Pricing Model Performance and the Two-Pass Cross-Sectional Regression Methodology

Raymond Kan, Cesare Robotti, and Jay Shanken

First draft: April 2008
This version: November 2010

*Kan is from the University of Toronto; Robotti is from the Federal Reserve Bank of Atlanta and EDHEC Risk Institute; Shanken is from Emory University and the National Bureau of Economic Research. We thank Pierluigi Balduzzi, Christopher Baum, Tarun Chordia, Wayne Ferson, Nikolay Gospodinov, Olesya Grishchenko, Ravi Jagannathan, Ralitsa Petkova, Monika Piazzesi (the Editor), Yaxuan Qi, Tim Simin, Jun Tu, Chu Zhang, Guofu Zhou, two anonymous referees, seminar participants at the Board of Governors of the Federal Reserve System, Concordia University, Federal Reserve Bank of Atlanta, Federal Reserve Bank of New York, Penn State University, University of Toronto, and participants at the 2009 Meetings of the Association of Private Enterprise Education, the 2009 CIREQ-CIRANO Financial Econometrics Conference, the 2009 FIRS Conference, the 2009 SoFiE Conference, the 2009 Western Finance Association Meetings, the 2009 China International Conference in Finance, and the 2009 Northern Finance Association Meetings for helpful discussions and comments. Kan gratefully acknowledges financial support from the Social Sciences and Humanities Research Council of Canada and the National Bank Financial of Canada. The views expressed here are the authors’ and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Corresponding author: Jay Shanken, Goizueta Business School, Emory University, 1300 Clifton Road, Atlanta, Georgia, 30322, USA; telephone: (404)727-4772; fax: (404)727-5238. E-mail: jay.shanken@bus.emory.edu.
Pricing Model Performance and the Two-Pass Cross-Sectional Regression Methodology

Abstract

Since Fama and MacBeth (1973), the two-pass cross-sectional regression (CSR) methodology has been the most popular approach for estimating and testing asset pricing models. Over the years, many studies have employed the CSR $R^2$ as a measure of model performance. We derive the asymptotic distribution of this statistic and develop associated model comparison tests, taking into account the inevitable impact of model misspecification on the variability of the CSR estimates. This provides a formal alternative to the common heuristic of simply comparing the point estimates of $R^2$. Indeed, we encounter several examples of large $R^2$ differences that are not statistically significant. A version of the intertemporal CAPM exhibits the best overall performance in our tests, followed by the more empirically motivated “three-factor model” of Fama and French (1993). Interestingly, the performance of several prominent consumption CAPMs proves to be sensitive to variations in design, such as the portfolios used as test assets or economic restrictions imposed on model parameters.
I. Introduction

A fundamental principle of microeconomics is the notion that if consumers allocate their income in an optimal fashion, then the relative prices of two goods should equal the ratio of their marginal utilities. This principle extends to financial assets, with the corresponding implication that the price of an asset should be lower (its expected return higher) if the asset tends to provide higher returns in low marginal utility states of the economy (see Rubinstein, 1976 and Breeden, 1979). These are states in which aggregate consumption is high, either because wealth is high or investment opportunities are such that less investment is desirable.

The classic capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965) considers the special case in which investment opportunities are constant and investors hold efficient portfolios so as to maximize their expected return (future wealth) for a given level of variance. The CAPM predicts that an asset’s risk premium will be proportional to its beta — the measure of return sensitivity to the aggregate market portfolio return. Merton (1973) and Long (1974) derive multi-beta expected return relations that reflect not only the market beta, but also an asset’s sensitivity to factors that relate to shifts in the investment opportunity set. As Breeden (1979) shows, Merton’s intertemporal model (ICAPM) is actually equivalent to a single-beta consumption model (CCAPM) since the chosen level of consumption endogenously reflects the various demand effects of the ICAPM through the simple marginal utility principle mentioned above.

The earliest tests of the CAPM averaged stock returns over time and then examined the cross-sectional relation between these averages and estimates of stock betas. Beta estimates were obtained from first-pass time-series regressions of stock returns on the returns of a broad market index. Regressing the cross-section of average returns on these betas (the “second pass”) then provides an estimate of the risk premium. Various problems in conducting inference with this “two-pass” time-series and cross-sectional approach were identified in the early years of the literature. The primary statistical methodology that emerged was due in large part to the classic Fama and MacBeth (1973) study, published in this journal.1 A key feature of the Fama-MacBeth approach, used in hundreds of papers since, is the estimation of a risk premium each month (via cross-sectional regressions), with inference ultimately based on the time-series mean and standard error of the monthly risk-premium estimates.

1Also see the related paper by Black, Jensen and Scholes (1972).
A formal econometric analysis of the two-pass methodology was first provided by Shanken (1992). He shows how the asymptotic standard error of the second-pass risk premium estimator is influenced by estimation error in the first-pass betas, requiring an adjustment to the traditional Fama-MacBeth standard errors.\footnote{Jagannathan and Wang (1998) extend this asymptotic analysis by relaxing the assumption that returns are homoskedastic conditional on the model’s factors. See Jagannathan, Skoulakis and Wang (2008) for a synthesis of the two-pass methodology.} A test of the validity of the pricing model’s constraint on expected returns can also be derived from the cross-sectional regression residuals, e.g., Shanken (1985).\footnote{Generalized method of moments (GMM) and maximum likelihood approaches for estimation and testing have also been developed. See Shanken and Zhou (2007) for detailed references to this literature and a discussion of relations between the different methodologies.} As a practical matter, however, models are at best approximations to reality. Therefore, it is desirable to have a measure of “goodness-of-fit” with which to assess the performance of a risk-return model. The most popular measure, given its simple intuitive appeal, has been the $R^2$ for the cross-sectional relation. This $R^2$ indicates the extent to which the model’s risk measures (betas) account for the cross-sectional variation in average returns, typically for a set of asset portfolios.

In this paper, our interest is in rigorously evaluating and comparing the performance of several prominent empirical asset pricing models. In addition to the basic CAPM and CCAPM, the theory-based models considered are the CAPM with labor income of Jagannathan and Wang (1996), the CCAPM conditioned on the consumption-wealth ratio of Lettau and Ludvigson (2001), the ultimate consumption risk model of Parker and Julliard (2005), the durable consumption model of Yogo (2006), and the five-factor ICAPM of Petkova (2006). We also study the well-known “three-factor model” of Fama and French (1993). Although this model was primarily motivated by empirical observation, its size and book-to-market factors are sometimes viewed as proxies for more fundamental economic factors.

Surprisingly, despite the widespread use of the two-pass methodology and the fact that $R^2$’s for competing models have routinely been compared in empirical asset pricing studies, to our knowledge no prior attempt has been made to derive the asymptotic distribution of this important statistic. Moreover, no formal model comparison test based on the $R^2$ measure has previously been proposed. Rather, the cross-sectional $R^2$ has been used mainly as a descriptive statistic. However, since this statistic is subject to considerable sampling variation, those comparisons can hardly be considered conclusive.
Against this background, the central methodological contribution of our paper is the introduction of statistical procedures for testing whether two beta-pricing models have the same population $R^2$, and whether a given model dominates others in the context of multiple model comparison. Our procedures account for the fact that each model’s parameters must be estimated and that these estimates will typically be correlated across models. Both ordinary least squares (OLS) and generalized least squares (GLS) $R^2$s are considered. OLS is more relevant if the focus is on the expected returns for a particular set of assets or test portfolios, but the GLS $R^2$ may be of greater interest from an investment perspective in that it is directly related to the relative efficiency of portfolios that “mimick” a model’s economic factors.\(^4\)

Model comparison essentially presumes that deviations from the implied restrictions are likely for some or all models, which seems particularly apt given the inherent limitations of asset pricing theory. Yet, researchers often conduct inferences about risk premia or other asset pricing model parameters while imposing the null hypothesis that the model is correctly specified. Indeed, it is not uncommon to see this done even when a model is clearly rejected by the data — a logical inconsistency. Therefore, the asymptotic properties of the two-pass methodology are derived here under quite general assumptions that allow for model misspecification, extending the results of Shanken and Zhou (2007) under normality.

Our main empirical analysis uses the “usual” 25 size and book-to-market portfolios of Fama and French (1993) plus five industry portfolios as the assets. Specification tests reject the hypothesis of a perfect fit for the majority of the models, further pointing to the need for robust statistical methods. We show empirically, for the first time, that misspecification-robust standard errors can be substantially higher than the usual ones when a factor is “non-traded,” i.e., is not some benchmark portfolio return. As one example, consider the $t$-statistic on the GLS risk premium estimator for the consumption growth factor in the durable consumption model of Yogo (2006). The Fama-MacBeth $t$-statistic declines from 2.50 to 2.20 with the usual adjustment for errors in the betas, but it is further reduced to only 1.36 when misspecification is taken into account.

Although there is still some evidence of pricing, significance is often substantially reduced for

---

\(^4\)Kandel and Stambaugh (1995) show that there is a direct relation between the GLS $R^2$ and the relative efficiency of a market index. They also argue, as do Roll and Ross (1994), that there is virtually no relation at all for the OLS $R^2$ unless the index is exactly efficient. Lewellen, Nagel and Shanken (2010) provide a multi-factor generalization of the GLS result, with “mimicking portfolios” substituted for factors that are not returns.
consumption and ICAPM factors. In the model comparison tests, the basic CAPM and CCAPM specifications are clearly the worst performers, with low CSR $R^2$s that are often statistically dominated by those of other models at the 5% level. The conditional CCAPM based on the consumption-wealth ratio is also a poor performer.

Across the various specifications considered, the ICAPM has the best overall fit, yet the three-factor model is found to statistically dominate other models far more frequently. This is due in part to the fact that the ICAPM $R^2$ is sometimes not very precisely estimated. Indeed, we see many cases in which large differences between $R^2$s are not reliably different from zero. For example, the ICAPM OLS fit exceeds that of the CAPM by a full 65 percentage points and is still not statistically significant. This highlights the difficulty of distinguishing between models and the limitations of simply comparing point estimates of $R^2$s. In this respect, our work reinforces and extends the simulation-based conclusion of Lewellen, Nagel and Shanken (2010), who focus on individual $R^2$s, rather than differences across models.\footnote{Jagannathan, Kubota and Takehara (1998), Kan and Zhang (1999), and Jagannathan and Wang (2007) use simulations to examine the sampling errors of the cross-sectional $R^2$ and risk premia estimates under the assumption that one of the factors is “useless,” i.e., independent of returns.}

We find that the durable goods version of CCAPM performs about as well as the top models in our basic analysis. Its relative performance deteriorates substantially, however, when we constrain the zero-beta rate (expected return for risky assets with no systematic risk) to equal the risk-free T-bill rate. Our exploration of this modification of the usual CSR approach is motivated by the observation that most of the estimated zero-beta rates are far too high to be consistent with plausible spreads between borrowing and lending rates, as required by theory.

Another issue concerns the fact that when a model is misspecified, its fit will generally vary with the test assets employed. Therefore, we examine the sensitivity of our model comparison results to an alternative set of assets — 25 portfolios formed by ranking stocks on size and CAPM beta. Interestingly, the conditional CCAPM and ICAPM are the best performers in this context, both dominating the three-factor model at the 5% level in the OLS case. Again, precision plays an important role here, as other models with lower $R^2$s than the three-factor model are not statistically dominated.

Finally, an important related question is whether a particular factor in a multi-factor model makes an incremental contribution to the model’s overall explanatory power, given the presence
of the other factors. We show that this question cannot be answered by examining the usual risk premium coefficients on the multiple regression betas, which have been the exclusive focus of most prior CSR analyses. Rather, one must determine whether a factor’s “covariance price of risk” differs from zero in the CSR model. Our empirical investigation of this issue results in a surprising finding for the three-factor model. With an unconstrained zero-beta rate, the much heralded book-to-market factor is not statistically significant in terms of covariance risk, but the size factor is.

The rest of the paper is organized as follows. Section II presents an asymptotic analysis of the zero-beta rate and risk premia estimates under potentially misspecified models. In addition, we provide an asymptotic analysis of the sample cross-sectional $R^2$s. Section III introduces tests of equality of cross-sectional $R^2$s for two competing models and provides the asymptotic distributions of the test statistics for different scenarios. Section IV presents our main empirical findings. Section V introduces a new test of multiple model comparison. We explore the small-sample properties of the various tests in Section VI and Section VII summarizes our main conclusions.

II. Asymptotic Analysis under Potentially Misspecified Models

A. Population Measures of Pricing Errors and Cross-Sectional $R^2$s

As discussed in the introduction, an asset pricing model seeks to explain cross-sectional differences in average asset returns in terms of asset betas computed relative to the model’s systematic economic factors. Thus, let $f$ be a $K$-vector of factors and $R$ a vector of returns on $N$ test assets. We define $Y = [f', R']'$ and its mean and covariance matrix as

$$\mu = E[Y] \equiv \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad (1)$$

$$V = \text{Var}[Y] \equiv \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}, \quad (2)$$

where $V$ is assumed to be positive definite.\footnote{For most of our analysis, we only need to assume $V_{11}$ is nonsingular and $V_{21}$ is of full column rank. For the case of GLS CSR, we also need to assume $V_{22}$ is nonsingular.} The multiple regression betas of the $N$ assets with respect to the $K$ factors are defined as $\beta = V_{21}V_{11}^{-1}$. These are measures of systematic risk or the
sensitivity of returns to the factors. In addition, we denote the covariance matrix of the residuals of the $N$ assets by $\Sigma = V_{22} - V_{21}V_{11}^{-1}V_{12}$.

The proposed $K$-factor beta pricing model specifies that asset expected returns are linear in the betas, i.e.,

$$\mu_2 = X\gamma,$$

where $X = [1_N, \beta]$ is assumed to be of full column rank, $1_N$ is an $N$-vector of ones, and $\gamma = [\gamma_0, \gamma_1]'$ is a vector consisting of the zero-beta rate ($\gamma_0$) and risk premia on the $K$ factors ($\gamma_1$).\(^7\) The zero-beta rate may be higher than the risk-free interest rate if risk-free borrowing rates exceed lending rates in the economy.

