statistic construction in the Online Appendix A.2.1. The 5^{th} principal component appears to be the most important variable which obtains the relative influence weighting of 26.19%. The next five most important predictors of the future risk premium come from the Goyal and Welch's variable set with the relative influence scores ranging from 25.6% (corporate bond returns) to 3.5% (dividend price ratio).[14] The top five predictors dominate the sample and aggregate to a combined weighting of close to three-quarters. This evidence suggests the importance of some hidden factors that help to span investors' information set in forming expectations.

[13] The tenth component, PC10, is the linear combination of predictors that capture the largest variation in the predictor space, after imposing orthogonality conditions with respect to PC1, through to PC9.

[14]Note that the relative influence measure provides the ranking of 23 variables in the predictor space (13 variables from Goyal and Welch (2008) and 10 components from step 1).

4.3 Partial Dependence Plots

Turning to Figure 2 which presents the partial dependence plots of 5 predictors, including corporate bond return, 5^{th} principal component (PC5),[15] stock variance, term structure, and dividend price ratio in modelling the equity risk premium, we gain some insight into the existence of non-linearity in the relationship between the equity premium and information variables.[16]

With respect to the corporate bond return, the relation is highly non-linear. In particular, when returns fall in the range of -10% to -5%, the risk premium does not change and remains around -1%. In contrast, when the bond returns increase from -4% to -2%, a sharp increase in the risk premium is observed. The relation again becomes flat when the predictor passes into its positive domain. Turning to the 5^{th} principal component (*PC5*) plot, there is a flat structure in the negative and extreme positive levels. Yet, in the intermediate range, a higher value of the component is associated with a lower risk premium reaching its minimum at -0.6%. The non-linear pattern persists in the term structure of the interest rate. Although there is a general positive relation between the term structure and the equity risk premium, the slope is particularly steep across the mid-range observations. Finally, the dividend price ratio shows the strongest positive relation across the high value range, whereas it remains constant mostly across lower values. Clearly, the

[15] The fifth component, PC5, is the linear combination of predictors that capture the largest variation in the predictor space, after imposing orthogonality conditions with respect to PC1, PC2, PC3 and PC4.

[16]For brevity, we present the plots for 5 predictors. Other predictors also exhibit non-linear relations with the risk premium. The discussion of the plot construction is presented in the Online Appendix A.2.1.

partial dependence plots raise a concern over the assumption of using a linear functional form in modelling the conditional expected return in the prior literature. As the non-linearities are evident across all predictors, accounting for these effects might help capture the conditional risk premium. We turn next to the out-of-sample evaluation to provide formal evidence in relation to this claim.

[Insert Figure 2 here]

4.4 Model Validity

4.4.1 Out-of-sample Tests

We start the out-of-sample comparison by comparing the out-of-sample R^2_{OOS} and the associated p-value of our BRT method, as opposed to those of the KS, Mean, and LAR. We also report the economic significance of these R^2_{OOS} through the gain in utility for mean-variance investors, with three different degrees of relative risk aversion γ ($\gamma = 1, 3, \text{and } 5$).[17]

As can be seen from Table 2 Panel A, the boosted trees estimation provides superior predictive performance. The R^2_{OOS}, 1.00%, with the associated p-value of 3%, is consistently larger than its counterparts. In particular, the KS performs the worst in predicting future stock excess returns, with R^2_{OOS} being -9.71%. This is high in absolute value, so the historical average outperforms predictions generated by the OLS model. The low R^2_{OOS} of the OLS might be due to the linearity assumption and degrees-of-freedom issues, which likely cause the model to be highly unstable. This result is not surprising and is documented in the previous literature. The mean

[17] The out-of-sample R^2_{OOS} and its economic significance i.e. utility gain construction is discussed in section A.2.2.

combination and LAR show a large improvement compared to OLS, with the associated R^2_{OOS} equal to 0.52%, and 0.7% and statistically significant at the 10% level.

With respect to economic significance, the utility gains based on the superior BRT forecasts are consistently positive and economically sizable, ranging from around 0.57% ($\gamma = 5$) to nearly 3.63% ($\gamma = 1$). Again, these utility gains are larger than the other methods, across most degrees of risk aversion. For example, although the LAR's R^2_{OOS} is positive, investors with the degree of risk aversion above 3 experience utility losses if they allocate their capital between stocks and bonds based on the LAR forecasts. Similarly, the KS forecast yields utility loss at risk aversion coefficient of 5. Among the alternatives, the mean combination is the only one that consistently delivers utility gains across various degrees of risk aversion.

Panel B of Table 2 further reports the Mincer-Zarnowitz test. The *p-value* near 0% indicates that the traditional OLS is highly misspecified, whereas the BRT model, mean combination, and LAR show no evidence of misspecification (the associated *p-values* of 81%, 68% and 65%, respectively). The evidence suggests that a linear model with unrestricted coefficients as in the KS method is unlikely to capture the complexity of the data generating process. The Mean and LAR methods provide different mechanisms to impose restrictions on the coefficients in the linear system, while the BRT deviates from the non-linearity.

We present the pair-wise correlation between the forecasts produced by different methods in Panel C of Table 2 The BRT forecasts have high correlation with the mean combination and LAR forecasts. The correlations are 0.71 and 0.59, respectively. On the other hand, the historical average forecasts exhibit lowest correlation with other methods. These results suggest that forecasts produced by different methods might contain different information. We now turn to the

formal encompassing test evidence to investigate whether forecasts produced by a given model provide additional information relative to those of the other models.

Next, we employ forecast encompassing test statistic by Harvey, Leybourne and Newbold (1998). The test statistic assess the null hypothesis that model i encompasses model j against the one-sided alternative hypothesis that model i does not encompass model j. We provide the statistic construction in the Online Appendix A.2.3. Panel D of Table 2 presents the *p-value* of the test statistics for each pair. Each entry corresponds to the null hypothesis that the forecasts given in the row heading encompasses the forecasts presented in the column heading. For example, the *p-values* in the Mean row indicate that the forecasts produced by the mean combination method encompass those by the historical average (*p-value*=0.87) and the LAR (*p-value*=0.33). In contrast, at 10% level of significance we can reject the null hypothesis that mean forecasts encompass the forecasts of the BRT (*p-value*=0.09) and KS (*p-value*=0.08). None of the other existing methods can encompass the BRT forecasts. Across historical, KS, and LAR rows the p-values in the BRT column are all less than 10%. On the other hand, in the row BRT, the *p-values* across columns are consistently greater than 0.2, indicating that the null that the BRT forecasts encompass those of other methods cannot be rejected. The evidence, collectively, suggests that the forecasts in the existing models do not contain useful information beyond that already incorporated in the BRT.

[Insert Table 2 here]

4.4.2 Further Diagnostic Specifications

Table 3 provides a more detailed picture of how the BRT method fares in different specifications. We first show the importance of identifying the true investors' information set by presenting the model's predictive performance associated with varying numbers of principal components entering stage two. The label "0 component" represents the case when we use the raw

input (156 variables) directly in step two, without performing step one (PCA). We then increase the number of components entering the second step to 1, 2, 5, 10, 50, 100, and 120.

Panel A of Table 3 shows that without step 1, PCA (0 components), the BRT performs quite poorly in predicting future excess returns. The R^2_{OOS} stays at 0.31% and is not statistically significant (*p-value*=0.11). However, once the first principal component is placed in the second stage with the 13 predictors of Goyal and Welch (2008), the predictive performance improves considerably to 0.95%, significant at the 5% level. It also generates a sizable average utility gain of 1.01%. As the number of components increases, BRT continues to beat the historical average, and attains its best performance at 10 components. It is consistent with the evidence in Panel C of Table 1 in which *PC5* and *PC8* are among the top three most important predictors in forming the expected risk premium. The combining evidence highlights the importance of better identifying investors' true information set, which is essentially what the first stage tries to address. Of particular interest, the BRT method with the PCA can be thought as a conjunction of forecast combination and information combination suggested by Rapach, Strauss and Zhou (2010).

[Insert Table 3 here]

Turning attention to Panel B of Table 3, we illustrate the model performance with varying numbers of iterations to 2,000, 3,000, 4,000, and 5,000 boosting trees. There is a potential threat of overfitting once the number of trees increases to 4000 as the R^2_{OOS} declines to -0.6%. The overall advice is that care should be taken as one increases the number of boosting iterations. Since we do not use a validation period to find the best set of parameters, we adopt the common practice to start our specifications i.e. the learning rate, the boosting iterations and the tree depth. Additionally, it is not obvious whether the overfitting would bias against or for our main non-negativity

hypothesis. The simulation reported later in the Section 4.**Error! Reference source not found.**, suggests that our results remain statistically significance after accounting for generated parameters.

Next, we examine the robustness of the predictive performance as we vary the lengths of estimation and evaluation periods. Specifically, Panel C of Table 3 presents the out-of-sample performance when the training periods increases from 10 years to 20 years, and 30 years, and thus the evaluation periods are effectively 1970:1-2016:12, 1980:1-2016:12, and 1990:1-2016:12. Additionally, following Rapach, Strauss and Zhou (2010) other evaluation samples include Post Oil Shock (1976:1-2016:12), Technology Bubble (1:2000-2016:12), and the recent Great Recession (1970:1-2008:1). Overall, across different scenarios, we consistently find the out-of-sample R^2_{OOS} positive, ranging from 1.29% to 1.58%. Most cases are statistically significant at least at the 10% level of significance, with the exception of the period between 2000:1 and 2016:12. This seems to be the strictest period, because within the course of only 17 years there are two major crises. On the other hand, the utility gains are economically large, peaking at 1.21% in the full sample period with the initial 10-year training period.

