Journal of Banking & Finance xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Banking & Finance

journal homepage: www.elsevier.com/locate/jbf

On comparing zero-alpha tests across multifactor asset pricing models

Lieven De Moor [a,*], Geert Dhaene [b], Piet Sercu [c]

[a] Faculty of Economic and Social Sciences and Solvay Business School, Vrije Universiteit Brussel, Belgium
[b] Department of Economics, KU Leuven, Belgium
[c] Leuven School of Business and Economics, KU Leuven, Belgium

ARTICLE INFO

Article history:
Received 31 January 2015
Accepted 22 August 2015
Available online xxxx

Keywords:
Asset pricing
Pricing errors
Model comparison
Multifactor models

ABSTRACT

Evaluating competing multifactor asset pricing models involves comparing the statistical significance of their mean pricing errors (alphas). Unfortunately, this comparison favors imprecisely estimated models because *p*-values tend to be higher in more noisy models. To avoid false impressions of relative success at tests for zero mean pricing errors, we develop a notion of comparative *p*-values and suggest comparing these instead of the raw *p*-values. This comparison gives more precisely estimated models a fairer chance or, equivalently, quantifies how much easier it is for imprecisely estimated models, by comparison, to pass the test.

1. Introduction

To be deemed successful, a multifactor asset pricing model should meet two requirements. First, it should explain the expected excess returns better than alternative models that use different factors. Measures for this are the R^2 and significance tests for the factor exposures (for time-series regressions) or for the estimated prices of covariance risks (for cross-sectional regressions). Second, the covariance risks should explain *all* of the expected excess returns. This means not just that additional factors should be insignificant but also that, in regressions of excess returns on the factors, there should be no role for intercepts (the alphas). The latter condition can be tested by means of significance tests of the estimated alphas, separately or jointly.[1] This paper focuses on zero-alpha tests and particularly on the issue of unequal power of the competing regressions.

In Fama and French (2012), the term "power" occurs 21 times and is concentrated in the discussions of the test results and the comparison across models. They distinguish two problem situations, those with "too much" and those with "too little" power. The first arises when a model's R^2 is comparatively high and the power is, therefore, probably adequate. The null hypothesis of zero alphas is then sometimes rejected even when the pricing errors are small. Second, it may occur that "the regressions fit less tightly and power is a problem" (Fama and French, 2012, p. 458) because then the zero-alpha null hypothesis is easily accepted even when the pricing errors are relatively large.

Thus, the twin criteria of a good fit and zero alphas often contradict each other, which has implications for the comparison of competing models. In particular, comparing *p*-values of the zero-alpha null across models is contentious because the precision of the estimated alphas differs across models. Similarly, in a GMM framework, comparing Hansen's (1982) *J* statistics across alternative asset pricing models is controversial because it favors more noisy models (Jagannathan and Wang, 1996; Hansen and Jagannathan, 1997). Jagannathan and Wang (1996, p. 18) argue that "if a model contains 'more noise,' [...] then the value of the quadratic form will be smaller. In this case, it would be misleading to conclude that the 'more noisy' the model, the better it performs." In the same vein, Cochrane (2005, p. 216) notes that "it has proved nearly irresistible for authors to claim success for a new model over previous ones by noting improved J_T statistics, despite different weighting matrices, different moments, and sometimes much larger pricing errors." Again, the underlying problem is the unequal precision of the competing model estimates. In short, if a new model has "too little power" (i.e., its alpha estimates have greater standard errors than those in competing models) and rejects the zero-alpha null less often, one may well wonder about the extent to which this reflects

* Corresponding author.
E-mail addresses: lieven.de.moor@vub.ac.be (L. De Moor), geert.dhaene@kuleuven.be (G. Dhaene), piet.sercu@kuleuven.be (P. Sercu).

[1] There is a very large body of literature on (one- and multifactor) asset pricing models and testing. See, e.g., Jensen (1968), Black et al. (1972), Fama and MacBeth (1973), Basu (1977), Banz (1981), Jobson and Korkie (1982), Gibbons (1982), Stambaugh (1982), Gibbons et al. (1989), Kandel and Stambaugh (1989), MacKinlay and Richardson (1991), Shanken (1992), Fama and French (1992, 1993, 1996, 2012, 2015), Jegadeesh and Titman (1993), Jagannathan and Wang (1996), Hansen and Jagannathan (1997), Carhart (1997), Pastor and Stambaugh (2003), Liu (2006), Ray et al. (2009), Lewellen et al. (2010), Asness et al. (2013) and Beaulieu et al. (2013).

http://dx.doi.org/10.1016/j.jbankfin.2015.08.032
0378-4266/

poor precision rather than a genuine improvement. Likewise, "too much power" is a potential issue when the new model has more tightly estimated alphas and rejects the zero-alpha null more often.

In this paper, we compare zero-alpha p-values across models in a way that equalizes the precision of the estimated alphas. Our model comparisons are pairwise and we treat the models, new and existing, symmetrically. For a given pair of models and a single-alpha comparison, we revise the p-value of the model with the highest precision of the estimated alpha. Specifically, we ask what the p-value would have been had the precision been as poor as in the less-precise model. This means that we add to the estimated alpha of the more precise model a zero-mean random term with a standard deviation chosen such that the precision of the estimated alphas of the two models become equal. Each possible value of this random term implies a corresponding p-value of the zero-alpha null, and integrating this p-value over its distribution gives the *comparative p-value*, as we call it, of the more precise model. For the less-precise model, the comparative p-value is the same as the unmodified p-value, so that the two models can then be compared in terms of that metric. The comparison will reveal whether any success at the zero-alpha test of the lower-precision model relative to the more precise model is primarily due to its low precision rather than to a genuine improvement. For joint zero-alpha tests, the comparative p-values are defined similarly: for each model, we augment the estimated vector of alphas with a zero-mean random vector to equalize the precision of the model estimates and integrate the p-value derived from the ensuing Wald statistic over its distribution.