When the model is misspecified, the pricing-error vector, $\mu_2 - X\gamma$, will be nonzero for all values of $\gamma$. In that case, it makes sense to choose $\gamma$ to minimize some aggregation of pricing errors. Denoting by $W$ an $N \times N$ symmetric positive definite weighting matrix, we define the (pseudo) zero-beta rate and risk premia as the choice of $\gamma$ that minimizes the quadratic form of pricing errors:

$$\gamma_W \equiv \begin{bmatrix} \gamma_{W,0} \\ \gamma_{W,1} \end{bmatrix} = \text{argmin}_\gamma (\mu_2 - X\gamma)'W(\mu_2 - X\gamma) = (X'WX)^{-1}X'W\mu_2. \quad (4)$$

The corresponding pricing errors of the $N$ assets are then given by

$$e_W = \mu_2 - X\gamma_W = [I_N - X(X'WX)^{-1}X'W]\mu_2. \quad (5)$$

In addition to aggregating the pricing errors, researchers are often interested in a normalized goodness-of-fit measure for a model. A popular measure is the cross-sectional $R^2$. Following Kandel and Stambaugh (1995), this is defined as

$$\rho^2_W = 1 - \frac{Q}{Q_0}, \quad (6)$$

where

$$Q = e'_W We_W, \quad (7)$$

$$Q_0 = e'_0 W e_0, \quad (8)$$

and $e_0 = [I_N - 1_N(1'_N W 1_N)^{-1}1'_N W]\mu_2$ represents the deviations of mean returns from their cross-sectional average. In order for $\rho^2_W$ to be well defined, we need to assume that $\mu_2$ is not proportional

\(^7\)Portfolio characteristics that are constant over time can easily be accommodated in the CSR without creating any additional complication. The analysis that includes such characteristics is available upon request.
to $1_N$ (the expected returns are not all equal) so that $Q_0 > 0$. Note that $0 \leq \rho_W^2 \leq 1$ and it is a decreasing function of the aggregate pricing-error measure $Q = e'_W W e_W$. Thus, $\rho_W^2$ is a natural measure of goodness of fit.

While multiple regression betas or “factor loadings” are typically used as the regressors in the second-pass CSR, an alternative specification in terms of the covariances $V_{21}$ (equivalently, the simple regression betas) provides important information about a given factor’s contribution to the model’s explanatory power when $K > 1$. Thus, let $C = [1_N, V_{21}]$ and $\lambda_W$ be the choice of coefficients that minimizes the quadratic form of pricing errors:

$$\lambda_W \equiv \begin{bmatrix} \lambda_{W,0} \\ \lambda_{W,1} \end{bmatrix} = \underset{\lambda}{\text{argmin}} \lambda (\mu_2 - C \lambda)' W (\mu_2 - C \lambda) = (C' W C)^{-1} C' W \mu_2.$$  \hfill (9)

It is easy to show that the pricing errors from this alternative second-pass CSR, $e_W = \mu_2 - C \lambda_W$, are the same as those in (5) and thus that the $\rho_W^2$ for these two CSRs are also identical. However, as we will discuss in Section III.A, there are important differences in the economic interpretation of the pricing coefficients.\footnote{See Jagannathan and Wang (1998) and Cochrane (2005, Chapter 13.4) for a discussion of this issue. Another solution to this problem is to use simple regression betas as the regressors in the second-pass CSR, as in Chen, Roll, and Ross (1986) and Jagannathan and Wang (1996, 1998). Kan and Robotti (2009) provide asymptotic results for the CSR with simple regression betas under potentially misspecified models.}

It should be emphasized that unless the model is correctly specified, $\gamma_W$, $\lambda_W$, $e_W$, and $\rho_W^2$ depend on the choice of $W$. We consider two popular choices of $W$ in the literature, $W = I_N$ (OLS CSR) and $W = V_{22}^{-1}$ (GLS CSR). To simplify the notation, we suppress the subscript $W$ from $\gamma_W$, $\lambda_W$, $e_W$, and $\rho_W^2$ when the choice of $W$ is clear from the context.

B. Sample Measures of Pricing Errors and Cross-Sectional $R^2$’s

Let $Y_t = [f_t', R_t']'$, where $f_t$ is the vector of $K$ proposed factors at time $t$ and $R_t$ is a vector of returns on $N$ test assets at time $t$. Throughout the paper, we assume the time series $Y_t$ is jointly stationary and ergodic, with finite fourth moment. Suppose we have $T$ observations on $Y_t$ and denote the sample moments of $Y_t$ by

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix} = \frac{1}{T} \sum_{t=1}^{T} Y_t,$$

$$\hat{V} = \begin{bmatrix} \hat{V}_{11} & \hat{V}_{12} \\ \hat{V}_{21} & \hat{V}_{22} \end{bmatrix} = \frac{1}{T} \sum_{t=1}^{T} (Y_t - \hat{\mu})(Y_t - \hat{\mu})'.$$  \hfill (10)
The popular two-pass method first estimates the betas of the $N$ assets by running the following multivariate regression:

$$R_t = \alpha + \beta f_t + \epsilon_t, \quad t = 1, \ldots, T.$$  \hspace{1cm} (12)

The estimated betas from this first-pass time-series regression are given by the matrix $\hat{\beta} = \hat{V}_{21}\hat{V}_{11}^{-1}$. We then run a single CSR of $\hat{\mu}_2$ on $\hat{X} = [1_N, \hat{\beta}]$ to estimate $\gamma_W$ in the second pass.\(^9\)

When the weighting matrix $W$ is known, as in OLS CSR, we can estimate $\gamma_W$ in (4) by

$$\hat{\gamma} = (\hat{X}'W\hat{X})^{-1}\hat{X}'W\hat{\mu}_2.$$  \hspace{1cm} (13)

Similarly, letting $\hat{C} = [1_N, \hat{V}_{21}]$, we estimate $\lambda_W$ in (9) by

$$\hat{\lambda} = (\hat{C}'W\hat{C})^{-1}\hat{C}'W\hat{\mu}_2.$$  \hspace{1cm} (14)

In the GLS case, we need to substitute the consistent estimate of $W$, $\hat{W} = \hat{V}_{22}^{-1}$, in (13) and (14).

The sample measure of $\rho^2$ is similarly defined as

$$\rho^2 = 1 - \frac{\hat{Q}}{\hat{Q}_0},$$  \hspace{1cm} (15)

where $\hat{Q}_0$ and $\hat{Q}$ are obtained by substituting the sample counterparts of the parameters in (7) and (8).

C. Asymptotic Distribution of $\hat{\gamma}$ under Potentially Misspecified Models

When computing the standard error of $\hat{\gamma}$, researchers typically rely on the asymptotic distribution of $\hat{\gamma}$ under the assumption that the model is correctly specified. Shanken (1992) presents the asymptotic distribution of $\hat{\gamma}$ under the conditional homoskedasticity assumption on the residuals. Jagannathan and Wang (1998) extend Shanken’s results by allowing for conditional heteroskedasticity as well as autocorrelated errors.

Two recent papers have investigated the asymptotic distribution of $\hat{\gamma}$ under potentially misspecified models. Hou and Kimmel (2006) derive the asymptotic distribution of $\hat{\gamma}$ for the case of GLS CSR with a known value of $\gamma_0$, and Shanken and Zhou (2007) present asymptotic results for

\(^9\)Some studies allow $\hat{\beta}$ to change throughout the sample period. For example, in the original Fama and MacBeth (1973) study, the betas used in the CSR for month $t$ were estimated from data prior to that month. It has become more customary in recent decades to use full-period beta estimates for portfolios formed by ranking stocks on various characteristics.
the OLS, weighted least squares, and GLS cases with $\gamma_0$ unknown. However, both analyses are somewhat restrictive, as they rely on the i.i.d. normality assumption. We relax this assumption and provide general expressions for the asymptotic variances of both $\hat{\gamma}$ and $\hat{\lambda}$ under potential model misspecification in the appendix.\footnote{White (1994) and Hall and Inoue (2003) provide an asymptotic analysis of the parameter estimates for the case of misspecified GMM. However, their results cannot be directly applied to the two-pass CSR estimator since $\beta$ and $\gamma$ are not estimated jointly. Instead, the two-pass procedure can be interpreted as a sequential GMM that first estimates $\beta$ from one set of moment conditions and then estimates $\gamma$ using a different set of moment conditions by plugging in the estimated $\beta$ (see Pagan (1984) for an analysis of regressions with generated regressors and Newey (1984) for a discussion of sequential GMM).}

In this subsection, we examine a special case in which the variance formula simplifies considerably, permitting several important insights to emerge.

**Lemma 1.** When the factors and returns are i.i.d. multivariate elliptically distributed with kurtosis parameter $\kappa$, the asymptotic variance of the GLS estimator $\hat{\gamma} = (\hat{X}'\hat{V}_{22}^{-1}\hat{X})^{-1}\hat{X}'\hat{V}_{22}^{-1}\bar{\mu}_2$ is given by

$$V(\hat{\gamma}) = \Upsilon_w + \Upsilon_{w2},$$

where

$$\Upsilon_w = H + (1 + \kappa)\gamma_1'V_{11}^{-1}\gamma_1(X'S^{-1}X)^{-1},$$

$$\Upsilon_{w2} = (1 + \kappa)Q\left[(X'S^{-1}X)^{-1}\tilde{V}_{11}^{-1}(X'S^{-1}X)^{-1} + (X'S^{-1}X)^{-1}\right],$$

with $H = (X'V_{22}^{-1}X)^{-1}$, $Q = e'V_{22}^{-1}e$, and $\tilde{V}_{11}^{-1} = \begin{bmatrix} 0 & 0'K \\ 0K & V_{11}^{-1} \end{bmatrix}$.

The base term $H$ is the usual variance of a GLS estimator, ignoring measurement errors in the betas and model misspecification. This term is consistently estimated using the time-series sample variance of the monthly estimators, as in Fama-MacBeth. Each of the variance adjustment terms has an impact that is magnified when stock returns are fat-tailed.

The second term in $\Upsilon_w$ adjusts for the errors-in-variables (EIV) feature of the model, while the term $\Upsilon_{w2}$ adjusts for model misspecification. In the latter case, the variance increases with the degree of misspecification, as measured by $Q$. When $Q > 0$, $\Upsilon_{w2}$ is positive definite since it is the sum of two matrices, the first positive semidefinite and the second positive definite. In the proof of Lemma 1, we show that the misspecification adjustment term crucially depends on the
variance of the residuals from projecting the factors on the returns. For factors that have very low correlation with returns (e.g., macroeconomic factors), the impact of misspecification on the asymptotic variance of $\hat{\gamma}_1$ can be very large. This new insight will be helpful in understanding the empirical results in Section IV.

D. Asymptotic Distribution of the Sample Cross-Sectional $R^2$

The sample $R^2 (\hat{\rho}^2)$ in the second-pass CSR is a popular measure of goodness of fit for a model. A high $\hat{\rho}^2$ is viewed as evidence that the model under study does a good job of explaining the cross-section of expected returns. Lewellen, Nagel and Shanken (2010) point out several pitfalls in this approach and explore simulation techniques to obtain approximate confidence intervals for $\rho^2$. In this subsection, we provide the first formal statistical analysis of $\hat{\rho}^2$.

In the following proposition, we show that the asymptotic distribution of $\hat{\rho}^2$ crucially depends on whether (1) the population $\rho^2$ is 1 (i.e., a correctly specified model), (2) $0 < \rho^2 < 1$ (a misspecified model that provides some explanatory power for the expected returns on the test assets), or (3) $\rho^2 = 0$ (a misspecified model that does not explain any of the cross-sectional variation in expected returns for the test assets).

**Proposition 1.** In the following, we set $W$ to be $V_{22}^{-1}$ for the GLS case.

(1) When $\rho^2 = 1$, the limiting distribution of $T(\hat{\rho}^2 - 1)$ is that of a linear combination of $N - K - 1$ independent $\chi^2_1$ random variables.

(2) When $0 < \rho^2 < 1$,

$$\sqrt{T}(\hat{\rho}^2 - \rho^2) \overset{d}{\sim} N\left(0, \sum_{j=-\infty}^{\infty} E[n_t n_{t+j}]\right), \tag{19}$$

where

$$n_t = \frac{2[-u_t y_t + (1 - \rho^2)v_t]}{Q_0} \text{ for known } W, \tag{20}$$

$$n_t = \frac{[u_t^2 - 2u_t y_t + (1 - \rho^2)(2v_t - v_t^2)]}{Q_0} \text{ for } \hat{W} = \hat{V}_{22}^{-1}, \tag{21}$$

with $u_t = e'W(R_t - \mu_2)$, $v_t = e_0'W(R_t - \mu_2)$, and $y_t = 1 - \lambda_1' (f_t - \mu_1)$ is the normalized stochastic discount factor (SDF).
(3) When $\rho^2 = 0$, the limiting distribution of $T\hat{\rho}^2$ is that of a linear combination of $K$ independent $\chi_1^2$ random variables.

The first asymptotic distribution in the proposition allows us to perform a specification test of the beta pricing model. This is an alternative to the various multivariate asset pricing tests that have been developed in the literature. Although all of these tests focus on an aggregate pricing-error measure, the $R^2$-based test examines pricing errors in relation to the cross-sectional variation in expected returns, allowing for a simple and appealing interpretation. At the other extreme, the last asymptotic distribution permits a test of whether the model has any explanatory power for expected returns.

When $0 < \rho^2 < 1$, the case of primary interest, Proposition 1 shows that $\hat{\rho}^2$ is asymptotically normally distributed around its true value. It is readily verified that the expressions for $n_t$ approach zero when $\rho^2 \to 0$ or $\rho^2 \to 1$. Consequently, $se(\hat{\rho}^2)$ tends to be lowest when $\rho^2$ is close to zero or one, and thus it is not monotonic in $\rho^2$. Note that the asymptotic normal distribution of $\hat{\rho}^2$ breaks down for the two extreme cases ($\rho^2 = 0$ or 1). Intuitively, the normal distribution fails because, by construction, $\hat{\rho}^2$ will always be above zero (even when $\rho^2 = 0$) and below one (even when $\rho^2 = 1$).