Figure 3 plots the R^2_{OOS}, with varying sample splits, for the actual sample and the 90%, 95%, and 99% confidence intervals obtained from simulated samples in the Online Appendix A.2.4. Two observations are worth noting. First, the simulated R^2_{OOS} are largely negative, demonstrated by a large red region below 0. It means that under the null of no predictability, our two-stage method does not mechanically generate positive out-of-sample R^2_{OOS}. Second, as the test sample becomes smaller, the upper critical values start to increase, whereas the actual R^2_{OOS} slightly decreases, yet remains positive and consistently above the 5% simulated cut-off R^2_{OOS}. This is consistent with the statistically significant R^2_{OOS} at 5% level reported in Table 2 and Table 3.

Therefore, we remain confident that our significant R^2_{OOS} is not a statistical artefact resulting from finite samples and data-mined sample splits.[18]

[Insert Figure 3 here]

Our final validity test pertains to long-horizon predictability. If state variables tracking risk premium are persistent, the economic significance of forecastability is more visible for longer horizons (Cochcrane, 2011). We investigate the forecasting power of our method over horizons spanning from 1 month to 36 months. Panel D of Table 3 reports R^2_{OOS} with Clark-West *p-values*, corrected for overlapping observations using the Newey and West (1987) method.[19] The R^2_{OOS} increases monotonically with horizons, attaining 19% and 31% for 12-month and 24-month horizons, respectively.

4.5 Inequality Test

The evidence above supports the validity of the two-stage method. We now apply the forecasts generated by the model to the multiple inequalities framework and test the main hypothesis: the positivity of the ex-ante risk premium. We use forecasts from 2,000 boosting iterations since this specification delivers the highest economic significance. In addition to testing the full sample (1970:1-2016:12), we consider a sub-period 1970:1-2007:12 which excludes the

[18] Rossi and Inoue (2012) demonstrate that sample splits can be data-mined.

[19] Long-run horizon forecasts are notorious for statistical biases (Boudoukh, Richardson and Whitelaw (2008)). Although we report the p-value, we do not wish to establish statistical significance since the long-horizon R^2_{OOS} statistical properties are relatively unknown and beyond the scope of our paper.

post Great Recession period based on previous evidence that the predictive performance deteriorates after the 2008 recession (see Table 3 Panel C). For comparative purposes, we also conduct the same analysis yet using the realised risk premium as a proxy for the expected risk premium.

4.5.1 Instrumental Variables

Although the principal components have strong statistical power in predicting risk premium, we refrain from choosing them because they carry little economic interpretation to why they are related risk premium. One way to "assign" an economic identity to the PCs is to regress the PCs on some economic factors e.g. GDP growth, stock market volatility and so on. We do not pursue this exercise for several reasons. First, our premise is that the true information set is unknown, thus correlations with some existing variables do not necessarily connect a principal component to an economic measure. Second, we do not have theories to guide us. Accordingly, our goal is less ambitious. As the information set is large, we simply hope to avoid omitted variable bias. At the same time, we want to utilise the commonality among the predictors which seem to be particularly useful in the return predictability context (Kozak, Nagel, and Santosh 2018).

On the other hand, to provide a test against the null that the positivity risk premium holds, our set of instrumental variables should be powerful and economically motivated to support the alternative models i.e. the negative risk premium states exists. To this end, we carefully consider the existing evidence based in a substantial literature on time-varying risk premium. Our final set of seven variables all belong to the list of predictors in Goyal and Welch (2008). In particular, we include corporate bond return ($corpr$), long term government bond returns (ltr), inflation rate ($infl$), T-bill ($Rfree$), the term structure of interest rates (tms), volatilities ($svar$) and lagged risk premium (nrp). For example, the downward sloping term structure, high T-bill, and high

inflation rate contains useful information about regime shifting states in business cycles (Boudoukh, Richardson, Whitelaw, 1997). These states are likely to be associated with the negative risk premium periods (Whitelaw, 2000). Volatility arises as a natural candidate and a large literature focuses on testing the inter-temporal relation between risk and expected return. We choose the corporate bond return because the variable seems to be the strong predictor in the predicting model (see relative influence measure in Panel C of Table 1). The choice of long-term bond return is the government counterpart to corporate bond return. Finally, we choose the lagged risk premium following Ostdiek (1998).

The variables need to be constructed to be non-negative, so that the inequality restrictions in the testing framework are preserved. In this regard, we employ two transformation methods that correspond, a priori, to periods of low implied premium (i.e. low corporate and government bond returns, lagged negative risk premium, and downward-sloping term structure, low volatilities, high risk free rate, high inflation rate). We employ dummy variable and magnitude methods to transform the instrument variables following Boudoukh, Richardson and Smith (1993). The transformation methods are discussed in Online Appendix B.

4.5.1.1 Dummy Transformation

Table 4 reports several summary statistics of the instrumental variables. Notably, most of the states occur quite frequently. For example, the high T-bill rate and low corporate bond return periods occur 57% and 49% of the time, respectively. In contrast, the downward sloping term structure infrequently occurs in our sample period, at around 9% of the time. However, as can be seen below, the term structure is one of the most successful instruments in capturing the negative risk premium states.

All of the instruments exhibit some order of autocorrelation. In particular, the T-bill rate presents the highest level of autocorrelation with the coefficient equal to 0.94. The autocorrelation coefficients of the inflation rate, the term structure, and the lagged negative risk premium are 0.23, 0.27, and 0.12, respectively. The two bond returns instruments exhibit the lowest autocorrelation at around -0.02. Additionally, looking at cross-correlation, the conditioning variables are strongly correlated. For example, corporate bond return and government return exhibit the highest correlation at 0.84, followed by the correlation between T-bill and term structure (0.61). Therefore, low bond return, high T-bill rate, or downward term structure are not independent events. As a result, the conditional mean estimators are likely to be correlated, and it is necessary to estimate a consistent variance-covariance matrix. Our multiple inequality testing incorporates the matrix information into the calculation of test statistic distribution to assess the significance of the departure from the null.

[Insert Table 4 here]

In Table 5, we present the main empirical results of our hypothesis for the full period (1970:1-2016:12) and for the pre-recession period (1970:1-2007:12). The test results with respect to the transformation 1 (dummy) are reported in Panel A. The downward sloping term structure is the best instrument, followed by the lagged negative risk premium in capturing negative risk premium periods; the annual mean risk premiums are –3.52% (-3.57%) and -1.30% (-1.38%), respectively for the full (pre-recession) period. Conditioning on the periods of low corporate bond return, the risk premium is 0.69%, whereas the conditional mean risk premiums in the pre-recession period are -0.11% and -0.12%.

Some caution should be applied – we need to avoid hastily interpreting these univariate results as evidence for a departure from the null. The high cross-correlations indicate that the

instrumental variables might pick up the same events or the same sampling errors. Therefore, it is necessary that our formal multiple inequality test statistic accounts for these autocorrelations and cross-correlations. The joint *W* statistic is 3.30 (2.23) with an asymptotic *p-value* of 13% (20%) in the full (sub) sample period. Overall, the asymptotic joint statistics suggest that we cannot reject the positivity restriction of the ex-ante risk premium at 10% level of significance.

Notably, the empirical *p-values* under the simulation arrive at the same conclusion. The *p-values* are 9% and 10% for the full and sub samples, respectively. Recall that we create the distribution of 1,000 *W* statistics under the null that the risk premium is unconditionally positive and unpredictable. Therefore, we cannot reject the positive risk premium restriction at 5% level. Given the similar conclusions reached from asymptotic and empirical *p-values*, our method (PCA + BRT + inequality) does not seem to suffer from finite-sample statistical bias.

[Insert Table 5 here]

4.5.1.2 Magnitude Transformation

A potential disadvantage with the dummy transformation is that it ignores the magnitude of the information variables. Therefore, the test statistic might not have enough power to reject against the null. Accordingly, we turn our attention to transformation 2, which utilises this feature of the data – Panel B of Table 5 presents the empirical results. The univariate results are largely consistent with the dummy transformation. The low corporate bond return, the low inflation rate, the low long term bond return, the downward sloping term structure and the lagged negative risk premium all succeed in identifying potential negative risk premium periods with the weighted conditional risk premiums -3.44%, -0.51%, -2.90%, -1.74%, -3.51% in the full sample period, respectively. Notably, except for conditioning on the downward sloping term structure, the negative conditional risk premiums decreases in the other conditioning periods, which suggest that

the magnitudes of the conditioning variables provide additional relevant information. One observation is that while the high inflation rate and low long term bond rate do not capture the negative risk premium under the dummy transformation, they are now able to pick up the negative risk premium periods. In contrast, high short-term rate and low volatility appear to provide little relevant information for the negative risk premium states, for either transformation method.