Since most of the autocorrelations of the relevant terms ($n_t$) in this and our other propositions are small (under 0.1 and frequently under 0.05) and statistically insignificant, consistent with much of the literature, we conduct inference assuming these terms are serially uncorrelated. However, we have also explored the impact of a one-lag adjustment and find very little effect on our inferences about pricing and $R^2$s. These additional results are summarized briefly in the footnotes.

III. Tests for Comparing Competing Models

One way to think about model comparison and selection is to ask whether two competing beta pricing models have the same population cross-sectional $R^2$. In this section, we derive the asymptotic distribution of the difference between the sample $R^2$s of two models. We show that this distribution depends on whether the two models are nested or non-nested and whether the models are correctly specified or not.

---

11 This is because when $\rho^2 = 1$, we have $e = 0_N$ and $u_t = 0$, so $n_t$ in (20) and (21) becomes zero. Similarly, when $\rho^2 = 0$, we have $y_t = 1$, $e = e_0$ and $u_t = v_t$, so again $n_t$ in (20) and (21) vanishes.
Our analysis is related to the model selection tests of Kan and Robotti (2009) and Li, Xu and Zhang (2010) who, building on the earlier statistical work of Vuong (1989), Rivers and Vuong (2002), and Golden (2003), develop tests of equality of the Hansen and Jagannathan (1997) distances of two competing asset pricing models. As emphasized by Kan and Zhou (2004), however, the Hansen-Jagannathan distance evaluates a model’s ability to explain prices rather than expected returns and need not rank models in the same manner as $R^2$. Thus, an investigation of model comparison with the more frequently employed cross-sectional $R^2$ metric is needed.

We consider two competing beta pricing models. Let $f_1$, $f_2$, and $f_3$ be three sets of distinct factors, where $f_i$ is of dimension $K_i \times 1$, $i = 1, 2, 3$. Assume that model A uses $f_1$ and $f_2$, while Model B uses $f_1$ and $f_3$ as factors. Therefore, model A requires that the expected returns on the test assets are linear in the betas or covariances with respect to $f_1$ and $f_2$, i.e.,

$$
\mu_2 = 1_N \lambda_{A,0} + \text{Cov}[R, f'_1] \lambda_{A,1} + \text{Cov}[R, f'_2] \lambda_{A,2} = C_A \lambda_A,
$$

where $C_A = [1_N, \text{Cov}[R, f'_1], \text{Cov}[R, f'_2]]$ and $\lambda_A = [\lambda_{A,0}, \lambda_{A,1}', \lambda_{A,2}']'$. Model B requires that expected returns are linear in the betas or covariances with respect to $f_1$ and $f_3$, i.e.,

$$
\mu_2 = 1_N \lambda_{B,0} + \text{Cov}[R, f'_1] \lambda_{B,1} + \text{Cov}[R, f'_3] \lambda_{B,3} = C_B \lambda_B,
$$

where $C_B = [1_N, \text{Cov}[R, f'_1], \text{Cov}[R, f'_3]]$ and $\lambda_B = [\lambda_{B,0}, \lambda_{B,1}', \lambda_{B,3}']'$.

In general, both models can be misspecified. Following the development in Section II.A, given a weighting matrix $W$, the $\lambda_i$ that maximizes the $\rho^2$ of model $i$ is given by

$$
\lambda_i = (C_i' W C_i)^{-1} C_i' W \mu_2,
$$

where $C_i$ is assumed to have full column rank, $i = A, B$. For each model, the pricing-error vector $e_i$, the aggregate pricing-error measure $Q_i$, and the corresponding goodness-of-fit measure $\rho_i^2$ are all defined as in Section II.

When $K_2 = 0$, model B nests model A as a special case. Similarly, when $K_3 = 0$, model A nests model B. When both $K_2 > 0$ and $K_3 > 0$, the two models are non-nested. We study the nested models case in the next subsection and deal with non-nested models in Section III.B.
A. Nested Models

When models are nested, it is natural to suppose that the explanatory power of the larger model will exceed that of the smaller model precisely when expected returns are related to the betas on the additional factors. Our next result demonstrates that this is true, but only if we formulate this condition in terms of the simple betas or covariances with the factors. Without loss of generality, we assume $K_3 = 0$, so that model A nests model B.

Lemma 2. $\rho^2_A = \rho^2_B$ if and only if $\lambda_{A,2} = 0_{K_2}$.

Note that Lemma 2 is applicable even when the models are misspecified. By the lemma, to test whether the models have the same $\rho^2$, one can simply perform a test of $H_0 : \lambda_{A,2} = 0_{K_2}$ based on the CSR estimate and its misspecification-robust covariance matrix. Alternatively, in keeping with the common practice of comparing cross-sectional $R^2$s, we can use $\hat{\rho}^2_A - \hat{\rho}^2_B$ to test $H_0 : \rho^2_A = \rho^2_B$. In the appendix, we show that the asymptotic distribution of this statistic is that of a linear combination of $K_2$ independent $\chi^2_1$ random variables.

Before moving on to the case of non-nested models, we highlight an important issue about risk premia which does not appear to be widely understood. Empirical work on multi-factor asset pricing models typically focuses on whether factors are “priced” in the sense that coefficients on the multiple regression betas are nonzero in the CSR relation. While the economic interpretation of these risk premia can be of interest for other reasons, Lemma 2 tells us that if the question is whether the extra factors $f_2$ improve the cross-sectional $R^2$, then what matters is whether the prices of covariance risk associated with $f_2$ are nonzero.

Indeed, empirical papers frequently use factors that equal the spread between two investment returns, e.g., small-firm and large-firm returns. In this case, provided the model is correctly specified for all assets, the risk-premium coefficient on the multiple-regression beta must equal the expected value of the factor. This is true regardless of what other factors are present in the model. Thus, a nonzero coefficient, in this familiar context, indicates only that the expected return spread is nonzero. It says nothing about the incremental role of that factor in the cross-sectional model.

To establish this point more generally, we suppress the model subscript $A$ and let $\gamma = [\gamma_0, \gamma_1', \gamma_2']'$ be the zero-beta rate and risk premia for $f_1$ and $f_2$. The standard relation between multiple re-
gression betas and covariances, along with (4) and (9), then implies that there is a one-to-one correspondence between $\gamma$ and $\lambda$; the zero-beta rates are identical and the usual risk premia are obtained by multiplying the prices of covariance risk by the factor covariance matrix:

$$
\begin{bmatrix}
\gamma_1 \\
\gamma_2 \\
\end{bmatrix} = 
\begin{bmatrix}
\text{Var}[f_1] & \text{Cov}[f_1, f_2'] \\
\text{Cov}[f_2, f_1'] & \text{Var}[f_2] \\
\end{bmatrix}
\begin{bmatrix}
\lambda_1 \\
\lambda_2 \\
\end{bmatrix}.
$$

Hence, when $\lambda_2$ is zero, the risk premia associated with $f_2$ are $\gamma_2 = \text{Cov}[f_2, f_1']\lambda_1$. Clearly, $\gamma_2$ can still be nonzero unless $f_1$ and $f_2$ are uncorrelated. Similarly, we can show that $\gamma_2 = 0_{K_2}$ does not imply $\lambda_2 = 0_{K_2}$ unless $f_1$ and $f_2$ are uncorrelated.

B. Non-Nested Models

Testing $H_0: \rho_A^2 = \rho_B^2$ is more complicated for non-nested models. The reason is that under $H_0$, there are three possible asymptotic distributions for $\hat{\rho}_A^2 - \hat{\rho}_B^2$, depending on why the two models have the same cross-sectional $R^2$. To see this, we first define the normalized SDFs for models A and B as

$$
y_A = 1 - (f_1 - E[f_1])'\lambda_{A,1} - (f_2 - E[f_2])'\lambda_{A,2}, \quad y_B = 1 - (f_1 - E[f_1])'\lambda_{B,1} - (f_3 - E[f_3])'\lambda_{B,3}.
$$

At first sight, it may appear that $y_A = y_B$ is equivalent to the joint restriction $\lambda_{A,1} = \lambda_{B,1}$, $\lambda_{A,2} = 0_{K_2}$ and $\lambda_{B,3} = 0_{K_3}$. The following lemma shows that the first equality is redundant, however, since it is implied by the other two.

**Lemma 3.** For non-nested models, $y_A = y_B$ if and only if $\lambda_{A,2} = 0_{K_2}$ and $\lambda_{B,3} = 0_{K_3}$.

In the appendix, we show that $y_A = y_B$ implies that the two models have the same pricing errors and hence $\rho_A^2 = \rho_B^2$. If $y_A \neq y_B$, there are additional cases in which $\rho_A^2 = \rho_B^2$. A second possibility is that both models are correctly specified (i.e., $\rho_A^2 = \rho_B^2 = 1$). This occurs, for example, if model A is correctly specified and the factors $f_3$ in model B are given by $f_3 = f_2 + \epsilon$, where $\epsilon$ is pure “noise” — a vector of measurement errors with mean zero, independent of returns. In this case, we have $C_A = C_B$ and both models produce zero pricing errors. A third possibility is that the two models produce different pricing errors but the same overall goodness of fit. Intuitively,

---

12When $\lambda_2 = 0_{K_2}$, we see that $\gamma_1 = \text{Var}[f_1]\lambda_1$. Consequently, the risk premia for $f_1$ stay the same when we add $f_2$ to the model.

13Some numerical illustrations of these points are provided in the appendix.
one model might do a good job of pricing some assets that the other prices poorly and vice versa, such that the aggregation of pricing errors is the same in each case \( \rho_A^2 = \rho_B^2 < 1 \). As it turns out, each of these three scenarios results in a different asymptotic distribution for \( \hat{\rho}_A^2 - \hat{\rho}_B^2 \).

Given the three distinct cases described above, testing \( H_0 : \rho_A^2 = \rho_B^2 \) for non-nested models entails a sequential test, as suggested by Vuong (1989). In our context, this involves first testing \( H_0 : y_A = y_B \). If we reject \( H_0 : y_A = y_B \), then we go on to test \( H_0 : \rho_A^2 = \rho_B^2 = 1 \). This second test can be viewed as a generalization of the cross-sectional regression test (CSRT) of Shanken (1985) and later multivariate tests of the validity of the expected return relation for a single pricing model. Finally, if the hypothesis that both models are correctly specified is also rejected, we proceed to evaluate \( H_0 : 0 < \rho_A^2 = \rho_B^2 < 1 \) using a normal test that will be discussed shortly. Let \( \alpha_1, \alpha_2, \) and \( \alpha_3 \) be the significance levels employed in these three tests. Then the sequential test has an asymptotic significance level that is bounded above by \( \max[\alpha_1, \alpha_2, \alpha_3] \). Thus, if \( \alpha_1 = \alpha_2 = \alpha_3 = 0.05 \), the significance level of this procedure for testing \( H_0 : \rho_A^2 = \rho_B^2 \) is asymptotically no larger than 5%.

Another approach is to simply perform the final test of \( H_0 : 0 < \rho_A^2 = \rho_B^2 < 1 \). This amounts to assuming that \( y_A \neq y_B \) and that both models are misspecified. By Lemma 3, the first assumption rules out the unlikely scenario that the additional factors are completely irrelevant for explaining cross-sectional variation in expected returns. The second assumption is sensible because asset pricing models are approximations of reality and we do not expect them to be perfectly specified. In our empirical work, we will conduct both the sequential test and the normal test when comparing non-nested models. We focus mainly on the normal test, however, as this test will be more powerful insofar as the assumptions above are valid.

The tests of \( y_A = y_B \) and of \( \rho_A^2 = \rho_B^2 = 1 \) are developed in the appendix. Next, we present the details of the normal test.

1. Testing \( H_0 : 0 < \rho_A^2 = \rho_B^2 < 1 \)

**Proposition 2.** Suppose \( y_A \neq y_B \) and \( 0 < \rho_A^2 = \rho_B^2 < 1 \).\(^{14}\) We have:

\[
\sqrt{T}(\hat{\rho}_A^2 - \hat{\rho}_B^2) \sim N \left( 0, \sum_{j=-\infty}^{\infty} E[d_d dt_j] \right). \tag{27}
\]

\(^{14}\)Since \( \rho_A^2 = \rho_B^2 = 0 \) implies \( y_A = y_B = 1 \), this case is already covered by the test based on Lemma 3.
When the weighting matrix $W$ is known,

$$d_t = 2Q_0^{-1} \left[ u_{Bt} y_{Bt} - u_{At} y_{At} - (\rho_A^2 - \rho_B^2)v_t \right],$$

(28)

where $u_{At} = e'_A W (R_t - \mu_2)$, $u_{Bt} = e'_B W (R_t - \mu_2)$, and $v_t = e'_0 W (R_t - \mu_2)$. With the GLS weighting matrix $\hat{W} = \hat{V}_2^{-1}$,

$$d_t = Q_0^{-1} \left[ u_{At}^2 - 2u_{At} y_{At} - u_{Bt}^2 + 2u_{Bt} y_{Bt} - (\rho_A^2 - \rho_B^2)(2v_t - v_t^2) \right],$$

(29)

where $u_{At} = e'_A V_2^{-1} (R_t - \mu_2)$ and $u_{Bt} = e'_B V_2^{-1} (R_t - \mu_2)$.$^{15}$

Note that if $y_{At} = y_{Bt}$, then $\rho_A^2 = \rho_B^2$, $u_{At} = u_{Bt}$, and hence $d_t = 0$. Or, if $y_{At} \neq y_{Bt}$, but both models are correctly specified (i.e., $u_{At} = u_{Bt} = 0$ and $\rho_A^2 = \rho_B^2 = 1$), then again $d_t = 0$. Thus, the normal test cannot be used in these cases, consistent with the maintained assumptions in the proposition.