Turning to the joint statistic which takes into account the auto-cross correlations across the estimators, we find sufficient evidence against the positivity restriction of the risk premium. For the full period, the *W* statistic is 6.87. This value is: (a) significant at the 5% level based on the asymptotic distribution; and (b) significant at the 10% level (but only marginally above $p = 0.5$) for the simulation distribution. For the sub period, the *W* statistic of 5.37 delivers similar inferences, namely, rejection at the: (a) 10% level asymptotically; and (b) 10% level based on the empirical *p-value*.

Collectively, the joint test agrees with the univariate results that the negative risk premium can occur in some states of the economy. These states are associated with the periods when corporate and long government bond returns are low, when the inflation rates are high, when the long term rate is lower than short-term rate, and when the premium in previous period is negative. These periods are documented to link with the change in business cycles, which essentially the states when theories predict the negative risk premium can occur (Boudoukh, Richardson and Whitelaw, 1997, and Whitelaw, 2000).

4.5.1.3 Realised Risk Premium

For comparison purposes, in Table 6, we repeat the main empirical tests using the ex-post realised risk premium as the proxy for the ex-ante risk premium, instead of the forecasts from the two-stage model. With respect to the univariate tests, most of the instrumental variables are able

capture the negative realised risk premium periods, with the exception of the volatility. These results are consistent with the evidence above in the predicted premium case. Additionally, the conditional realised risk premiums are all noticeably larger than the conditional predicted risk premium from our two-stage method. For example, for the full sample and magnitude transformation, conditioning on the downward sloping term structure, the predicted mean of the risk premium is -1.74 (Panel B of Table 6), whereas the realised risk premium is -9.46%. Yet, the associated standard errors for the realised case are also larger than its predicted counterpart. This highlights a widely acknowledged important point that the realised risk premium is a noisy proxy for the ex-ante risk premium.

Indeed, although there is strong descriptive evidence that the negative realised risk premium can occur, we fail to formally reject the null hypothesis that the positivity restriction on the risk premium is violated. With respect to the dummy transformation, the *W* statistics are 2.07 (asymptotic *p-value* equal to 28%) to 3.18 (asymptotic *p-value* equal to 16%) in the full and pre-recession periods, respectively. Moreover, using magnitude transformation, we are likewise unable to reject the hypothesis at the 5% level. The corresponding asymptotic *p-values* are 28% and 37%. The realised risk premium seems too noisy and our joint statistics which incorporate the information of the variance-covariance matrix to account for noise and the auto-cross correlations of the estimators, are unable to reject the null.

To shed further light on the power of the test, we conduct a simulation experiment to assess the power of our PCA+BRT tests. The simulation procedure and results are reported in the Online Appendix A.2.5. The collectively evidence suggests that our method has reasonable power to detect a false null.

[Insert Table 6 here]

5. Concluding Remarks

A large number of asset pricing studies focus on testing the linear restrictions imposed by theoretical models, yet they mostly ignore the positivity condition of the risk premium. In this paper, we bring a fresh perspective and novel empirical testing strategy to this important restriction. Our study's main contribution lies in the effective and highly practical manner of modelling the conditional risk premium, in which we address two highly challenging issues surrounding the traditional linear instrumental variable approach. Specifically, in the context of a substantial "degrees of freedom" problem (having such a vast number of potential predictors, >> 100, relative to the time series sample size), we adopt Principal Component Analysis and the Boosted Regression Tree (BRT) techniques to alleviate criticisms pertaining to the identity of the investors' information set, and the non-linear structure of the return data generating process.

The empirical evidence suggests our two-stage methodology is advantageous compared to other linear methods in modelling the conditional risk premium. More importantly, we show that, in the US market, the positive risk premium condition is violated in some states of the economy, such as low corporate and government bond returns, downward-sloping term structure, high inflation, and lag negative risk premium periods. This key empirical result implies a rejection of the conditional CAPM. Accordingly, we are concerned about the current common practice of directly imposing the positive risk premium constraint in predictive models. As such, we urge researchers in this area to entertain the methods that we illustrate herein.

Online Appendices

Appendix A: Regression Trees and Model Validity Statistics

A.1 Regression Tree Technicalities

This appendix provides the interested reader, some further technical details on decision trees. A more detailed discussion can be found in Hastie, Tibshirani and Friedman (2009)).

Generally, the conditional expectation of a variable Y, the regression problem seeks to approximate the unknown function from basis function expansions of input X (information variables): $f(x) = E(Y|X = x) = \sum_{i=1}^{p} \beta_i h(x; \gamma)$, where β is the expansion coefficients and h is the basis functions of the input X, parameterized by γ. Traditionally, the linear regression is convenient because it approximates f(x) by first-order Taylor expansion and $h(x, \gamma) = x$. On the other hand, the regression trees method seeks to estimate the unknown functional form by dividing the predictor space into non-overlapping regions and simply models the response variable as a constant within each region. The splitting variables, splitting points and constant fit in each region form the nature of basis function $h(x, \gamma)$. Formally, in the first step the tree algorithm searches through the entire predictor space, and chooses an independent variable j and a splitting point s so that the two defined planes are given by:

$$R_1(j, s) = \{X|X_j \leq s\} \text{ and } R_2(j, s) = \{X|X_j > s\} \tag{A.1}$$

which minimises the sum of squared residuals criterion in the resulting regions. Specifically, the optimal parameters set (j, s) is the solution of:

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (rp_{t+1} - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (rp_{t+1} - c_2)^2 \right] \tag{A.2}$$

For a given pair (j, s), $(\hat{c}_1, \hat{c}_2)$ is the solution of the inner minimisation, which is just the average of the response variable in each region:[20]

$$\hat{c}_1 = \frac{1}{\sum_t^T I(x_t \in R_1(j,s))} \sum_t^T rp_{t+1} I(x_t \in R_1(j,s))$$
$$\hat{c}_2 = \frac{1}{\sum_t^T I(x_t \in R_2(j,s))} \sum_t^T rp_{t+1} I(x_t \in R_2(j,s)) \quad \text{(A.3)}$$

where rp is the risk premium or the response variable, and

$I(x \in R_i)$ is an indicator function equal to 1 if x falls into region R_i, and is 0 otherwise.

In a similar manner, the subsequent optimal split (a splitting point and a predictor) is decided by minimising the sum of squared residuals in the resulting regions from previous splits, instead of searching on the entire sample space. The process continues until some stopping criteria are satisfied. Suppose that there are M regions $R_1, R_2, \ldots, R_m$ at the final step, then the fitted value of the regression trees is:

$$\hat{f}(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \quad \text{(A.4)}$$

where $I(x \in R_m)$ is an indicator variable equal to 1 if $x \in R_m$ and is 0 otherwise,

$\mathbf{c_m}$ is the average of the response variable in region $\mathbf{R_m}$.

[20]That is what is meant by modelling the response variable as constant within each region.

A.2 Model Assessment

A.2.1 Relative Influence Measure and Partial Dependence Plots

Breiman et al. (1984) propose the influence measure to evaluate the contribution of the predictors in a single regression tree. The idea of this measure is straightforward. For example, at the j^{th} step, a predictor and a splitting point are chosen to partition an identified region from previous steps into two sub-regions. The predictor is one that satisfies the objective function [equation (A.2)], i.e. minimises the sum of squared errors in partitioned regions. Therefore, if a variable is chosen multiple times, its importance in the model can be measured as the sum of the reduction in squared errors $\hat{\imath}^2$ across the regions where it is chosen as the splitting variable, $ד_k^2$. Taking the average $ד_k^2$ across boosting iterations obtains the measure of influence of variable k, $\widehat{ד_k^2}$. Finally, the relative influence measure, RI_k of the variable k is calculated by dividing the influence measure $\widehat{ד_k^2}$ by the total influence of all variables across the boosting iterations.[21]

While the relative influence shows the importance of predictors in fitting future stock returns, partial dependence plots inform on the marginal effect of individual variables on the conditional expected return. The idea of this measure is that for each value of X_k, it averages the effect of all variables in X_C (the information set that excludes variable X_k), and thus tracks the effect of X_k on the predicted value of the response variable.

[21] To conserve space, the interested reader is referred to Breiman, Friedman, Stone and Ohlsen (1984) for the specific details.

A.2.2 Out-of-Sample R^2_{OOS}

Although the above measures offer useful information about the validity of the two-stage method, they do not directly indicate that the model is in fact superior in modelling conditional risk premium. To assess the performance of the two-stage method in modelling the conditional risk premium, we adopt the out-of-sample R^2 statistic that is commonly used in the literature (see, e.g., Li, Ng and Swaminathan (2013)):

$$R^2_{OOS} = 1 - \frac{\sum_{t=1}^{T}(rp_{t+1} - \hat{f}(x))^2}{\sum_{t}^{T}(rp_{t+1} - \overline{rp}_{t+1})^2} \quad (A.5)$$

where $\hat{f}(x)$ is the two-stage method's predicted value of expected risk premium,

$\overline{rp}_t$ is the historical average of stock excess returns until time $t-1$.

We assess the statistical significance of the R^2_{OOS} using the Clark and West (2007) adjusted statistic:

$$f_{t+1} = (rp_{t+1} - \overline{rp}_{t+1})^2 - ((rp_{t+1} - \widehat{rp}_{t+1})^2 - (\overline{rp}_{t+1} - \widehat{rp}_{t+1})^2) \quad (A.6)$$

The *p-value* of the R^2_{OOS} can be obtained by regressing the out-of-sample value f_t on a constant, and calculating the p-value of the one-sided test associated with the constant.

To gauge the economic significance of the R^2_{OOS}, we calculate the utility gains for mean-variance investors with relative risk aversion γ, who form monthly portfolios between stocks and bonds using forecasts from the predictive models as opposed to forecasts based on the historical mean. Specifically, the allocation to stocks in the next period based on historical forecasts of expected return and variance is illustrated below:

$$w_{0,t} = \left(\frac{1}{\gamma}\right)\left(\frac{\overline{rp}_{t+1}}{\widehat{\sigma}^2_{t+1}}\right) \quad (A.7)$$

Similarly, $w_{1,t}$ represents the allocation to stocks in the next period if the investors employ a predictive model of returns:

$$w_{1,t} = \left(\frac{1}{\gamma}\right)\left(\frac{\widehat{rp}_{t+1}}{\hat{\sigma}^2_{t+1}}\right) \tag{A.8}$$

In both investment decisions, we forecast variance of stock returns $\hat{\sigma}^2_{t+1}$ using a 10-year rolling window of monthly returns (Li, Ng and Swaminathan (2013)). The investors' average utility, based on the allocation using the historical mean and the predictive model over the out-of-sample period, are illustrated below:

$$U_0 = \mu_0 - \frac{1}{2}\gamma\,\hat{\sigma}^2_0 \tag{A.9}$$

and

$$U_1 = \mu_1 - \frac{1}{2}\gamma\,\hat{\sigma}^2_1 \tag{A.10}$$

where $(\mu_0, \hat{\sigma}^2_0)$ and $(\mu_1, \hat{\sigma}^2_1)$ are the means and variances of the portfolios' returns based on the historical mean and the predictive model forecasts, respectively.

The utility gain is the difference between U_1 and U_0. To report the annualised percentage return, we multiply $(U_1 - U_0)$ by 1200. The risk aversion coefficients $\gamma = 1, 3,$ and 5 are chosen for the main results (see, e.g., Campbell and Thompson (2008), among others).

In addition to R^2_{OOS}, we conduct the Mincer and Zarnowitz (1969) regression test of unbiased forecasts by simply regressing rp_{t+1} on the forecasts $\widehat{rp}_{t+1}$, out-of-sample, and jointly testing the null that the intercept is equal to 0 and the coefficient is equal to 1.

A.2.3 Forecast Encompassing Tests

We conduct the encompassing test to assess whether our forecasts provide additional information to that of alternative models. Following Rapach, Strauss and Zhou (2010) and others,

we use Harvey, Leybourne and Newbold (1998) *HLN-statistic* to test the null hypothesis that model i encompasses model j against the one-sided alternative hypothesis that model i does not encompass model j. Define $d_{t+1} = (e_{t+1}^{i} - e_{t+1}^{j})e_{t+1}^{i}$ where $e_{i,t+1} = rp_{t+1} - \widehat{f}_{i}(x)$ and $e_{j,t+1} = rp_{t+1} - \widehat{f}_{j}(x)$. The HLN-statistic, which follows a t_{q-1} distribution, can be calculated as follows:

$$\mathrm{HLN} = \mathrm{q}/\mathrm{q} - 1\left[\widehat{\mathrm{V}}(\bar{\mathrm{g}})^{-\frac{1}{2}}\right]\bar{\mathrm{g}} \tag{A.11}$$

where $\bar{\mathrm{g}} = \frac{1}{\mathrm{q}}\sum_{\mathrm{k}=1}^{\mathrm{q}} \mathrm{g}_{\mathrm{t+k}}$,

$\widehat{\mathrm{V}}(\bar{\mathrm{g}})^{-\frac{1}{2}} = \frac{1}{\mathrm{q}^{2}}\sum_{\mathrm{k}-1}^{\mathrm{q}}(\mathrm{g}_{\mathrm{t+k}} - \bar{\mathrm{g}})^{2}$,

q is the number of out-of-sample forecasts.

A.2.4 Small-Sample Bias and Estimated Parameters

First, we generate T observations of the risk premium from a normal distribution with the mean and the variance equal to those of the realised sample. We set T=684 and 576 to match the number of observations in the full sample and subsample (pre-2008), respectively. Note that the sample unconditional mean of the risk premium is positive, therefore we maintain that the null is true. Second, the instrumental variables X are simulated from the vector autoregression equation:

$$\begin{pmatrix} \mathrm{X}_{1,\mathrm{t}} \\ \mathrm{X}_{2,\mathrm{t}} \\ \vdots \\ \mathrm{X}_{\mathrm{K,t}} \end{pmatrix} = \begin{pmatrix} \rho_1 & 0 & \cdots & 0 \\ 0 & \rho_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \rho_{\mathrm{K}} \end{pmatrix} \begin{pmatrix} \mathrm{X}_{1,\mathrm{t-1}} \\ \mathrm{X}_{2,\mathrm{t-1}} \\ \vdots \\ \mathrm{X}_{\mathrm{K,t-1}} \end{pmatrix} + \sum \tag{A.12}$$

The autocorrelation parameters $(\rho_1, \rho_2, \ldots, \rho_k)$ are the full sample estimates. The instruments' initial values are set at their unconditional means. To incorporate cross-dependencies, we draw the error terms $\sum$ from a multivariate normal distribution with mean zero and sample

covariance matrix of the predictors. To pick the instruments in the inequality testing, we select a random subset of seven variables without replacement from the full set of 156 predictors to match our choice in the empirical implementation of the realised sample. In each simulation we run the two-stage method and inequality test, keeping the exact specifications as in our main empirical test to account for estimated parameters. In this setting, due to extreme computational burden, we rely on 1,000 simulations and record the associated statistics. The empirical *p-value* for the *W* statistics is calculated as the fraction of simulated *W* values exceeding the *W* statistics obtained from our actual samples.[22]

A.2.5 Test Power

In the simulation experiment, we incorporate predictability into our testing by assuming the risk premium is generated from a small set of predictors. First, we generate the instruments as in the first experiment. We then generate the risk premium from the linear predictive regression below:

$$\mathrm{Y_{t+1}} = \beta' \mathrm{X_t} + \epsilon_{t+1} \qquad \text{(A.13)}$$

[22] Generally, a larger number of simulations deliver a more accurate test, but at a diminishing rate. In reality, the test accuracy should be traded off against the computational intensity, linked to the computational "cost". Indeed, echoing the arguments of Harvey and Liu (2017, fn 9), our choice here of 1,000 simulations, contrasting the 10,000 simulations executed in estimating the distribution weights, is purely because of the computational burden in running the boosted regression trees. For consistency, we also re-estimate the distribution weights under 1,000 simulations. The results are qualitatively the same.

The X variables are a constant term, dividend price ratio (*dp*), earning price (*ep*) ratio, net payout ratio (*ntis*). We choose the constant term so that the unconditional mean is equal to the sample average of the risk premium. The predicting variables are selected because they are highly autocorrelated which is the source of our problem. The X variables are all standardised to have mean zero and unit standard deviation. Assuming all the coefficients are equal, we calculate them so that the R^2 of the regression A.13 is 10% and the variance of the risk premium is equal to its sample counterpart. To do so, we start with the $R^2 = \frac{\beta'\Omega\beta}{\beta'\Omega\beta+\sigma_\epsilon^2} = \frac{\beta'\Omega\beta}{\sigma_Y^2}$ where Ω is the covariance matrix of the predictors. We then fix the R^2 and σ_Y^2 and solve for β. Therefore, the error terms come from a normal distribution with mean zero and variance equal to $9\beta'\Omega\beta$. To ensure our null is true at all times, we can replace the negative predicted risk premium values with zero. i.e. $E_t(Y_{t+1}) = max\big(0, E_t(Y_{t+1})\big)$. Note that we only truncate predicted value to ensure that we can observe negative risk premium values due to noise. We run 1,000 simulations and record the W distribution, therefore *W* cut-off points, W_{crit_sim}.

We assess the power of the test, by entertaining the alternative that the risk premium can be negative. In other words, we do not restrict the predicted risk premium to be always positive. Therefore, for a given size, the fraction of times the simulated *W* statistics in this experiment exceeding the critical cut-off values generated under the true null W_{crit_sim} determines the power of the test. We assess the power at the 5% size test.

At 5% level, the test rejects the null hypothesis 27% of the time. Recall that our procedure only chooses a random set of 7 variables out of 156 variables. Ideally, we would like to detect the negative risk premium using variables which are motivated by sound theory(ies). The advantage of focusing on a set of "theory-driven" variables is that it is more likely to contain the "true" set of variables that predict the risk premium. In fact, if we choose a random set of 7 variables from

Goyal and Welch, 2008 which contains our 3 true predictive variables, with the 5% critical cut-off value, we can reject the null 40% of the time. Collectively, the simulation result suggests that the test has power to detect a false null.