IV. Empirical Analysis

We use our methodology to evaluate the performance of several prominent asset pricing models. First, we describe the data used in the empirical analysis and outline the different specifications of the beta pricing models considered. Then we present our results. Simulation results supporting the use of our tests are deferred to a later section so that we can get right to the empirical analysis.

A. Data and Beta Pricing Models

The return data are from Kenneth French’s website and consist of the monthly returns on the 25 Fama-French size and book-to-market ranked portfolios plus five industry portfolios. The industry portfolios are included to provide a more challenging test of the models, as suggested by Lewellen, Nagel and Shanken (2010). The data are from February 1959 to July 2007 (582 monthly observations). The beginning date of our sample period is dictated by the consumption data availability.

$^{15}$One could impose $H_0 : \rho_A^2 = \rho_B^2$ in (28) and (29) and the $v_t$ terms would drop out of these expressions. However, our simulation results indicate that not imposing $H_0 : \rho_A^2 = \rho_B^2$ in the computation of the standard errors leads to improved finite-sample properties of the normal test. Similarly, we obtain better finite-sample performance when, in the GLS case, we multiply $u_t$ and $v_t$ by $(T - N - 2)/T$. 
We analyze eight asset pricing models starting with the simple static CAPM. The cross-sectional specification for this model is

\[ \mu_2 = \gamma_0 + \beta_{vw} \gamma_{vw}, \]

where \( vw \) is the excess return (in excess of the one-month T-bill rate from Ibbotson Associates) on the value-weighted stock market index (NYSE-AMEX-NASDAQ) from Kenneth French’s website. The CAPM performed well in the early tests, e.g., Fama and MacBeth (1973), but has fared poorly since.

One extension that has performed better is our second model, the conditional CAPM (C-LAB) of Jagannathan and Wang (1996). This model incorporates measures of the return on human capital as well as the change in financial wealth and allows the conditional betas to vary with a state variable, \( prem \), the lagged yield spread between BAA and AAA rated corporate bonds from the Board of Governors of the Federal Reserve System.\(^\text{16}\) The cross-sectional specification is

\[ \mu_2 = \gamma_0 + \beta_{vw} \gamma_{vw} + \beta_{lab} \gamma_{lab} + \beta_{prem} \gamma_{prem}, \]

where \( lab \) is the growth rate in per capita labor income, \( L \), defined as the difference between total personal income and dividend payments, divided by the total population (from the Bureau of Economic Analysis). Following Jagannathan and Wang (1996), we use a two-month moving average to construct the growth rate \( lab_t = (L_{t-1} + L_{t-2})/(L_{t-2} + L_{t-3}) - 1 \), for the purpose of minimizing the influence of measurement error.

Our third model (FF3) extends the CAPM by including two empirically-motivated factors. This is the Fama-French (1993) three-factor model with

\[ \mu_2 = \gamma_0 + \beta_{vw} \gamma_{vw} + \beta_{smb} \gamma_{smb} + \beta_{hml} \gamma_{hml}, \]

where \( smb \) is the return difference between portfolios of small and large stocks, where size is based on market capitalization, and \( hml \) is the return difference between portfolios of stocks with high and low book-to-market ratios (“value” and “growth” stocks, respectively) from Kenneth French’s website.

The fourth model (ICAPM) is an empirical implementation of Merton’s (1973) intertemporal extension of the CAPM based on Campbell (1996), who argues that innovations in state variables

\(^{16}\)All bond yield data are from this source unless noted otherwise.
that forecast future investment opportunities should serve as the factors. The five-factor specification proposed by Petkova (2006) is

\[ \mu_2 = \gamma_0 + \beta_{vw}\gamma_{vw} + \beta_{\text{term}}\gamma_{\text{term}} + \beta_{\text{def}}\gamma_{\text{def}} + \beta_{\text{div}}\gamma_{\text{div}} + \beta_r\gamma_{rf}, \]

where \( \text{term} \) is the difference between the yields of ten-year and one-year government bonds, \( \text{def} \) is the difference between the yields of long-term corporate Baa bonds and long-term government bonds (from Ibbotson Associates), \( \text{div} \) is the dividend yield on the Center for Research in Security Prices (CRSP) value-weighted stock market portfolio, and \( rf \) is the one-month T-bill yield (from CRSP, Fama Risk Free Rates). The actual factors for \( \text{term}, \text{def}, \text{div}, \) and \( rf \) are their innovations from a VAR(1) system of seven state variables that also includes \( vw, smb, \) and \( hml \).\(^{17}\)

Next, we consider consumption-based models. Our fifth model (CCAPM) is the unconditional consumption model, with

\[ \mu_2 = \gamma_0 + \beta_{cg}\gamma_{cg}, \]

where \( cg \) is the growth rate in real per capita nondurable consumption (seasonally adjusted at annual rates) from the Bureau of Economic Analysis. This model has generally not performed well empirically. Therefore, we also examine other consumption models that have yielded more encouraging results.

One such model (CC-CAY) is a conditional version of the CCAPM due to Lettau and Ludvigson (2001). The relation is

\[ \mu_2 = \gamma_0 + \beta_{cay}\gamma_{cay} + \beta_{cg}\gamma_{cg} + \beta_{cg\cdot cay}\gamma_{cg\cdot cay}, \]

where \( cay \), the conditioning variable, is a consumption-aggregate wealth ratio.\(^ {18}\) This specification is obtained by scaling the constant term and the \( cg \) factor of a linearized consumption CAPM by a constant and \( cay \). Scaling factors by instruments is one popular way of allowing factor risk premia and betas to vary over time. See Cochrane (1996), among others.

Our seventh model (U-CCAPM) is the ultimate consumption model of Parker and Julliard (2005), which measures asset systematic risk as the covariance with future, as well as contemporaneous consumption, allowing for slow adjustment of consumption to the information driving

\(^{17}\)In contrast to Petkova (2006), we do not orthogonalize the innovations since the \( R^2 \) of the model is the same whether we orthogonalize or not.

\(^{18}\)Following Jørgensen and Attanasio (2003), we linearly interpolate the quarterly values of \( cay \) to permit analysis at the monthly frequency.
returns. The specification is
\[ \mu_2 = \gamma_0 + \beta_{cg36}\gamma_{cg36}, \]
where \( cg36 \) is the growth rate in real per capita nondurable consumption over three years starting with the given month.

The last model (D-CCAPM), due to Yogo (2006), highlights the cyclical role of durable consumption in asset pricing. The specification is
\[ \mu_2 = \gamma_0 + \beta_{vw}\gamma_{vw} + \beta_{cg}\gamma_{cg} + \beta_{cgdur}\gamma_{cgdur}, \]
where \( cgdur \) is the growth rate in real per capita durable consumption (seasonally adjusted at annual rates) from the Bureau of Economic Analysis.

B. Results

We start by estimating the sample cross-sectional \( R^2 \)s of the various pricing models just described. Then we analyze pricing and the impact of potential model misspecification on the statistical properties of the estimated \( \gamma \) and \( \lambda \) parameters. Next, we present the results of our pairwise tests of equality of the cross-sectional \( R^2 \)s for different models. Finally, we examine the sensitivity of our findings to requiring that the zero-beta rate equal the risk-free rate.

1. Sample Cross-Sectional \( R^2 \)s of the Models

In Table 1, we report \( \hat{\rho}^2 \) for each model and investigate whether the model does a good job of explaining the cross-section of expected returns. We denote the \( p \)-value of a specification test of \( H_0 : \rho^2 = 1 \) by \( p(\rho^2 = 1) \), and the \( p \)-value of a test of \( H_0 : \rho^2 = 0 \) by \( p(\rho^2 = 0) \). Both tests are based on the asymptotic results in Proposition 1 for the sample cross-sectional \( R^2 \) statistic. We also provide an approximate \( F \)-test of model specification for comparison. Next, we report the asymptotic standard error for \( R^2 \), \( \text{se}(\hat{\rho}^2) \), computed under the assumption that \( 0 < \rho^2 < 1 \). Finally, the number of parameters in each asset pricing model is \text{No. of par}.

The \( F \)-test is a generalized version of the CSRT of Shanken (1985). It is based on a quadratic form in the model’s deviations, \( \hat{Q}_c = \hat{\epsilon}'\hat{V}(\hat{\epsilon})^+\hat{\epsilon} \), where \( \hat{V}(\hat{\epsilon}) \) is a consistent estimator of the asymptotic variance of the sample pricing errors and \( \hat{V}(\hat{\epsilon})^+ \) its pseudo-inverse. When the model is
correctly specified (i.e., \( e = 0_N \) or \( \rho^2 = 1 \)), we have \( T\hat{Q}_c \overset{d}{\sim} \chi^2_{N-K-1} \). Following Shanken, the reported \( p \)-value, \( p(\hat{Q}_c = 0) \), is for a transformation of \( \hat{Q}_c \) that has an approximate \( F \) distribution: \( \hat{Q}_c \overset{\text{app}}{\sim} (N-K-1) F_{N-K-1,T-N+1} \).  

Following Shanken, the reported \( p \)-value, \( p(\hat{Q}_c = 0) \), is for a transformation of \( \hat{Q}_c \) that has an approximate \( F \) distribution: \( \hat{Q}_c \overset{\text{app}}{\sim} (N-K-1) F_{N-K-1,T-N+1} \).

Table 1 about here

In Panels A and B of Table 1, we provide results for the OLS and GLS CSRs, respectively. First, we consider the specification tests. The OLS \( F \)-test rejects five of the eight models at the 1% level, with four of those five also rejected by the \( R^2 \) test. Using GLS, all models are rejected at the 5% level and all but one at the 1% level. For OLS, D-CCAPM has the highest \( R^2 \) of 77.2%, with ICAPM and FF3 close behind. The same three models have the highest GLS \( R^2 \)s, with FF3 the highest at 29.8%. Turning to the test of \( \rho^2 = 0 \), we see that this null hypothesis is rejected at the 5% level for five of the eight models using OLS and for just three models with GLS.

Note that FF3, with an OLS \( R^2 \) of 74.7%, is rejected at the 1% level by both tests, whereas C-LAB, with a lower \( R^2 \) of 54.8%, is not rejected at the 10% level by the \( R^2 \) test. This is understandable when we observe that the FF3 OLS \( R^2 \) has the lowest standard error of all the models. Thus, a strong rejection by the specification test may be driven by relatively small deviations from a model if those deviations are precisely estimated. As a result, the specification test is not useful for model comparison. An alternative test will be needed to determine whether a model like FF3 significantly outperforms other models.

Another issue is the number of factors in a model. While the ICAPM has five factors, the other models considered have at most three factors. The extra degrees of freedom will be an advantage for the ICAPM in any given sample, holding true explanatory power constant across models. Our formal test will take this sampling variation into account and enable us to infer whether the model is superior in population, i.e., whether it better explains true expected returns.

Assuming that \( 0 < \rho^2 < 1 \), \( se(\hat{\rho}^2) \) captures the sampling variability of \( \hat{\rho}^2 \). In Table 1, we observe that the \( \hat{\rho}^2 \)s of several models are quite volatile. In particular, the ICAPM GLS \( R^2 \) is not

\(^{19}\)Our \( \hat{Q}_c \) is more general than the CSRT of Shanken (1985) because we can use sample pricing errors from any CSR, not just the ones from the GLS CSR. In addition, we allow for conditional heteroskedasticity and autocorrelated errors. Proofs of the results related to \( \hat{Q}_c \) are available upon request.

\(^{20}\)Simulation evidence suggests that this test has better size properties than the asymptotic test, especially when \( N \) is large relative to \( T \).
significantly different from zero; despite being the second highest of eight $R^2$s, its standard error is the largest. Four of the eight OLS standard errors exceed 0.2, with U-CCAPM’s the highest at 0.244. This high volatility will make it hard to distinguish between models.\footnote{With a one-lag Newey-West adjustment, the $p$-values under 0.10 for the $R^2$-based specification test barely change (OLS and GLS). For the $F$-test, the only noteworthy change is a decline in the OLS D-CCAPM $p$-value from 0.077 to 0.037. $P$-values for the OLS and GLS tests of $\rho^2 = 0$ hardly change. Finally, most of the standard errors of $\hat{\rho}^2$ barely change. The largest change across all specifications is an increase for U-CCAPM from 0.244 to 0.269 (OLS).}

Several observations emerge from the results in Table 1. First, there is strong evidence of the need to incorporate model misspecification into our statistical analysis. Second, there is considerable sampling variability in $\hat{\rho}^2$ and so it is not entirely clear whether one model truly outperforms the others. Finally, specification-test results are sometimes sensitive to whether we employ OLS or GLS estimation, and it is not always the case that models with high $\hat{\rho}^2$s pass the specification test.

2. Properties of the $\gamma$ and $\lambda$ Estimates under Correctly Specified and Potentially Misspecified Models

Before turning to model comparison, we examine the pricing results based on the $\gamma$ and $\lambda$ estimators. As far as we know, all previous CSR studies except the recent paper by Shanken and Zhou (2007) have used standard errors that assume the model is correctly specified. As we argued in the introduction, it is difficult to justify this practice because (as we just saw empirically) some, if not all, of the models are bound to be misspecified. In this subsection, we investigate whether inferences about pricing are affected by using an asymptotic standard error that is robust to such model misspecification.

In Table 2, we focus on the zero-beta rate and risk premia estimates, $\hat{\gamma}$, of the beta pricing models. For each model, we report $\hat{\gamma}$ and associated $t$-ratios under correctly specified and potentially misspecified models. For correctly specified models, we give the $t$-ratio of Fama and MacBeth (1973), followed by that of Shanken (1992) and Jagannathan and Wang (1998), which account for estimation error in the betas. Last, is the $t$-ratio under a potentially misspecified model, based on our new results provided in the appendix. The various $t$-ratios are identified by subscripts $fm$, $s$, $jw$, and $pm$, respectively.
We see evidence in panel A (OLS) that the ultimate consumption factor $c_{g36}$, the value-growth factor $h_{ml}$ and the prem state variable have coefficients that are reliably positive at the 5% level. In panel B (GLS), $h_{ml}$ is again positively priced. Consistent with many past studies, the market factor $vw$ is negatively priced in several specifications, contrary to the usual theoretical prediction. In addition, the zero-beta rates exceed the risk-free rate by large amounts that would seem hard to reconcile with theory. We return to these issues later on.

Consistent with our theoretical results, we find that the $t$-ratios under correctly specified and potentially misspecified models are similar for traded factors, e.g., the FF3 factors, but they can differ substantially for factors that have low correlations with asset returns. As an example of the latter, consider the consumption factor $c_{g}$ in D-CCAPM. With GLS estimation, we have $t$-ratio$_{fm} = 2.50$, $t$-ratio$_{s} = 2.20$, $t$-ratio$_{jw} = 2.14$, and $t$-ratio$_{pm} = 1.36$, which shows that the misspecification adjustment can make a significant difference. The ICAPM provides another illustration of the different conclusions that one can reach by using misspecification-robust standard errors. While the $t$-ratios under correctly specified models in Panel B suggest that $\hat{\gamma}_{term}$ is highly statistically significant ($t$-ratio$_{fm} = 3.07$, $t$-ratio$_{s} = 2.58$, and $t$-ratio$_{jw} = 2.59$), the robust $t$-ratio is only 1.61, not quite significant at the 10% level. The scale factor $cay$ is one more example. In short, both model misspecification and beta estimation error materially affect inference about the expected return relation.

As discussed in Section III.A, there are issues with testing whether an individual factor risk premium is zero or not in a multi-factor model. Unless the factors are uncorrelated, only the price of covariance risk (elements of $\lambda_1$) allows us to identify factors that improve the explanatory power of the expected return model (equivalently, simple regression betas can be used). The usual risk premium for a given factor does not permit such an inference. Table 3 presents estimation results for $\lambda$. To conserve space, only the OLS $t$-ratios under potentially misspecified models are presented.

As an illustration of our point that risk premia and prices of covariance risk can deliver different messages, consider FF3. The size-factor coefficient $\hat{\lambda}_{smb}$ is statistically significant at the 1% level,

$^{22}$The market premium is positive in the CAPM. In the ICAPM, it is positive after controlling for the market’s exposure to the hedging factors. See, for example, Fama (1996).
with a robust $t$-ratio of 2.79. In contrast, $\hat{\gamma}_{smb}$ in Table 2 has a robust $t$-ratio of only 1.19. The reverse occurs for the value-growth factor, with $\hat{\lambda}_{hml}$ not quite significant at the 10% level, yet $\hat{\gamma}_{hml}$ commanding a $t$-ratio of 3.42 earlier. Hence, by focusing on the usual risk premia (the $\gamma$s), one might think that $smb$ is not an important factor in FF3 and that $hml$ is. However, results for the prices of covariance risk (the $\lambda$s) imply that $smb$ has explanatory power for the cross-section of expected returns above and beyond the other factors in FF3, while the role of $hml$ is questionable.\footnote{As expected, for one-factor models, $\hat{\gamma}_1$ and $\hat{\lambda}_1$ result in similar inferences. In this case, the $t$-ratios of the $\hat{\gamma}_1$ and $\hat{\lambda}_1$ would be identical if we imposed the null hypotheses of $\gamma_1 = 0$ and $\lambda_1 = 0$, so that the EIV adjustment terms drop out of the analysis.}

To summarize, accounting for model misspecification often makes a qualitative difference in determining whether estimates of the risk premia or the prices of covariance risk are statistically significant, especially when the factor has low correlation with asset returns. This is the case for several of the models (ICAPM, D-CCAPM, CC-CAY) that include macroeconomic or scaled factors. In addition, we have seen that focusing on the $\hat{\gamma}$s, rather than $\hat{\lambda}$s, as is typical in the literature, can lead to erroneous conclusions as to whether a factor is helpful in explaining the cross-section of expected returns.\footnote{Most of the changes in $t$-statistics with a one-lag Newey-West adjustment are trivial, with the largest across all specifications being a drop from 2.34 to 2.10 for the OLS estimator $\hat{\gamma}_{cg}$.}

3. Tests of Equality of the Cross-Sectional $R^2$s of Two Competing Models

Recall that a $p$-value is the probability, under the null hypothesis, of obtaining a test statistic at least as extreme as the one observed. As such, the $p$-value provides no direct information about alternative hypotheses and the extent of deviations from the null. Therefore, $p$-values from the specification tests do not allow us to formally compare models. In this subsection, we explore relative goodness of fit by empirically testing whether competing beta pricing models have significantly different sample cross-sectional $R^2$s.

In Section III, we noted that the asymptotic distribution of the difference between the sample cross-sectional $R^2$s depends on whether the two competing models are correctly specified or not and whether they are nested or non-nested. For nested models, we use the weighted chi-squared test described below Lemma 2.\footnote{Results for the Wald test of $\lambda_{A,2} = 0_{K_2}$ (not reported in the paper) are consistent with those shown in Table 4.} For the reasons discussed in Section III.B, we focus on the normal test for non-nested models in Proposition 2, and comment briefly on the sequential test results.
In Table 4, we report pairwise tests of equality of cross-sectional $R^2$s for different models, some nested and others non-nested. Panel A is for the OLS CSR and Panel B is for the GLS CSR. Each panel shows the differences between the sample cross-sectional $R^2$s for various pairs of models and the associated $p$-values (in parentheses). In the case of non-nested models, the reported $p$-values are two-tailed $p$-values.

The main findings can be summarized as follows. First, the results show that the CAPM and CCAPM are often outperformed by other models at the 1% and 5% levels. Specifically, CCAPM is dominated at the 5% level by U-CCAPM, FF3, D-CCAPM and ICAPM in Panel A, and again by the last two models in Panel B. CAPM is dominated by C-LAB, FF3 and D-CCAPM in Panel A, and by FF3 in Panel B. In many cases the OLS $R^2$ differences with CAPM exceed 60 or 70 percentage points. In addition, FF3 dominates the consumption models CC-CAY (OLS and GLS) and U-CCAPM (GLS).\footnote{As noted earlier, all the $p$-values in Table 4 are computed under the assumption that the $d_t$ are serially uncorrelated. For $p$-values less than 0.10, the largest change observed with a one-lag Newey-West adjustment is an increase from 0.032 to 0.049 for the CCAPM/U-CCAPM comparison. Most other changes are trivial.}

There are several cases of large $R^2$ differences that do not give rise to statistical rejections due to limited precision of the estimates. Recall, for example, that U-CCAPM has the highest OLS standard error (0.244) in Table 1. Despite an $R^2$ difference of 27 percentage points in favor of FF3, the $p$-value in Panel A of Table 4 is 0.226. As another example, the ICAPM fit improves on the CAPM by a full 65 percentage points and still just misses being statistically significant at the 5% level. Clearly, the common practice of simply comparing sample $R^2$ values is not a reliable method for identifying superior models.\footnote{In one of nine cases, the sequential test no longer rejects at the 5% level.}

We have also explored the effect of including the three Fama-French factors, along with the 30 portfolios, as test assets in the various model comparisons. For models with one or more of these traded factors, inclusion requires that the estimated price of risk conform to the corresponding model restriction (i.e., equal the expected market premium over the zero-beta rate or equal the expected spread return for $smb$ and $hml$). As discussed by Lewellen, Nagel and Shanken (2010), this holds either exactly (GLS) or approximately (OLS). The changes in results here are minimal,
perhaps because the factors are closely mimicked by the original test assets. For comparison with other studies, we also performed the analysis using just the 25 size and book-to-market portfolios. The range in $R^2$s was slightly higher in this case and the standard error of $\hat{\rho}^2$ was higher for every model. Consistent with the lower precision, there were 11 instances of model comparison rejections at the 5% level, as compared to 15 in Table 4.

Inherent in model misspecification is the fact that one model may exhibit superior performance with some test assets, but poorer performance with other assets. To explore this possibility, we repeated the analysis with 25 portfolios formed by first sorting stocks into quintiles based on size and then, within each size quintile, by the estimate of the simple $vw$ beta. In contrast to the earlier results, the performance of CC-CAY is impressive with these portfolios. Its OLS $R^2$ is 0.874 (0.366 earlier), about the same as that for ICAPM, and CC-CAY actually dominates FF3 at the 5% level. With GLS estimation, its $R^2$ of 0.432 is the highest of all the models and CC-CAY is the only model not rejected by the specification tests. Thus, we see that model comparison can be very sensitive to the test assets employed.

4. Excess Returns Analysis

Consistent with standard practice in the literature, the CSR analysis thus far has proceeded with the zero-beta rate and risk premia coefficients unconstrained. The resulting $R^2$ is a reasonable measure of a model’s success in explaining cross-sectional differences in average returns. However, given the high values of the zero-beta rate and the negative market premia, we may not want to “credit” the theories for all of this explanatory power. One way of dealing with this issue is to constrain the zero-beta rate to equal the risk-free rate, a practice that is common in other parts of the empirical asset pricing literature. For example, studies that focus on time-series “alphas” when all factors are traded impose this restriction (see, for example, Gibbons, Ross and Shanken, 1989).

We implement the zero-beta restriction in the CSR context by working with test portfolio returns in excess of the T-bill rate, while excluding the constant from the expected return relations. As is typical for regression analysis without a constant, the corresponding $R^2$ measure involves (weighted) sums of squared values of the dependent variable (mean excess returns) in the denominator, not

---

28 All NYSE-AMEX-NASDAQ common stocks are considered. This is similar to the approach of Fama and French (1992). We use quintiles, rather than deciles, to mitigate potential finite-sample issues related to the inversion of a large sample covariance matrix. The results of this analysis are available upon request.
squared deviations from the cross-sectional average. This ensures that $R^2$ is always between 0 and 1.

Table 5 presents $R^2$s and other model information for the excess returns specification, in the same format as Table 1. The first thing that we notice are the large OLS values of $R^2$. These numbers are not directly comparable to the earlier values, however, for the reason just mentioned. Since the models do not include a constant, simply getting the overall level of mean returns right will now enhance a model’s $R^2$. In contrast, a positive value of the earlier goodness-of-fit measure indicates that a model has some ability to explain deviations of mean test-portfolio returns from the cross-sectional average.

Table 5 about here

The OLS sample $R^2$ measures in Panel A range from 0.858 (CAPM) to 0.972 (ICAPM), whereas the GLS values in Panel B are much lower, ranging from 0.044 (CCAPM) to 0.339 (ICAPM). Moreover, ICAPM is the only model that is not rejected at the 5% level by either OLS specification test. For brevity, we do not report the detailed pricing results, but note that the OLS $\hat{\gamma}$s for $cg$ (in CCAPM), $cg36$, $vw$, $hml$, and $term$ are significantly positive, while those for $rf$ and $div$ are significantly negative at the 5% level. The GLS results are similar, except that $cg$ is no longer significant.

The D-CCAPM factor $cgdur$ is no longer significantly priced using excess returns, and the model, which was the top OLS performer earlier, now has one of the lowest cross-sectional $R^2$ values. On the other hand, U-CCAPM moves up in the OLS rankings, now just behind ICAPM and FF3. Also noteworthy is the fact that $hml$ is the dominant factor in this FF3 specification, i.e., it now has a significant price of covariance risk, while $smb$ does not.

To conserve space, we simply summarize the model comparison results. There are fewer rejections now. FF3 dominates CAPM (OLS and GLS), CCAPM and D-CCAPM (both GLS), all at the 1% level. There are no additional rejections at the 5% level, although FF3 barely misses over C-LAB (GLS) and ICAPM comes close to dominating CAPM (OLS and GLS). It is interesting to note that while FF3 dominates more models statistically, ICAPM has the higher sample $R^2$s. Precision appears to play a role in this. The standard error of ICAPM’s GLS $R^2$ is the highest of all the models. Again, the importance of taking into account information about sampling variation
Earlier, we noted that the performance of CC-CAY was impressive with size-beta portfolios employed as the test assets. The model even dominated FF3 at the 5% level (OLS). This is no longer true in the excess-returns specification with size-beta portfolios. In this case, ICAPM is again the top OLS performer, followed by FF3. However, CAPM is the only model dominated at the 5% level. With GLS estimation, the CC-CAY $R^2$ of 0.5 is about twice that of the nearest competitor, but due to its large standard error (0.196), the model does not dominate FF3 or ICAPM, even at the 10% level.

V. Multiple Model Comparison

Thus far, we have considered comparison of two competing models. However, given a set of models of interest, one may want to test whether one model, the “benchmark,” has the highest $\rho^2$ of all models in the set. This gives rise to a common problem in applied work — if we focus on the statistic that provides the strongest evidence of rejection, without taking into account the process of searching across alternative specifications, there will be a tendency to reject the benchmark more often than the nominal size of the tests suggests. In other words, the true $p$-value will be larger than the one associated with the most extreme statistic. For example, in a head-on competition, we saw earlier that FF3 dominates C-LAB at the 5% level with GLS estimation. But will C-LAB still be statistically rejected from the perspective of multiple model comparison?

In this section, we develop and implement a formal test of multiple model comparison. This is a multivariate inequality test based on results in the statistics literature due to Wolak (1987, 1989). Suppose we have $p$ models. Let $\rho^2_i$ denote the population CSR $R^2$ of model $i$ and let $\delta \equiv (\delta_2, \ldots, \delta_p)$, where $\delta_i \equiv \rho^2_1 - \rho^2_i$: We are interested in testing the null hypothesis that the benchmark, model 1, performs at least as well as all others, i.e., $H_0: \delta \geq 0$, with $r = p - 1$. The alternative is that some model has a higher population $R^2$ than model 1.