Appendix B: **Instrumental Variable Transformation Methods**

In the dummy transformation I, the low corporate bond return state is defined when it lies below the unconditional mean, and takes the value of $z^{+}_{corp,t} = 1$, and 0 otherwise. Similarly, $z^{+}_{ltr,t} = 1$ indicates the state when long term government bond returns are below its long run mean, and 0 otherwise. $z^{+}_{nrp,t} = 1$ is when the previous period has a negative risk premium. The term structure in set B $z^{+}_{tms,t} = 1$ is when it is downward sloping. For both inflation rate and risk free rate, $z^{+}_{infl,t}$ and $z^{+}_{rfr,t}$ are equal to 1 when their values are greater than the long run mean. Finally, $z^{+}_{svar,t} = 1$ indicates states when svar is lower than the long run average.

For the magnitude transformation II, we aim to improve the power of the test and utilise the economic magnitude of states corresponding to low risk premium periods. Unlike the dummy transformation which put equal weights in calculating the conditional risk premium in the periods of predicted negative risk premium, transformation 2 attempts to calculate the conditional means of the risk premium, weighted by the magnitudes of the instruments. In particular, we define the low corporate bond return state as $z^{+}_{corp,t} = |\, min(0, corpr_t - E[corpr_t])|$, where $\mathrm{E[corpr_t]}$ is the long run mean of corporate bond returns. In a similar manner, the low long term government rate of return state and low volatility state are $z^{+}_{ltr,t} = |\, min(0, ltr_t - E[ltr_t])|$ and $z^{+}_{svar,t} = |\, min(0, svar_t - E[svar_t])|$, respectively. On the other hand, $z^{+}_{infl,t} = max(0, corpr_t - E[corpr_t])$ and $z^{+}_{rfr,t} = max(0, corpr_t - E[corpr_t])$ are defined as high inflation and high risk

free states. We define $z_{3,t}^{+} = max\left(0, -\widehat{(rp_t)}\right)$ as the lagged negative risk premium, and $z_{tms,t}^{+} = max(0, -tms_t)$ as downward-sloping term structure.

To provide an economic interpretation, we normalise these variable in both transformations as $z_{i,t}^{*} = \frac{z_{i,t}^{+}}{E\left(z_{i,t}^{+}\right)}$ where $E\left(z_{i,t}^{+}\right)$ is the sample mean of $z_{i,t}^{+}$ in the context $\hat{\theta}_{\mu z_i}^{+} = \frac{1}{T}\sum_{t=1}^{T}\left[\hat{f}(x) z_{it}^{*}\right]$. For example, with respect to transformation 1 and the normalised instrument $z_{\text{infl,t}}^{*}$, the estimator $\hat{\theta}_{\mu z_{infl}}^{+}$ is the sample mean of the risk premium, conditional on the high inflation rates. In transformation 2, the corresponding economic interpretation of $\hat{\theta}_{\mu z_{infl}}^{+}$is that it reflects the conditional mean of the risk premium, weighted most by the high inflation rates.

References

Andrews, D. W. K., 1991, Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 817-858.

Ang, A., Bekaert, G., 2007, Stock return predictability: Is it there? *Review of Financial Studies* 20, 651-707.

Bali, T. G., Hu, J., Murray, S., 2019, Option implied volatility, skewness, and kurtosis and the cross-section of expected stock returns. *Georgetown McDonough School of Business Research Paper*.

Boudoukh, J., Richardson, M., Smith, T., 1993, Is the ex ante risk premium always positive?: A new approach to testing conditional asset pricing models. *Journal of Financial Economics* 34, 387-408.

Boudoukh, J., Richardson, M., Whitelaw, R. F., 1997, Nonlinearities in the Relation Between the Equity Risk Premium and the Term Structure. *Management Science* 43, 371-385.

Boudoukh, J., Richardson, M., Whitelaw, R. F., 2008, The Myth of Long-Horizon Predictability. *Review of Financial Studies* 21, 1577-1605.

Brav, A., Lehavy, R., Michaely, R., 2005, Using expectations to test asset pricing models. *Financial management*, *34*, 31-64.

Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A., 1984, *Classification and regression trees*: (CRC press).

Campbell, J. Y., Thompson, S. B., 2008, Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *Review of Financial Studies* 21, 1509-1531.

Chen, L., Zhao, X., 2009, Return Decomposition. *Review of Financial Studies* 22, 5213-5249.

Chinco, A., Clark-Joseph, A. D., Ye, M., 2019, Sparse Signals in the Cross-Section of Returns. *The Journal of Finance* 74, 449-492.

Clark, T. E., West, K. D., 2007, Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics* 138, 291-311.

Cochrane, J. H., 2011, Presidential Address: Discount Rates. *The Journal of Finance* 66, 1047-1108.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004, Least angle regression. *Annals of Statistics* 32, 407-499.

Elton, E. J., 1999, Expected Return, Realized Return, and Asset Pricing Tests. *The Journal of Finance* 54, 1199-1220.

Foster, F. D., Smith, T., Whaley, R., 1997 Assessing Goodness-of-Fit of Asset Pricing Models : The Distribution of the Maximal R2." *The Journal of Finance* 52, 591-607.

Friedman, J. H., 2002 "Stochastic gradient boosting." *Computational Statistics and Data Analysis* 38, 367-378.

Goyal, A., Welch, I., 2008, A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21, 1455-1508.

Harvey, C. R., Liu, Y., Zhu, H., 2016, … and the Cross-Section of Expected Returns. *Review of Financial Studies* 29, 5-68.

Harvey, D. I., Leybourne, S. J., Newbold, P., 1998, Tests for Forecast Encompassing. *Journal of Business & Economic Statistics* 16, 254-259.

Hastie, T., Tibshirani, R., Friedman, J., 2009, *The Elements of Statistical Learning*. (New York USA: Springer).

Huang, D., Jiang, F., Tu, J., Zhou, G., 2015, Investor sentiment aligned: A powerful predictor of stock returns. *Review of Financial Studies* 28, 791-837.

Jurado, K., Ludvigson, S. C., Ng, S., 2015, Measuring Uncertainty. *American Economic Review* 105, 1177-1216.

Kelly, B., Pruit, S., 2013, Market Expectations in the Cross-Section of Present Values. *The Journal of Finance* 68, 1721-1756.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S., 2018, Human Decisions and Machine Predictions. *The Quarterly Journal of Economics* 133, 237-293.

Kozak, S., Nagel, S., Santosh, S., 2017, "Shrinking the cross section." *Journal of Financial Economics* Forthcoming.

Kozak, S., Nagel, S., Santosh, S., 2018, Interpreting Factor Models, *The Journal of Finance 73*, 1183-1223.

Li, Y., Ng, D. T., Swaminathan, B., 2013, Predicting market returns using aggregate implied cost of capital. *Journal of Financial Economics* 110, 419-436.

Ludvigson, S. C., Ng, S., 2007, The empirical risk-return relation: A factor analysis approach. *Journal of Financial Economics* 83, 171-222.

Lundblad, C., 2007, The risk return tradeoff in the long run: 1836-2003. *Journal of Financial Economics* 85, 123-150.

Martin, I., 2017, What is the Expected Return on the Market? *The Quarterly Journal of Economics 132*, 367-433.

Martin, I., Wagner, C., 2019 "What is the Expected Retrun on a Stock? *The Journal of Finance*, forthcoming.

Merton, R. C., 1980, On estimating the expected return on the market. *Journal of Financial Economics* 8, 323-361.

Mincer, J., Zarnowitz, V., 1969, The Evaluation of Economic Forecasts. *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance. NBER Working Papers*, 3-46.

Newey, W. K., West, K. D., 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix. *Econometrica* 55, 703-708.

Ostdiek, B., 1998, The world ex ante risk premium: an empirical investigation. *Journal of International Money and Finance* 17, 967-999.

Pástor, Ľ., Sinha, M., Swaminathan, B., 2008, Estimating the intertemporal risk–return tradeoff using the implied cost of capital. *The Journal of Finance 63*, 2859-2897.

Pettenuzzo, D., Timmermann, A., Valkanov, R., 2014, Forecasting stock returns under economic constraints. *Journal of Financial Economics* 114, 517-533.

Rapach, D. E., Strauss, J. K., Zhou, G., 2010, Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies* 23, 821-862.

Rossi, A. G., Timmermann, A., 2015, "Modeling Covariance Risk in Merton's ICAPM." *Review of Financial Studies* 28, 1428-1461.

Rossi, B., Inoue, A., 2012, Out-of-Sample Forecast Tests Robust to the Choice of Window Size. *Journal of Business & Economic Statistics* 30, 432-453.

Stock, H.J., Watson, M.K., 2002, Forecasting Using Principal Components from a Large Number of Predictors, *Journal of the American Statistical Association* 97, 1167-1179.

Walsh, K., 2015, The investment horizon and asset pricing models. *Australian Journal of Management* 40, 277-294.

Whitelaw, R. F., 2000, Stock Market Risk and Return: An Equilibrium Approach. *Review of Financial Studies* 13, 521-547.

Wolak, F. A., 1989, Testing inequality constraints in linear econometric models. *Journal of Econometrics* 41, 205-235.

Wolak, F. A., 1991, The Local Nature of Hypothesis Tests Involving Inequality Constraints in Nonlinear Models. *Econometrica* 59, 981-995.