The test is based on the sample counterpart, $\hat{\delta} \equiv (\hat{\delta}_2, \ldots, \hat{\delta}_p)$, where $\hat{\delta}_i \equiv \hat{\rho}^2_1 - \hat{\rho}^2_i$. We assume

---

29 The sequential test delivers the same conclusions as the normal test at the 5% level.

30 Chen and Ludvigson (2009) employ the “reality check” of White (2000) to draw inferences about multiple model comparison with the Hansen-Jagannathan distance.
that
\[ \sqrt{T}(\hat{\delta} - \delta) \overset{d}{\sim} N(0, r, \Sigma_{\hat{\delta}}). \] (30)

As in Proposition 2, sufficient conditions for asymptotic normality are i) \( 0 < \rho_i^2 < 1 \), and ii) the implied SDFs of the different models are distinct. Let \( \hat{\delta} \) be the optimal solution in the following quadratic programming problem:

\[ \min_{\delta} (\hat{\delta} - \delta)' \Sigma_{\hat{\delta}}^{-1} (\hat{\delta} - \delta) \quad \text{s.t.} \quad \delta \geq 0, \] (31)

where \( \hat{\Sigma}_{\hat{\delta}} \) is a consistent estimator of \( \Sigma_{\hat{\delta}} \). The likelihood ratio test of the null hypothesis is

\[ LR = T(\hat{\delta} - \tilde{\delta})' \Sigma_{\hat{\delta}}^{-1} (\hat{\delta} - \tilde{\delta}). \] (32)

Since the null hypothesis is composite, to construct a test with the desired size, we require the distribution of \( LR \) under the least favorable value of \( \delta \), which is \( \delta = 0, r \). Under this value, \( LR \) follows a “chi-bar distribution,” i.e.,

\[ LR \overset{d}{\sim} \sum_{i=0}^{r} w_i (\Sigma_{\hat{\delta}}^{-1}) X_i, \] (33)

where the \( X_i \)'s are independent \( \chi^2 \) random variables with \( i \) degrees of freedom and \( \chi^2_0 \) is simply defined as the constant zero. An explicit formula for the weights \( w_i \), which are functions of \( \Sigma_{\hat{\delta}} \), is given in Kudo (1963). In the appendix, we provide a numerically efficient procedure for obtaining these weights that greatly improves on methods employed in previous research. We use this procedure to obtain asymptotically valid \( p \)-values.

In comparing a benchmark model with a set of alternative models, we first remove those alternative models \( i \) that are nested by the benchmark model since, by construction, \( \delta_i \geq 0 \) in this case. If any of the remaining alternatives is nested by another alternative model, we remove the “smaller” model since the \( \rho^2 \) of the “larger” model will be at least as big. Finally, we also remove from consideration any alternative models that nest the benchmark, since the normality assumption on \( \hat{\delta}_i \) does not hold under the null hypothesis that \( \delta_i = 0 \). An alternative testing procedure for multiple model comparison is needed in this case. We return to this issue below. Table 6 provides our findings.
For the OLS comparisons in Panel A, only CCAPM is rejected, with \( p \)-value 0.000. The GLS results in Panel B provide additional evidence against the consumption models. CCAPM and CC-CAY are rejected at the 5% level and U-CCAPM just misses rejection with \( p \)-value 0.053. C-LAB which, as mentioned above, was dominated in the pairwise comparison with FF3 (\( p \)-value 0.025), is no longer rejected in the multiple model comparison. The GLS \( p \)-value of 0.073 in Table 6 is higher than before, since it takes into account the element of searching over alternative models.

Next, we turn to the topic of nested multiple model comparison. Although the LR test is no longer applicable here (\( \hat{\delta} \) is not asymptotically normally distributed), fortunately our earlier approach to testing for equality of \( R^2 \)s can easily be adapted to this context. One need only consider a single expanded model that includes all of the factors contained in the models that nest the benchmark. For example, in the case of CCAPM, this expanded model includes \( cg \), \( cay \), \( cay \cdot cg \), \( vw \), and \( cgdur \) from CC-CAY and D-CCAPM. Using Lemma 2, it is easily demonstrated that the expanded model dominates the benchmark model if and only if one or more of the “larger” models dominate it. Thus, the null hypothesis that the benchmark model has the same \( R^2 \) as these alternatives can be tested using the earlier methodology.\(^{31}\)

The results are as follows. The CCAPM \( p \)-values are 0.009 (OLS) and 0.092 (GLS), while the \( p \)-values for CAPM are 0.057 (OLS) and 0.458 (GLS). Given the \( p \)-value of 0.001 for the CAPM/FF3 comparison in Table 4 (OLS and GLS), a Bonferroni approach would have yielded a stronger rejection of CAPM in this case. With four models nesting CAPM, the Bonferroni (upper-bound) \( p \)-value is \( 4 \times 0.001 = 0.004 \). But of course, the decision about which joint test to perform should really be made a priori.

We conclude with a few observations about results for our alternative empirical specifications. Consistent with our earlier evidence that the performance of D-CCAPM declines in the excess-returns specification, the model is rejected at the 5% level in multiple comparison tests with excess returns, as is CCAPM (GLS). Also, multiple model comparison confirms the decline of FF3 when size-beta portfolios are employed. FF3 is rejected at the 5% level in this case, as are CAPM and CCAPM (OLS). Since several models with lower \( R^2 \)s are not rejected, again the greater precision

\(^{31}\)If one of the models has a higher \( R^2 \), then so will the expanded model. Conversely, if none of the models improves the \( R^2 \), then for each of the additional factors, the vector of asset covariances must be orthogonal to the CSR residuals of the benchmark model, as will any linear combination of these covariance vectors. Thus, the expanded model must have the same \( R^2 \) as the benchmark.
with which the FF3 $R^2$ is estimated contributes to this finding. With excess returns and size-beta portfolios, however, only CAPM (OLS and GLS) and CCAPM (GLS) are dominated at the 5% level.

VI. Simulation Evidence

In this section, we explore the small-sample properties of our various test statistics via Monte Carlo simulations. In all simulation experiments, the test assets are the 25 size and book-to-market portfolios plus five industry portfolios used in most of our analysis. The time-series sample size is taken to be $T = 600$, close to the actual sample size of 582 in our empirical work. The factors and the returns on the test assets are drawn from a multivariate normal distribution. Both OLS and GLS specifications are examined. We compare actual rejection rates over 10,000 iterations to the nominal 5% level of our tests. A more detailed description of the various simulation designs can be found in the appendix.

We start with the specification tests — the $R^2$ test based on Proposition 1 and the approximate $F$-test. To evaluate the size properties of these tests, we simulate data from a world in which FF3 is exactly true. The $F$-test performs very well in both cases, with just a slight tendency to over-reject (5.5% OLS, 5.6% GLS). The $R^2$ test is right on the money for OLS, but rejects a bit too much (7.8%) for GLS. To analyze the power of these tests, we simulate data assuming the population $R^2$s for FF3 are 0.747 (OLS) and 0.298 (GLS), the sample values observed earlier. The rejection rates are close to one in all cases when $T = 600$.\(^{32}\)

Both of our tests of the hypothesis $\rho^2 = 0$ have the correct size when simulating a world in which FF3 has no explanatory power. The tests also display good power against alternatives based on the FF3 sample $R^2$s. Likewise for the nested-models test of equality of $R^2$s, with CAPM nested in FF3.

Next, we turn to our main test, the normal test of equality of $R^2$s for non-nested models. In this experiment, $\rho^2$ is the same for FF3 and C-LAB, 0.647 (OLS) or 0.203 (GLS). These are averages of the sample $R^2$s obtained earlier. The size properties of the test are very good in the OLS case (5.8%), while there is a tendency to under-reject a bit (2.7%) with GLS (this tendency is more

\(^{32}\)When the nominal size of the test differs from the actual size, this should be interpreted as power corresponding to the latter.
pronounced for lower values of $T$).

The power of the normal test is explored using the sample $R^2$s of FF3 and C-LAB as the population $R^2$ values. These are 0.747 and 0.548 for OLS, 0.298 and 0.109 for GLS. The rejection rate is only 14.6% for OLS, but somewhat higher at 31.5% in the GLS case. This is a reflection of the limited precision of the sample $R^2$s, given the substantial noise (unexpected) component of returns. We also examined power using CCAPM and FF3 as the two models, with the CCAPM $\rho^2$ equal to 0.044 (OLS) or 0.011 (GLS). Naturally, power increases substantially, given these large differences in performance. The rejection rates are now 87% (OLS) and 76% (GLS).

Finally, we examine the multiple-comparison inequality test for non-nested models. Recall that the composite null hypothesis for this test maintains that $\rho^2$ for the benchmark model is at least as high as that for all other models under consideration. Therefore, to evaluate size, we consider the case in which all models have the same $\rho^2$ value, so as to maximize the likelihood of rejection under the null. We simulate six different single-factor models corresponding to the factors $vw$, $smb$, $cg36$, $lab$, $prem$, and $rf$ and implement the likelihood ratio test with $r = 5$. The rejection rates range from 3.3% to 5% (OLS) and from 2.7% to 6% (GLS). Thus, the tests are fairly well specified under the null.

To examine power, we simulate five of our original models, CCAPM, U-CCAPM, C-LAB, FF3, and ICAPM, with the earlier sample $R^2$s serving as the population $R^2$s. Since FF3 and ICAPM have the highest $R^2$s, power is assessed by examining rejections of the three other alternative models in the multiple model comparison test. The rejection rates for the OLS test are 13.8% (C-LAB), 35.9% (U-CCAPM), and 86.3% (CCAPM). The corresponding GLS numbers are 25.1%, 64.9%, and 72%, respectively. Naturally, power increases as the $\rho^2$ of the benchmark model decreases, and “good” power requires that the differences in model performance are fairly large.

Overall, these simulation results suggest that the tests should be fairly reliable for the sample size encountered in our empirical work.

VII. Conclusion

We have provided an analysis of the asymptotic statistical properties of the traditional cross-sectional regression methodology and the associated $R^2$ goodness-of-fit measure when an underlying
beta pricing model fails to hold exactly. The importance of adjusting standard errors for model misspecification has also been demonstrated empirically for several prominent asset pricing models. As far as we know, our study is the first to consider (analytically or even in simulations) the important sampling distribution of the difference between the sample $R^2$s of two competing models. As we show, the asymptotic distribution of this difference depends on whether the models are correctly specified and whether they are nested or non-nested.

Our main analysis employs the 25 Fama-French size and book-to-market portfolios plus five industry portfolios as the test assets. In this case, the ICAPM specification of Petkova (2006) is the best overall performer, with the three-factor model of Fama and French (FF3, 1993) right behind. The $R^2$ differences for comparing these two models are not reliably different from zero, however, whether OLS or GLS estimation is employed, and whether the zero-beta rate is constrained to equal the risk-free rate or not. The durable goods consumption model of Yogo (2006) is competitive with FF3 and ICAPM in the main analysis, but its performance declines dramatically when we impose economic restrictions on the zero-beta rates.

With an alternative set of test assets, 25 size-beta portfolios, some important changes emerge. The conditional CAPM of Lettau and Ludvigson (2001), one of the poorer performers in the main analysis, now competes with ICAPM for top honors in several specifications. On the other hand, FF3 exhibits some vulnerability with these test assets, and is dominated at the 5% level by both models.

The evidence discussed above involves pairwise model comparison. While these methods take us well beyond the common practice of simply comparing point estimates of $R^2$, those tests are open to a criticism common in applied work, i.e., that the process of searching over various models to identify interesting results can lead to an overstatement of statistical significance. We address this issue by introducing tests of multiple model comparison for nested and non-nested models. Naturally, this results in fewer rejections than were obtained earlier. Several of our main conclusions are reinforced, however: rejections of the basic CAPM and CCAPM, near-rejections (at the 5% level) of both the conditional and ultimate CCAPMs, rejection of the durable CCAPM in the excess-returns specification, and rejection of FF3 when the test assets are size-beta portfolios.

To sum up, the robust performance of the ICAPM and the fact that it is the only model that is never statistically dominated in any of our analyses is impressive. Nevertheless, we should keep
in mind that the ICAPM $R^2$ is sometimes not estimated very precisely. Also, the model achieves its superior explanatory power with five factors, two more than any of the competing models. Still, while this undoubtedly can be an advantage, additional risk measures certainly don’t have to be related to actual expected returns, and our statistical analysis does take into account sampling variation related to the larger degrees of freedom.