Figure 1

A Visual Demonstration of the Regression Tree Method

Figure 1 presents a visual demonstration of the Regression Tree method in modelling the risk premium during the full sample period 1960:1-2016:12. The Regression Tree method approximates the conditional risk premium by sequentially breaking the predictor space into sub-regions and fitting a constant value of the risk premium in each region. The squares represent the sub-regions. The value within each square represents the fitted constant value of the risk premium. The left (right) hand sides of the inequalities represent the variables (values) that are used to split the predictor space. The predictors are from Goyal and Welch (2008): Log dividend price ratio (*dp*); Earning price ratio (*ep*); Default yield spread calculated as the difference between BAA and AAA corporate bond yields (*def*); Term structure (*tms*); Long term government bond yield (*lty*); Long term bond return (*ltr*); Stock variance measured as squared daily returns (*svar*); Three-month T-bill rate (*tbl*); Inflation rate (*infl*); Lag excess returns (*lret*); Net equity expansion is measured as the sum of 12 months net issue for NYSE stocks divided by the market capitalisation of the stocks (*ntis*); Book to market ratio (*bm*), and Corporate bond return (*corpr*). *PC1* to *PC10* refer to the first 10 principal components.

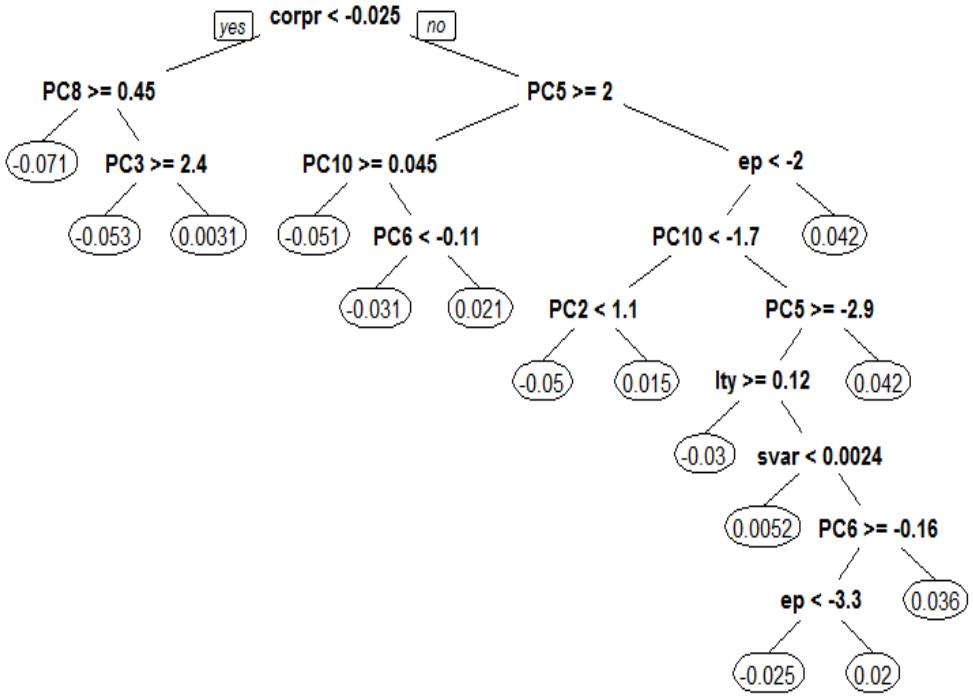

Figure 2
Marginal Effect of the Information Variables on the Expected Return

Figure 2 presents partial dependence plots for the risk premium, based on five predictors during the full sample period 1960:1-2016:12: namely, corporate bond returns ($corpr$), 5^{th}principal component ($PC5$), stock variance ($svar$), term structure (tms), and the dividend/price ratio (dp). The horizontal axis presents the sample values of the predictors. The vertical axis illustrates the conditional risk premium as a function of the individual predictor.

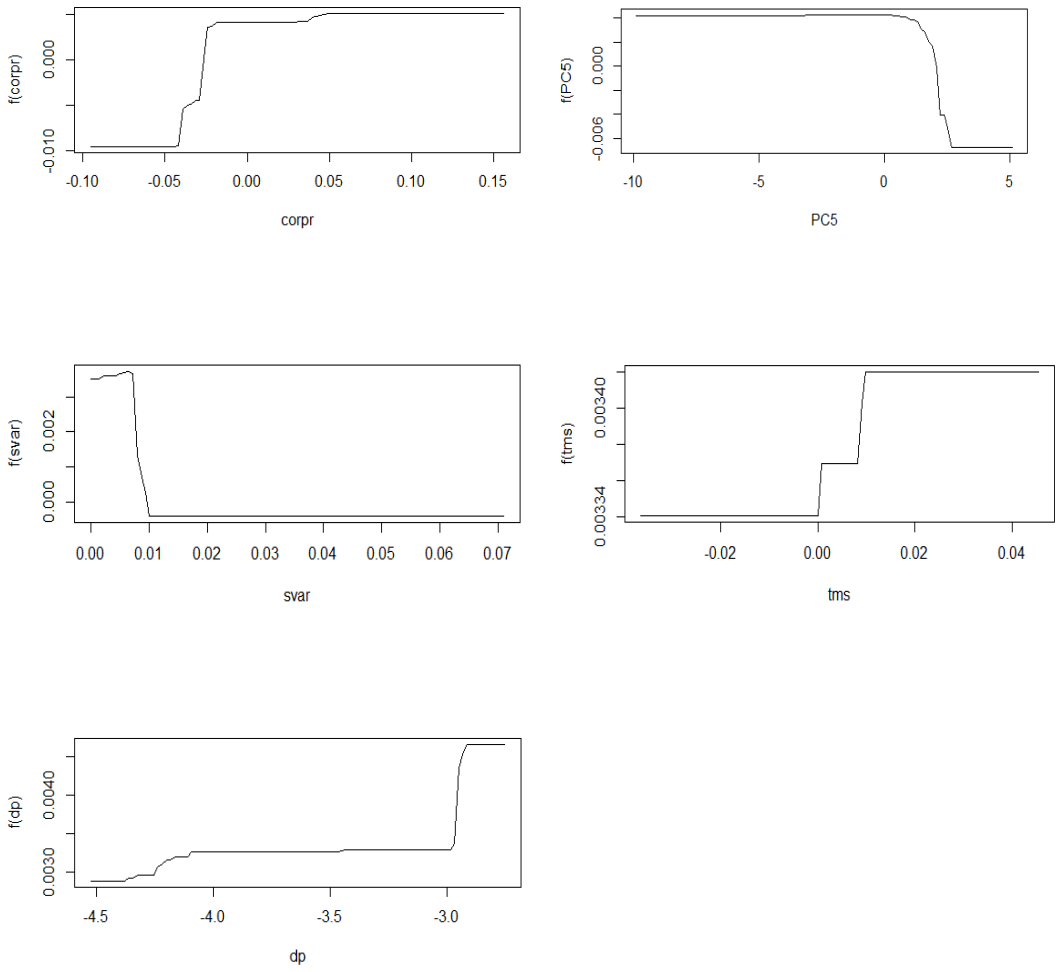

Figure 3
Out-of-sample R^2_{OOS} by sample split date
1-month returns

Figure 3 presents out-of-sample R^2_{OOS} of our method, based on monthly forecasts, across different sample split dates (solid line). We also report confidence intervals (90%, 95%, and 99%) of the R^2_{OOS} statistic based on simulated samples discussed in Online Appendix A.2.4.

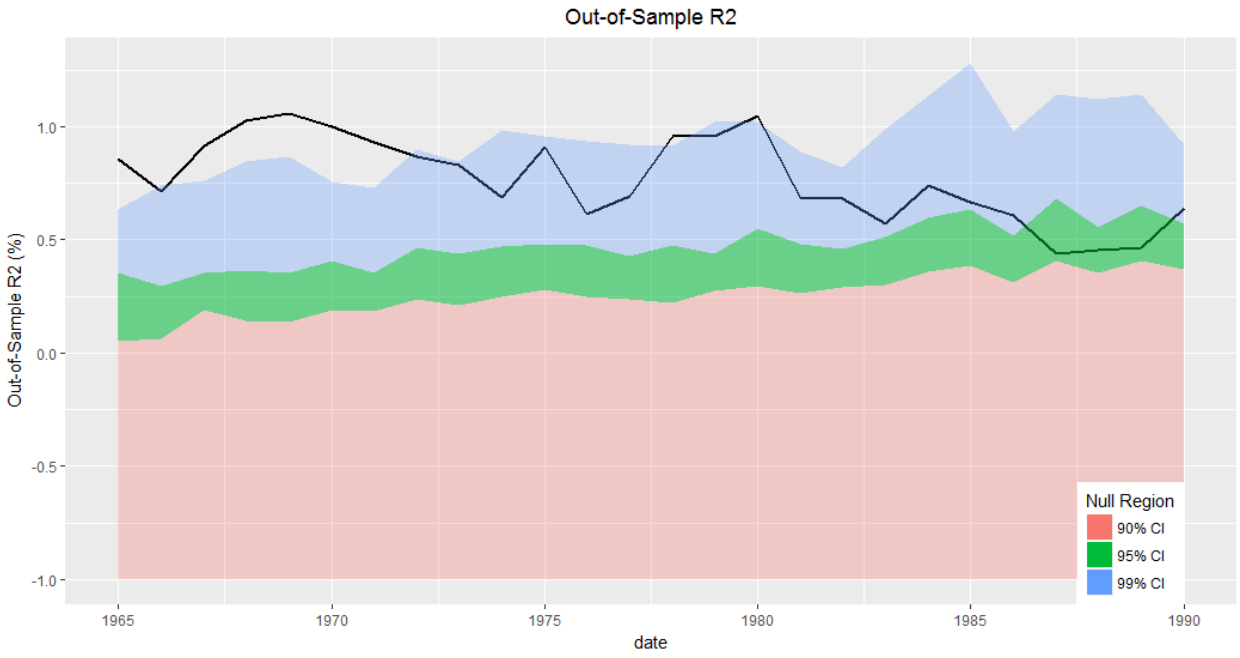

Table 1

Summary Statistics

Table 1 presents the summary statistics of the monthly risk premium on the CRSP US market $\hat{rp}$ from 1970:1 to 2016:12 estimated by the two-stage method comprising Principal Component Analysis and Boosted Regression Trees, and 13 information variables documented in Goyal and Welch (2008) (Panel A) from 1960:1 to 2016:12. These predictors are: Log dividend price ratio (*dp*); Earning price ratio (*ep*); Default yield spread calculated as the difference between BAA and AAA corporate bond yields (*def*); Term structure (*tms*); Long term government bond yield (*lty*); Long term bond return (*ltr*); Stock variance measured as squared daily returns (*svar*); Three-month T-bill rate (*tbl*); Inflation rate (*infl*); Lag excess returns (*lret*); Net equity expansion is measured as the sum of 12 months net issue for NYSE stocks divided by the market capitalisation of the stocks (*ntis*); Book to market ratio (*bm*), and corporate bond return (*corpr*). Panel B reports the relative and cumulative importance of the first 10 common components, R^2 and Cum, respectively. R^2 shows the fraction of total variance in the original information variables collected from Ludvigson and Ng (2007) and Goyal and Welch (2008). Panel C reports the relative influence (*RI%*), a statistic measuring the importance of the information variables in the Boosted Regression Trees stage, of 10 principal components and 13 information variables in Goyal and Welch (2008). The cumulative relative influence (Cum. RI) of the Top 3, 5, and 10 are also reported in Panel C.

Panel A: Instrumental Variables

	Mean	Min	Q1	Median	Q3	Max	Std. Dev.
$\hat{rp}$	**0.0033**	**-0.0144**	**0.0019**	**0.0045**	**0.0059**	**0.0156**	**0.0042**
$b.m$	0.5024	0.1205	0.3059	0.4592	0.6518	1.2065	0.2571
lty	0.0651	0.0175	0.0441	0.0612	0.0807	0.1482	0.0267
$ntis$	0.0113	-0.0577	0.0020	0.0142	0.0252	0.0512	0.0194
$Rfree$	0.0039	0.0000	0.0023	0.0040	0.0052	0.0136	0.0026
$infl$	0.0031	-0.0192	0.0007	0.0029	0.0051	0.0181	0.0036
ltr	0.0062	-0.1124	-0.0103	0.0043	0.0229	0.1523	0.0292
$corpr$	0.0063	-0.0949	-0.0071	0.0051	0.0191	0.1560	0.0259
$svar$	0.0021	0.0001	0.0006	0.0011	0.0021	0.0709	0.0044
ep	-2.8398	-4.8365	-3.0491	-2.8727	-2.6445	-1.8987	0.4281
dp	-3.5881	-4.5236	-3.9123	-3.5201	-3.3333	-2.7533	0.3943
def	0.0102	0.0032	0.0073	0.0090	0.0121	0.0338	0.0045
tms	0.0182	-0.0365	0.0072	0.0182	0.0301	0.0455	0.0146
$lret$	0.0040	-0.2482	-0.0195	0.0080	0.0314	0.1492	0.0428

Panel B: Principal Components (%)

	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	$PC6$	$PC7$	$PC8$	$PC9$	$PC10$
R^2	0.63	0.05	0.03	0.03	0.02	0.01	0.01	0.01	0.01	0.01
Cum. R^2	0.63	0.67	0.71	0.74	0.75	0.77	0.78	0.79	0.79	0.80

Panel C: Relative Influence (%)

Rank	1	2	3	4	5	6	7	8	9	10	11
Variable	$PC5$	$corpr$	$svar$	$b.m$	$lret$	dp	$PC8$	tms	$infl$	ep	ltr
RI	26.19	25.59	10.10	4.01	3.88	3.45	3.07	3.05	2.84	2.24	2.20
Rank	12	13	14	15	16	17	18	19	20	21	22	23
Variable	$PC9$	$PC7$	$PC1$	$PC6$	$PC4$	$ntis$	$PC10$	$PC3$	$Rfree$	def	lty	$PC2$
RI	2.05	1.91	1.81	1.80	1.64	1.37	0.88	0.63	0.42	0.33	0.31	0.23

	Top 3	Top 5	Top 10
Cum. RI	61.89	69.78	84.42

Table 2
Out-of-sample Prediction Performance of our Boosted Regression Tree Approach

Table 2 compares the out-of-sample predictive performance of the two-stage Boosted Regression Tree method described in Section 2.1, as opposed to those of kitchen sink ordinary least squares (KS), Least Angle Regression (LAR), historical average (Historical), and the mean combination (Mean), from 1970:1 to 2016:12. The benchmark Boosted Regression Trees uses 1,000 boosting iterations, 10 principal components as well as 13 economic variables in Goyal and Welch (2008). The LAR forms predictions based on the best 3 variables. Panel A reports the out-of-sample R^2_{OOS} and their Clark and West (2007) one-sided *p-values*. The economic significance of the R^2_{OOS}, measured by the average utility gain for mean-variance investors with three different relative risk aversion coefficients $\gamma = 1,3,$ and 5, are presented. Panel B displays the Mincer-Zarnowitz misspecification test. Panel C reports the correlation matrix between the forecasts formed by different methods. Panel D presents the p-value of the Harvey, Leybourne and Newbold (1998) *HLN* encompassing statistics discussed in Online Appendix A.2.3.

Panel A: Out of Sample R^2_{OOS}

	R^2_{OOS}	*p-value*	$\gamma = 1$	$\gamma = 3$	$\gamma = 5$
BRT	1.00%	0.03	3.63	1.21	0.57
KS	-9.71%	0.04	4.64	1.94	-0.60
LAR	0.52%	0.10	3.03	0.20	-0.48
Mean	0.70%	0.05	2.72	1.37	0.81

Panel B: Mincer-Zarnowitz Regression Test

	Intercept	Coefficient	*p-value*
BRT	0.0015	0.88	0.81
KS	0.0035	0.24	0.00
LAR	0.0021	0.63	0.68
Mean	-0.0001	1.56	0.65

Panel C: Correlation Matrix

	Mean	Historical	KS	LAR	BRT
Mean	1.00	0.56	0.46	0.59	0.71
Historical	-	1.00	0.08	0.21	0.20
KS	-	-	1.00	0.25	0.41
LAR	-	-	-	1.00	0.59
BRT	-	-	-	-	1.00

Panel D: Encompassing Test

	Mean	Historical	KS	LAR	BRT
Mean	-	0.87	0.08	0.33	0.09
Historical	0.05	-	0.04	0.10	0.03
KS	0.00	0.00	-	0.00	0.00
LAR	0.22	0.41	0.09	-	0.11
BRT	0.38	0.54	0.12	0.41	-

Table 3

Boosted Regression Tree Out-of-Sample Prediction with Different Specifications

Table 3 reports the predictive performance of the two-stage Boosted Regression Tree method, discussed in Section 2.1 under various specifications. Panel A presents R^2_{OOS}, Clark and West (2007) one-sided *p-values*, and the average utility gain for mean-variance investors with a relative risk aversion coefficient equal to 3, when 0, 1, 3, 5, 10, 20, 50, 100, and 120 components are included along with 13 economic variables from Goyal and Welch (2008) in the Boosted Regression Trees stage. Panel B reports similar statistics when the number of boosting iterations ranges from 1,000 to 5,000. Panel C shows the predictive performance of the method in different estimation and evaluation periods. The periods 1970:1-2016:12, 1980:1-2016:12 and 1990:1:2016:12 reflect that the initial training periods, observation numbers: 120, 240, and 360, respectively. Pre-Recession 1970:1-2007:12, Oil Shock 1976:1-2016:12, Oil Shock and Pre-Recession 1976:1-2007:12, Technology Bubble 2000:1-2016:12, and Technology Bubble Pre-Recession 2000:1-2007:12 periods are also under examination. Panel D reports R^2_{OOS} the long-horizon predictability. The Clark and West (2007) *p-value* is derived from Newey and West (1987) *t-statistic*.