Looking to the future, although our simulation results are encouraging, the small-sample properties of the test statistics proposed in this paper should be explored further. Other metrics for comparing models besides the CSR $R^2$ could also be considered. Finally, incorporating theoretical restrictions on the risk premia in the measure of model performance might be a way of enhancing power and providing more informative model comparison tests.
References


### TABLE 1
Sample Cross-Sectional $R^2$'s and Specification Tests of the Models

#### Panel A: OLS

<table>
<thead>
<tr>
<th></th>
<th>CAPM</th>
<th>C-LAB</th>
<th>FF3</th>
<th>ICAPM</th>
<th>CCAPM</th>
<th>CC-CAY</th>
<th>U-CCAPM</th>
<th>D-CCAPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\rho}^2$</td>
<td>0.115</td>
<td>0.548</td>
<td>0.747</td>
<td>0.766</td>
<td>0.044</td>
<td>0.366</td>
<td>0.473</td>
<td>0.772</td>
</tr>
<tr>
<td>$p(\rho^2 = 1)$</td>
<td>0.000</td>
<td>0.051</td>
<td>0.002</td>
<td>0.327</td>
<td>0.000</td>
<td>0.001</td>
<td>0.116</td>
<td>0.237</td>
</tr>
<tr>
<td>$p(\rho^2 = 0)$</td>
<td>0.258</td>
<td>0.042</td>
<td>0.009</td>
<td>0.009</td>
<td>0.510</td>
<td>0.256</td>
<td>0.005</td>
<td>0.007</td>
</tr>
<tr>
<td>se($\hat{\rho}^2$)</td>
<td>0.200</td>
<td>0.221</td>
<td>0.117</td>
<td>0.145</td>
<td>0.130</td>
<td>0.211</td>
<td>0.244</td>
<td>0.125</td>
</tr>
<tr>
<td>$\hat{Q}_c$</td>
<td>0.131</td>
<td>0.060</td>
<td>0.098</td>
<td>0.058</td>
<td>0.137</td>
<td>0.102</td>
<td>0.100</td>
<td>0.067</td>
</tr>
<tr>
<td>$p(Q_c = 0)$</td>
<td>0.000</td>
<td>0.170</td>
<td>0.001</td>
<td>0.135</td>
<td>0.000</td>
<td>0.001</td>
<td>0.002</td>
<td>0.077</td>
</tr>
<tr>
<td>No. of par.</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

#### Panel B: GLS

<table>
<thead>
<tr>
<th></th>
<th>CAPM</th>
<th>C-LAB</th>
<th>FF3</th>
<th>ICAPM</th>
<th>CCAPM</th>
<th>CC-CAY</th>
<th>U-CCAPM</th>
<th>D-CCAPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\rho}^2$</td>
<td>0.107</td>
<td>0.109</td>
<td>0.298</td>
<td>0.242</td>
<td>0.011</td>
<td>0.015</td>
<td>0.034</td>
<td>0.239</td>
</tr>
<tr>
<td>$p(\rho^2 = 1)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.024</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.007</td>
</tr>
<tr>
<td>$p(\rho^2 = 0)$</td>
<td>0.005</td>
<td>0.337</td>
<td>0.000</td>
<td>0.076</td>
<td>0.547</td>
<td>0.933</td>
<td>0.280</td>
<td>0.016</td>
</tr>
<tr>
<td>se($\hat{\rho}^2$)</td>
<td>0.069</td>
<td>0.071</td>
<td>0.101</td>
<td>0.137</td>
<td>0.036</td>
<td>0.040</td>
<td>0.059</td>
<td>0.133</td>
</tr>
<tr>
<td>$\hat{Q}_c$</td>
<td>0.126</td>
<td>0.128</td>
<td>0.099</td>
<td>0.086</td>
<td>0.143</td>
<td>0.141</td>
<td>0.149</td>
<td>0.084</td>
</tr>
<tr>
<td>$p(Q_c = 0)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.004</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.010</td>
</tr>
<tr>
<td>No. of par.</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

**Note.**–The table presents the sample cross-sectional $R^2$ ($\hat{\rho}^2$) and the generalized CSRT ($\hat{Q}_c$) of eight beta pricing models. The models include the CAPM, the conditional CAPM (C-LAB) of Jagannathan and Wang (1996), the Fama and French (1993) three-factor model (FF3), the intertemporal CAPM (ICAPM) of Petkova (2006), the consumption CAPM (CCAPM), the conditional consumption CAPM (CC-CAY) of Lettau and Ludvigson (2001), the ultimate consumption CAPM (U-CCAPM) of Parker and Julliard (2005), and the durable consumption CAPM (D-CCAPM) of Yogo (2006). The models are estimated using monthly returns on the 25 Fama-French size and book-to-market ranked portfolios and five industry portfolios. The data are from February 1959 to July 2007 (582 observations). $p(\rho^2 = 1)$ is the $p$-value for the test of $H_0 : \rho^2 = 1$. $p(\rho^2 = 0)$ is the $p$-value for the test of $H_0 : \rho^2 = 0$. se($\hat{\rho}^2$) is the standard error of $\hat{\rho}^2$ under the assumption that $0 < \rho^2 < 1$. $p(Q_c = 0)$ is the $p$-value for the approximate $F$-test of $H_0 : Q_c = 0$. No. of par. is the number of parameters in the model.
### TABLE 2
Estimates and t-ratios of Zero-Beta Rate and Risk Premia under Correctly Specified and Misspecified Models

#### Panel A: OLS

<table>
<thead>
<tr>
<th></th>
<th>CAPM</th>
<th>C-LAB</th>
<th>FF3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\gamma}_0$</td>
<td>$\hat{\gamma}_{vw}$</td>
<td>$\hat{\gamma}_0$</td>
</tr>
<tr>
<td>Estimate</td>
<td>1.61</td>
<td>-0.46</td>
<td>1.77</td>
</tr>
<tr>
<td>$t$-ratio$_{fm}$</td>
<td>4.81</td>
<td>-1.19</td>
<td>5.56</td>
</tr>
<tr>
<td>$t$-ratio$_{s}$</td>
<td>4.79</td>
<td>-1.18</td>
<td>3.52</td>
</tr>
<tr>
<td>$t$-ratio$_{lw}$</td>
<td>4.69</td>
<td>-1.17</td>
<td>3.72</td>
</tr>
<tr>
<td>$t$-ratio$_{pm}$</td>
<td>4.32</td>
<td>-1.11</td>
<td>3.71</td>
</tr>
</tbody>
</table>

#### ICAPM

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\gamma}_0$</th>
<th>$\hat{\gamma}_{vw}$</th>
<th>$\hat{\gamma}_{term}$</th>
<th>$\hat{\gamma}_{def}$</th>
<th>$\hat{\gamma}_{div}$</th>
<th>$\hat{\gamma}_{rf}$</th>
<th>$\hat{\gamma}_0$</th>
<th>$\hat{\gamma}_{cg}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>1.14</td>
<td>-0.15</td>
<td>0.20</td>
<td>-0.14</td>
<td>-0.02</td>
<td>-0.44</td>
<td>0.96</td>
<td>0.18</td>
</tr>
<tr>
<td>$t$-ratio$_{fm}$</td>
<td>4.29</td>
<td>-0.47</td>
<td>2.50</td>
<td>-2.69</td>
<td>-1.32</td>
<td>-3.13</td>
<td>4.80</td>
<td>0.75</td>
</tr>
<tr>
<td>$t$-ratio$_{s}$</td>
<td>2.76</td>
<td>-0.33</td>
<td>1.62</td>
<td>-1.75</td>
<td>-0.89</td>
<td>-2.03</td>
<td>4.68</td>
<td>0.73</td>
</tr>
<tr>
<td>$t$-ratio$_{lw}$</td>
<td>2.89</td>
<td>-0.35</td>
<td>1.56</td>
<td>-1.55</td>
<td>-0.91</td>
<td>-1.84</td>
<td>4.72</td>
<td>0.76</td>
</tr>
<tr>
<td>$t$-ratio$_{pm}$</td>
<td>2.56</td>
<td>-0.32</td>
<td>1.38</td>
<td>-1.50</td>
<td>-0.85</td>
<td>-1.85</td>
<td>3.99</td>
<td>0.65</td>
</tr>
</tbody>
</table>

#### CCAPM

<table>
<thead>
<tr>
<th></th>
<th>CC-CAY</th>
<th>U-CCAPM</th>
<th>D-CCAPM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\gamma}_0$</td>
<td>$\hat{\gamma}_{cay}$</td>
<td>$\hat{\gamma}_{cg}$</td>
</tr>
<tr>
<td>Estimate</td>
<td>1.46</td>
<td>-1.46</td>
<td>-0.02</td>
</tr>
<tr>
<td>$t$-ratio$_{fm}$</td>
<td>5.48</td>
<td>-2.42</td>
<td>-0.13</td>
</tr>
<tr>
<td>$t$-ratio$_{s}$</td>
<td>3.78</td>
<td>-1.67</td>
<td>-0.09</td>
</tr>
<tr>
<td>$t$-ratio$_{lw}$</td>
<td>4.69</td>
<td>-2.07</td>
<td>-0.10</td>
</tr>
<tr>
<td>$t$-ratio$_{pm}$</td>
<td>4.12</td>
<td>-1.21</td>
<td>-0.06</td>
</tr>
</tbody>
</table>
TABLE 2 (Continued)
ESTIMATES AND t-RATIOS OF ZERO-BETA RATE AND RISK PREMIA UNDER
CORRECTLY SPECIFIED AND MISSPECIFIED MODELS

<table>
<thead>
<tr>
<th>Panel B: GLS</th>
<th>CAPM</th>
<th>C-LAB</th>
<th>FF3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\gamma}_0$</td>
<td>$\hat{\gamma}_{vw}$</td>
<td>$\hat{\gamma}_0$</td>
</tr>
<tr>
<td>Estimate</td>
<td>1.75</td>
<td>−0.79</td>
<td>1.74</td>
</tr>
<tr>
<td>$t$-ratio$_{fm}$</td>
<td>9.40</td>
<td>−3.07</td>
<td>9.05</td>
</tr>
<tr>
<td>$t$-ratio$_s$</td>
<td>9.25</td>
<td>−3.04</td>
<td>8.88</td>
</tr>
<tr>
<td>$t$-ratio$_{jw}$</td>
<td>9.22</td>
<td>−3.03</td>
<td>8.85</td>
</tr>
<tr>
<td>$t$-ratio$_{pm}$</td>
<td>8.20</td>
<td>−2.83</td>
<td>7.53</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel C: ICAPM</th>
<th>CCAPM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CAPM</td>
</tr>
<tr>
<td>$t$-ratio$_{fm}$</td>
<td>6.75</td>
</tr>
<tr>
<td>$t$-ratio$_s$</td>
<td>5.63</td>
</tr>
<tr>
<td>$t$-ratio$_{jw}$</td>
<td>5.57</td>
</tr>
<tr>
<td>$t$-ratio$_{pm}$</td>
<td>4.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel D: CC-CAY</th>
<th>U-CCAPM</th>
<th>D-CCAPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$-ratio$_{fm}$</td>
<td>8.97</td>
<td>0.04</td>
</tr>
<tr>
<td>$t$-ratio$_s$</td>
<td>8.85</td>
<td>0.04</td>
</tr>
<tr>
<td>$t$-ratio$_{jw}$</td>
<td>8.67</td>
<td>0.04</td>
</tr>
<tr>
<td>$t$-ratio$_{pm}$</td>
<td>7.06</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Note.—The table presents the estimation results of eight beta pricing models. The models include the CAPM, the conditional CAPM (C-LAB) of Jagannathan and Wang (1996), the Fama and French (1993) three-factor model (FF3), the intertemporal CAPM (ICAPM) of Petkova (2006), the consumption CAPM (CCAPM), the conditional consumption CAPM (CC-CAY) of Lettau and Ludvigson (2001), the ultimate consumption CAPM (U-CCAPM) of Parker and Julliard (2005), and the durable consumption CAPM (D-CCAPM) of Yogo (2006). The models are estimated using monthly returns on the 25 Fama-French size and book-to-market ranked portfolios and five industry portfolios. The data are from February 1950 to July 2007 (582 observations). We report parameter estimates $\hat{\gamma}$ (multiplied by 100), the Fama and MacBeth (1973) $t$-ratio under correctly specified models ($t$-ratio$_{fm}$), the Shanken (1992) and the Jagannathan and Wang (1998) $t$-ratios under correctly specified models that account for the EIV problem ($t$-ratio$_s$ and $t$-ratio$_{jw}$, respectively), and our model misspecification-robust $t$-ratios ($t$-ratio$_{pm}$).
TABLE 3
Estimates and t-ratios of Zero-Beta Rate and Prices of Covariance Risk under Correctly Specified and Misspecified Models (OLS Case)

<table>
<thead>
<tr>
<th></th>
<th>CAPM</th>
<th>C-LAB</th>
<th>FF3</th>
<th>ICAPM</th>
<th>CCAPM</th>
</tr>
</thead>
</table>
|               | \(\hat{\lambda}_0\) | \(\hat{\lambda}_{vw}\) | \(\hat{\lambda}_0\) | \(\hat{\lambda}_{vw}\) | \(\hat{\lambda}_{lab}\) | \(\hat{\lambda}_{prem}\) | \(\hat{\lambda}_0\) | \(\hat{\lambda}_{vw}\) | \(\hat{\lambda}_{smb}\) | \(\hat{\lambda}_{hml}\) | \(\hat{\lambda}_0\) | \(\hat{\lambda}_{cay}\)
| Estimate      | 1.61 | -2.45 | 1.77     | -6.17 | 120.82 | 260.84 | 1.94     | -5.25 | 4.63 | 3.33 | 1.14 | -18.06 |
| \(t\)-ratio\(_{pm}\) | 4.32 | -1.12 | 3.71     | -2.09 | 0.82    | 2.75    | 6.71     | -2.25 | 2.79 | 1.60 | 2.56 | -1.89 |

<table>
<thead>
<tr>
<th></th>
<th>ICAPM</th>
<th>CCAPM</th>
</tr>
</thead>
</table>
|               | \(\hat{\lambda}_0\) | \(\hat{\lambda}_{term}\) | \(\hat{\lambda}_{def}\) | \(\hat{\lambda}_{div}\) | \(\hat{\lambda}_r\) | \(\hat{\lambda}_0\) | \(\hat{\lambda}_{cg}\)
| Estimate      | 1.14 | 147.13 | -325.08 | -605.05 | -108.40 | 0.96 | 29.19 |
| \(t\)-ratio\(_{pm}\) | 2.56 | -1.89 | 0.62     | -1.87  | -1.63   | -1.52 | 3.99   | 0.65 |

<table>
<thead>
<tr>
<th></th>
<th>CC-CAY</th>
<th>U-CCAPM</th>
<th>D-CCAPM</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>
|               | \(\hat{\lambda}_0\) | \(\hat{\lambda}_{cay}\) | \(\hat{\lambda}_{cg}\) | \(\hat{\lambda}_{cg-cay}\) | \(\hat{\lambda}_0\) | \(\hat{\lambda}_{cg36}\) | \(\hat{\lambda}_0\) | \(\hat{\lambda}_{vw}\) | \(\hat{\lambda}_{cg}\) | \(\hat{\lambda}_{cgdur}\)
| Estimate      | 1.46   | -65.77  | 0.75    | 4222.68   | 0.68       | 31.26     | 2.20       | -7.82     | 62.72     | 20.14 |
| \(t\)-ratio\(_{pm}\) | 4.12   | -1.43   | 0.01    | 0.43      | 2.78       | 2.30      | 6.39       | -2.56     | 0.94      | 1.57 |

Note.—The table presents the estimation results of eight beta pricing models. The models include the CAPM, the conditional CAPM (C-LAB) of Jagannathan and Wang (1996), the Fama and French (1993) three-factor model (FF3), the intertemporal CAPM (ICAPM) of Petkova (2006), the consumption CAPM (CCAPM), the conditional consumption CAPM (CC-CAY) of Lettau and Ludvigson (2001), the ultimate consumption CAPM (U-CCAPM) of Parker and Julliard (2005), and the durable consumption CAPM (D-CCAPM) of Yogo (2006). The models are estimated using monthly returns on the 25 Fama-French size and book-to-market ranked portfolios and five industry portfolios. The data are from February 1959 to July 2007 (582 observations). We report parameter estimates \(\hat{\lambda}\) (with \(\lambda_0\) multiplied by 100) and the model misspecification-robust t-ratio \((t\)-ratio\(_{pm}\)\).
TABLE 4
Tests of Equality of Cross-Sectional $R^2$s

Panel A: OLS

<table>
<thead>
<tr>
<th></th>
<th>C-LAB</th>
<th>FF3</th>
<th>ICAPM</th>
<th>CCAPM</th>
<th>CC-CAY</th>
<th>U-CCAPM</th>
<th>D-CCAPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAPM</td>
<td>-0.432</td>
<td>-0.631</td>
<td>-0.651</td>
<td>0.072</td>
<td>-0.251</td>
<td>-0.358</td>
<td>-0.657</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.001)</td>
<td>(0.055)</td>
<td>(0.818)</td>
<td>(0.440)</td>
<td>(0.367)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>C-LAB</td>
<td>-0.199</td>
<td>-0.219</td>
<td>0.504</td>
<td>0.182</td>
<td>0.075</td>
<td>-0.224</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.341)</td>
<td>(0.066)</td>
<td>(0.848)</td>
<td>(0.812)</td>
<td>(0.306)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FF3</td>
<td>-0.020</td>
<td>0.703</td>
<td>0.380</td>
<td>0.274</td>
<td>-0.025</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.865)</td>
<td>(0.000)</td>
<td>(0.031)</td>
<td>(0.226)</td>
<td>(0.742)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICAPM</td>
<td>0.723</td>
<td>0.400</td>
<td>0.293</td>
<td>0.006</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.067)</td>
<td>(0.279)</td>
<td>(0.967)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCAPM</td>
<td>-0.322</td>
<td>-0.429</td>
<td>-0.728</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.171)</td>
<td>(0.032)</td>
<td>(0.002)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC-CAY</td>
<td>-0.107</td>
<td>-0.406</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.701)</td>
<td>(0.037)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U-CCAPM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.299</td>
<td>(0.199)</td>
</tr>
</tbody>
</table>

Panel B: GLS

<table>
<thead>
<tr>
<th></th>
<th>C-LAB</th>
<th>FF3</th>
<th>ICAPM</th>
<th>CCAPM</th>
<th>CC-CAY</th>
<th>U-CCAPM</th>
<th>D-CCAPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAPM</td>
<td>-0.002</td>
<td>-0.191</td>
<td>-0.135</td>
<td>0.096</td>
<td>0.092</td>
<td>0.074</td>
<td>-0.132</td>
</tr>
<tr>
<td></td>
<td>(0.980)</td>
<td>(0.001)</td>
<td>(0.418)</td>
<td>(0.268)</td>
<td>(0.303)</td>
<td>(0.452)</td>
<td>(0.141)</td>
</tr>
<tr>
<td>C-LAB</td>
<td>-0.189</td>
<td>-0.133</td>
<td>0.098</td>
<td>0.094</td>
<td>0.075</td>
<td>-0.130</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.263)</td>
<td>(0.295)</td>
<td>(0.433)</td>
<td>(0.257)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FF3</td>
<td>0.055</td>
<td>0.287</td>
<td>0.283</td>
<td>0.264</td>
<td>0.059</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.696)</td>
<td>(0.008)</td>
<td>(0.008)</td>
<td>(0.021)</td>
<td>(0.608)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICAPM</td>
<td>0.231</td>
<td>0.227</td>
<td>0.209</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.110)</td>
<td>(0.117)</td>
<td>(0.170)</td>
<td>(0.986)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCAPM</td>
<td>-0.004</td>
<td>-0.023</td>
<td>-0.228</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.950)</td>
<td>(0.715)</td>
<td>(0.008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC-CAY</td>
<td>-0.019</td>
<td>-0.224</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.764)</td>
<td>(0.065)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U-CCAPM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.205</td>
<td>(0.140)</td>
</tr>
</tbody>
</table>

Note.—The table presents pairwise tests of equality of the OLS and GLS cross-sectional $R^2$s of eight beta pricing models. The models include the CAPM, the conditional CAPM (C-LAB) of Jagannathan and Wang (1996), the Fama and French (1993) three-factor model (FF3), the intertemporal CAPM (ICAPM) of Petkova (2006), the consumption CAPM (CCAPM), the conditional consumption CAPM (CC-CAY) of Lettau and Ludvigson (2001), the ultimate consumption CAPM (U-CCAPM) of Parker and Julliard (2005), and the durable consumption CAPM (D-CCAPM) of Yogo (2006). The models are estimated using monthly returns on the 25 Fama-French size and book-to-market ranked portfolios and five industry portfolios. The data are from February 1959 to July 2007 (582 observations). We report the difference between the sample cross-sectional $R^2$s of the models in row $i$ and column $j$, $\hat{\rho}_i^2 - \hat{\rho}_j^2$, and the associated p-value (in parentheses) for the test of $H_0 : \hat{\rho}_i^2 = \hat{\rho}_j^2$. The p-values are computed under the assumption that the models are potentially misspecified.
### TABLE 5
Sample Cross-Sectional $R^2$'s and Specification Tests of the Models Using Excess Returns

#### Panel A: OLS

<table>
<thead>
<tr>
<th></th>
<th>CAPM</th>
<th>C-LAB</th>
<th>FF3</th>
<th>ICAPM</th>
<th>CCAPM</th>
<th>CC-CAY</th>
<th>U-CCAPM</th>
<th>D-CCAPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\rho}^2$</td>
<td>0.858</td>
<td>0.893</td>
<td>0.958</td>
<td>0.972</td>
<td>0.880</td>
<td>0.886</td>
<td>0.946</td>
<td>0.883</td>
</tr>
<tr>
<td>$p(\rho^2 = 1)$</td>
<td>0.000</td>
<td>0.010</td>
<td>0.000</td>
<td>0.414</td>
<td>0.006</td>
<td>0.003</td>
<td>0.358</td>
<td>0.001</td>
</tr>
<tr>
<td>$p(\rho^2 = 0)$</td>
<td>0.001</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.003</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>se($\hat{\rho}^2$)</td>
<td>0.078</td>
<td>0.075</td>
<td>0.025</td>
<td>0.022</td>
<td>0.075</td>
<td>0.081</td>
<td>0.038</td>
<td>0.076</td>
</tr>
<tr>
<td>$\hat{Q}_c$</td>
<td>0.219</td>
<td>0.104</td>
<td>0.159</td>
<td>0.058</td>
<td>0.129</td>
<td>0.106</td>
<td>0.089</td>
<td>0.089</td>
</tr>
<tr>
<td>$p(Q_c = 0)$</td>
<td>0.000</td>
<td>0.001</td>
<td>0.000</td>
<td>0.166</td>
<td>0.000</td>
<td>0.001</td>
<td>0.014</td>
<td>0.007</td>
</tr>
<tr>
<td>No. of par.</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

#### Panel A: GLS

<table>
<thead>
<tr>
<th></th>
<th>CAPM</th>
<th>C-LAB</th>
<th>FF3</th>
<th>ICAPM</th>
<th>CCAPM</th>
<th>CC-CAY</th>
<th>U-CCAPM</th>
<th>D-CCAPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\rho}^2$</td>
<td>0.058</td>
<td>0.091</td>
<td>0.274</td>
<td>0.339</td>
<td>0.044</td>
<td>0.105</td>
<td>0.110</td>
<td>0.083</td>
</tr>
<tr>
<td>$p(\rho^2 = 1)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.075</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$p(\rho^2 = 0)$</td>
<td>0.005</td>
<td>0.314</td>
<td>0.000</td>
<td>0.003</td>
<td>0.190</td>
<td>0.387</td>
<td>0.028</td>
<td>0.216</td>
</tr>
<tr>
<td>se($\hat{\rho}^2$)</td>
<td>0.039</td>
<td>0.071</td>
<td>0.076</td>
<td>0.166</td>
<td>0.068</td>
<td>0.098</td>
<td>0.095</td>
<td>0.060</td>
</tr>
<tr>
<td>$\hat{Q}_c$</td>
<td>0.220</td>
<td>0.189</td>
<td>0.158</td>
<td>0.078</td>
<td>0.197</td>
<td>0.146</td>
<td>0.175</td>
<td>0.196</td>
</tr>
<tr>
<td>$p(Q_c = 0)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.017</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>No. of par.</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

**Note.**—The table presents the sample cross-sectional $R^2$ ($\hat{\rho}^2$) and the generalized CSRT ($\hat{Q}_c$) of eight beta pricing models. The models include the CAPM, the conditional CAPM (C-LAB) of Jagannathan and Wang (1996), the Fama and French (1993) three-factor model (FF3), the intertemporal CAPM (ICAPM) of Petkova (2006), the consumption CAPM (CCAPM), the conditional consumption CAPM (CC-CAY) of Lettau and Ludvigson (2001), the ultimate consumption CAPM (U-CCAPM) of Parker and Julliard (2005), and the durable consumption CAPM (D-CCAPM) of Yogo (2006). The models are estimated using monthly excess returns on the 25 Fama-French size and book-to-market ranked portfolios and five industry portfolios. The data are from February 1959 to July 2007 (582 observations). $p(\rho^2 = 1)$ is the $p$-value for the test of $H_0 : \rho^2 = 1$. $p(\rho^2 = 0)$ is the $p$-value for the test of $H_0 : \rho^2 = 0$. se($\hat{\rho}^2$) is the standard error of $\hat{\rho}^2$ under the assumption that $0 < \rho^2 < 1$. $p(Q_c = 0)$ is the $p$-value for the approximate $F$-test of $H_0 : Q_c = 0$. No. of par. is the number of parameters in the model.


**TABLE 6**  
**Multiple Model Comparison Tests**

**Panel A: OLS**

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>$\hat{\rho}^2$</th>
<th>$r$</th>
<th>$LR$</th>
<th>p-value</th>
<th>s</th>
<th>$\hat{\rho}_M^2 - \hat{\rho}^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAPM</td>
<td>0.115</td>
<td>2</td>
<td>0.844</td>
<td>0.259</td>
<td>4</td>
<td>0.734</td>
<td>0.057</td>
</tr>
<tr>
<td>C-LAB</td>
<td>0.548</td>
<td>5</td>
<td>1.056</td>
<td>0.330</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FF3</td>
<td>0.747</td>
<td>5</td>
<td>0.129</td>
<td>0.901</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICAPM</td>
<td>0.766</td>
<td>5</td>
<td>0.002</td>
<td>0.825</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCAPM</td>
<td>0.044</td>
<td>4</td>
<td>21.12</td>
<td>0.000</td>
<td>2</td>
<td>0.733</td>
<td>0.009</td>
</tr>
<tr>
<td>CC-CAY</td>
<td>0.366</td>
<td>5</td>
<td>4.728</td>
<td>0.059</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U-CCAPM</td>
<td>0.473</td>
<td>5</td>
<td>1.646</td>
<td>0.222</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-CCAPM</td>
<td>0.772</td>
<td>5</td>
<td>0.000</td>
<td>0.921</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Panel B: GLS**

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>$\hat{\rho}^2$</th>
<th>$r$</th>
<th>$LR$</th>
<th>p-value</th>
<th>s</th>
<th>$\hat{\rho}_M^2 - \hat{\rho}^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAPM</td>
<td>0.107</td>
<td>2</td>
<td>0.000</td>
<td>0.607</td>
<td>4</td>
<td>0.337</td>
<td>0.458</td>
</tr>
<tr>
<td>C-LAB</td>
<td>0.109</td>
<td>5</td>
<td>5.399</td>
<td>0.073</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FF3</td>
<td>0.298</td>
<td>5</td>
<td>0.000</td>
<td>0.866</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICAPM</td>
<td>0.242</td>
<td>5</td>
<td>0.153</td>
<td>0.567</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCAPM</td>
<td>0.011</td>
<td>4</td>
<td>7.349</td>
<td>0.021</td>
<td>2</td>
<td>0.248</td>
<td>0.092</td>
</tr>
<tr>
<td>CC-CAY</td>
<td>0.015</td>
<td>5</td>
<td>7.695</td>
<td>0.025</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U-CCAPM</td>
<td>0.034</td>
<td>5</td>
<td>5.456</td>
<td>0.053</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-CCAPM</td>
<td>0.239</td>
<td>5</td>
<td>0.264</td>
<td>0.563</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.**—The table presents multiple model comparison tests of the OLS and GLS cross-sectional $R^2$s of eight beta pricing models. The models include the CAPM, the conditional CAPM (C-LAB) of Jagannathan and Wang (1996), the Fama and French (1993) three-factor model (FF3), the intertemporal CAPM (ICAPM) of Petkova (2006), the consumption CAPM (CCAPM), the conditional consumption CAPM (CC-CAY) of Lettau and Ludvigson (2001), the ultimate consumption CAPM (U-CCAPM) of Parker and Julliard (2005), and the durable consumption CAPM (D-CCAPM) of Yogo (2006). The models are estimated using monthly returns on the 25 Fama-French size and book-to-market ranked portfolios and five industry portfolios. The data are from February 1959 to July 2007 (582 observations). We report the benchmark models in column 1 and their $R^2$s in column 2. $r$ in column 3 denotes the number of alternative models in each multiple non-nested model comparison. $LR$ in column 4 is the value of the likelihood ratio statistic with $p$-value given in column 5. $s$ in column 6 denotes the number of models that nest the benchmark model. Finally, $\hat{\rho}_M^2 - \hat{\rho}^2$ in column 7 denotes the difference between the sample $R^2$ of the expanded model ($M$) and the sample $R^2$ of the benchmark model with $p$-value given in column 8.