Panel A: Predictive Performance while increasing retention of Principal Components

	R^2_{OOS}	*p-value*	$\gamma = 3$
0 Component	0.31%	0.11	-3.01
1 Component	0.95%	0.05	1.10
3 Components	0.94%	0.05	0.72
5 Components	0.96%	0.03	0.82
10 Components	1.00%	0.03	1.21
20 Components	0.72%	0.05	1.032
50 Components	0.28%	0.13	0.51
100 Components	0.38%	0.10	0.45
120 Components	0.43%	0.09	0.77

Table 3 (*Continued*)

Panel B: Predictive Performance while increasing the Number of Boosting Iterations

	R^2_{OOS}	*p-value*	$\gamma = 3$
1,000 Iterations	1.00%	0.03	1.21
2,000 Iterations	0.8%	0.03	1.22
3,000 Iterations	0.2%	0.02	1.02
4,000 Iterations	-0.6%	0.02	0.74
5,000 Iterations	-1.63%	0.03	0.37

Panel C: Predictive Performance with changing Estimation and Evaluation Periods

	R^2_{OOS}	*p-value*	$\gamma = 3$
120 Periods 1970:1-2016:12	1.00%	0.03	1.21
240 Periods 1980:1-2016:12	1.00%	0.02	0.44
360 Periods 1990:1-2016:12	0.64%	0.10	0.81
Pre-Recession 1970:1-2007:12	1.2%	0.03	1.13
Oil Shock 1976:1-2016:12	0.91%	0.03	0.34
Oil Shock Pre-Recession 1976:1-2007:12	1.32%	0.03	-0.10
Technology Bubble 2000:1-2016:12	0.14%	0.26	0.30
Technology Bubble Pre-Recession 2000:1-2007:12	0.91%	0.14	-0.28

Panel D: Predictive Performance for Increasing Horizons

	R^2_{OOS}	*p-value*
1-month	1.00%	0.03
3-month	4.10%	0.01
6-month	9.97%	<0.01
9-month	15.32%	<0.01
12-month	19.44%	<0.01
24-month	31.17%	0.01
36-month	42.24%	0.05

Table 4

Instrumental Variables Summary Statistics

Table 4 reports the summary statistics for the instrumental variables employed in the multiple inequalities framework. The instruments are the high T-bill rate z_{rfr}, low corporate bond return z_{corpr}, high inflation rate z_{infl}, low long-term government bond return z_{ltr}, downward sloping term structure of interest rates z_{tms}, low volatility z_{vol}, and lagged negative risk premium z_{nrp}. "Low" ("High") is defined as falling below (above) the long run mean. The table uses the abbreviations *Autocor.* and *Prob. of States* for autocorrelation and probability of states, respectively. The autocorrelation reports the autocorrelation coefficient order of 1. The probability of states indicates the probability that the states occur in time. The statistics are calculated for the whole out-of-sample period from 1970:01 and 2016:12. Details of the instruments construction are discussed in Online Appendix B.

	Autocor.	*Prob. of States*	*Correlation Matrix*						
			z_{rfr}	z_{corpr}	z_{infl}	z_{ltr}	z_{tms}	z_{vol}	z_{nrp}
z_{rfr}	0.94	56.56	1.00	0.17	0.50	0.13	0.61	-0.16	0.22
z_{corp}	-0.01	48.94	-	1.00	0.18	0.84	0.08	-0.09	0.14
z_{infl}	0.23	48.23	-	-	1.00	0.12	0.42	-0.12	0.34
z_{ltr}	-0.02	49.29	-	-	-	1.00	0.08	-0.05	0.10
z_{tms}	0.27	8.87	-	-	-	-	1.00	-0.06	0.21
z_{vol}	0.09	70.57	-	-	-	-	-	1.00	-0.19
z_{nrp}	0.12	22.16	-	-	-	-	-	-	1.00

Table 5
Inequality Tests on the Positivity Restriction of the Ex-ante Risk Premium

Table 5 reports the tests of the hypothesis: whether or not the ex-ante equity risk premium is always positive. The tests use two sample periods: 1970:1-2016:12 and 1970:1-2007:12. The predicted values from the two-stage method serve as the proxy for the expected risk premium. The empirical results for transformation 1 (dummy) and transformation 2 (magnitude) are presented in Panels A and B, respectively. Transformation methods are discussed in Online Appendix B. The annualised conditional risk premiums and the associated standard errors conditioning on high T-bill rate z_{rfr}, low corporate bond return z_{corpr}, high inflation rate z_{infl}, low long term government bond return z_{ltr}, downward sloping term structure of interest rates z_{tms}, low volatility z_{vol}, and lagged negative risk premium z_{nrp}. The standard errors are annualised and shown in percentage terms. We use the Quadratic Spectral kernel with bandwidth parameter estimated by fitting a univariate AR(1) model to each element of $\theta_{\mu z^+}$ (Andrews (1991)), to account for conditional cross-correlation and autocorrelation.. Asymptotic *W Stat* and *p-value* are the joint test statistics of multiple inequality restrictions. Empirical p-value is the fraction of times that the simulated *W* values exceed the sample *W* statistic. The simulation details are discussed in Online Appendix A.2.4.

Panel A: Transformation 1 (Dummy)

	1970:01-2016:12		1970:01-2007:12	
	Conditional Risk Premium	Standard Error	Conditional Risk Premium	Standard Error
z_{rfr}	4.99	1.09	6.03	1.30
z_{corp}	0.69	0.64	-0.11	0.67
z_{infl}	2.62	0.97	2.14	1.13
z_{ltr}	0.78	0.63	0.72	0.67
z_{tms}	-3.52	1.94	-3.57	2.39
z_{vol}	4.15	0.57	3.48	0.67
z_{nrp}	-1.30	1.16	-1.38	1.32
Joint Statistics	3.30		2.23	
Asym. *p-value*	0.13		0.20	
Empirical *p-value*	0.09		0.10	

Panel B: Transformation 2 (Magnitude)

	1970:01-2016:12		1970:01-2007:12	
	Conditional Risk Premium	Standard Error	Conditional Risk Premium	Standard Error
z_{rfr}	3.44	2.38	4.24	2.98
z_{corp}	-3.44	1.31	-2.87	1.25
z_{infl}	-0.51	1.66	-1.20	2.04
z_{ltr}	-2.90	1.12	-2.31	1.04
z_{tms}	-1.74	2.83	-2.14	3.50
z_{vol}	3.49	0.48	3.45	0.55
z_{nrp}	-3.51	2.04	-3.01	2.00
Joint Statistics				
W Stat	6.87		5.37	
Asym. *p-value*	0.03		0.06	
Empirical *p-value*	0.05		0.05	

Table 6
Inequality Tests on the Positivity Restriction applied to the Realised Risk Premium

Table 6 reports the tests of the hypothesis: whether or not the ex-ante equity risk premium is always positive. The tests use two sample periods: 1970:1-2016:12 and 1970:1-2007:12. The ex-post realised risk premium serves as the proxy for the expected risk premium. The empirical results are shown for transformation 1 (dummy) and transformation 2 (magnitude) in Panels A and B, respectively. Detailed discussion of the transformation methods is presented in Online Appendix B. The annualised conditional risk premiums and the associated standard errors conditioning on high T-bill rate z_{rfr}, low corporate bond return z_{corpr}, high inflation rate z_{infl}, low long term government bond return z_{ltr}, downward sloping term structure of interest rates z_{tms}, low volatility z_{vol}, and lagged negative risk premium z_{nrp}. The standard errors are annualised and shown in percentage terms. All of the estimates are corrected for conditional cross-correlation and autocorrelation. We use the Quadratic Spectral kernel with bandwidth parameter estimated by fitting a univariate AR(1) model to each element of $\theta_{\mu z^+}$ (Andrews, 1991). *W stat* and *p-value* are the joint test statistics of multiple inequality restrictions. Crit. Val 5% represents the critical value of *W* statistic at the 5% level of significance.

Panel A: Transformation 1 (Dummy)

	1970:01-2016:12		1970:01-2007:12	
	Conditional Risk Premium	Standard Error	Conditional Risk Premium	Standard Error
z_{rfr}	2.98	3.62	7.58	4.48
z_{corp}	-2.95	3.25	-1.38	3.47
z_{infl}	-0.93	3.67	-0.96	4.44
z_{ltr}	1.48	2.99	-1.51	3.48
z_{tms}	-10.13	7.25	-15.94	8.93
z_{vol}	1.96	1.98	4.06	2.29
z_{nrp}	5.40	4.19	1.83	4.69
Statistics				
W Stat	2.07		3.18	
Asymp. *p-value*	0.28		0.16	

Panel B: Transformation 2 (Magnitude)

	1970:01-2016:12		1970:01-2007:12	
	Conditional Risk Premium	Standard Error	Conditional Risk Premium	Standard Error
z_{rfr}	-0.10	5.24	-0.32	6.49
z_{corp}	-10.60	7.22	-7.21	6.14
z_{infl}	-3.11	5.21	-5.33	6.30
z_{ltr}	-2.24	4.83	-2.16	4.97
z_{tms}	-9.46	11.37	-11.01	14.00
z_{vol}	4.42	1.82	4.40	2.14
z_{nrp}	-0.13	7.41	3.06	7.35
Joint Statistics				
W Stat	2.55		1.70	
Asym. *p-value*	0.28		0.37	