We emphasize that we make no claim whatsoever that the comparative p-values, taken individually, are any more valid than the original ones. Clearly, they are not: low precision cannot be conjured away. Our purpose is just to improve the way in which models with different precisions are compared and thus to avoid false impressions of relative success at zero-alpha tests when, in fact, the success is due to comparatively low precision. If we find, for instance, that a precisely estimated model is rejected given its low p-value but would be accepted in the light of its comparative p-value, we do not at all claim that the model's pricing errors are statistically insignificant. Rather, we note that, if they had been estimated as imprecisely as those in the imprecise model, we would not have worried about them. That conclusion qualifies the results for the imprecise model, not those for the precise one.

In Section 2, we review the link between R^2 and the power of zero-intercept tests. In Section 3, we introduce the comparative p-values and discuss their main properties. In Section 4, we give an illustration relating to the comparison between a liquidity-augmented CAPM and the Fama–French three-factor model. Section 5 concludes.

2. Sources of power

Fama and French (2012) treat R^2 and power almost as synonyms, but this misses a part of the picture that is relevant for our purposes. In a simple regression, it is true that, given the slope estimate, there is a one-to-one link between R^2 and the t-statistic for the slope, the first aspect of a CAPM test. But the relation with the second aspect, the t-test on the intercept, is more complicated. For illustrative purposes, consider a generic regression $Y = \alpha + \beta X + \varepsilon$ with, initially, one regressor and the classic OLS assumptions. We leave unspecified whether we consider a time-series or a cross-sectional test of $H_0 : \alpha = 0$.

Using standard notation, the OLS intercept estimate is $\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{X}$, telling us that the variance of $\widehat{\beta}$ affects the variance of $\widehat{\alpha}$ unless $\overline{X}$ equals zero. In a successful CAPM time-series regression, however, $\overline{X}$ is the risk premium, so one cannot count on a zero $\overline{X}$ to eliminate the effect of imprecise betas. Likewise, in a cross-section test, $\overline{X}$ is an average covariance or an average beta, which is usually nonzero. The link between the variances of $\widehat{\alpha}$ and $\widehat{\beta}$ is

$$\begin{aligned} \mathrm{var}(\widehat{\alpha}|X) &= \mathrm{var}(\overline{Y}|X) + \mathrm{var}(\widehat{\beta}|X)\overline{X}^2 \\ &= \frac{1}{n}\left(1 + \frac{\overline{X}^2}{\widehat{\mathrm{var}}(X)}\right)\mathrm{var}(\varepsilon), \end{aligned} \tag{1}$$

with n the number of observations, $\widehat{\mathrm{var}}(X) = \sum_i (X_i - \overline{X})^2/n$, and $\mathrm{var}(\varepsilon)$ the variance of the regression error (the unexplained excess return). In our context, all models use the same test returns Y (in excess of the risk-free rate). So for a given $\widehat{\mathrm{var}}(Y)$, the regression R^2 determines the estimate of $\mathrm{var}(\varepsilon)$ in (1). This is the channel that Fama and French (2012) frequently refer to.

For a given error variance, however, the intercept is also estimated imprecisely when the sample variance of the regressor is small relative to its squared mean. For example, if the test returns all have very similar covariances with the market excess return, a zero cross-sectional alpha is more easily accepted.

In the multifactor case, imprecise estimates $\widehat{\alpha} = \overline{Y} - \widehat{\beta}'\overline{X}$ can also arise from multicollinearity. Eq. (1) generalizes to

$$\mathrm{var}(\widehat{\alpha}|X) = \frac{1}{n}\left(1 + \overline{X}'\widehat{\mathrm{var}}(X)^{-1}\overline{X}\right)\mathrm{var}(\varepsilon),$$

with $\widehat{\mathrm{var}}(X) = \sum_i (X_i - \overline{X})(X_i - \overline{X})'/n$. Hence, α will be estimated imprecisely when the factors are highly collinear. In sum, the message is that anything that adversely affects the precision of the slope estimates also adversely affects the standard error of the intercept estimate.

3. Comparative *p*-values

Consider two competing asset pricing models, Model 1 and Model 2, each with its own set of factors, X_1 and X_2. For any asset excess return Y, Model j specifies the conditional mean excess return, given information set I, as $E(Y|I) = \alpha_j + \beta_j' X_j$ with $\alpha_j = 0$ (zero pricing error). Let $\widehat{\alpha}_j$ be an approximately unbiased, consistent, and asymptotically normal estimate of α_j, with consistent standard error estimate $\mathrm{se}(\widehat{\alpha}_j)$. The $\widehat{\alpha}_j$ may be OLS, GLS, or GMM estimates, and $\mathrm{se}(\widehat{\alpha}_j)$ may be a heteroskedasticity and/or autocorrelation robust estimate. We first define comparative p-values for the t-test and then for the Wald test.

3.1. Comparative p-values for a single alpha: definition

Using normal-theory asymptotic approximations, the p-value of the null $H_0 : \alpha_j = 0$ is

$$p_j = \Pr\left[|Z| > |t_j|\right] = 2\Phi(-|t_j|),$$

where $t_j = \widehat{\alpha}_j/\mathrm{se}(\widehat{\alpha}_j)$ is the t-statistic, Z is a standard normal variate, and $\Phi(\cdot)$ is the standard normal distribution function.

Now, if one compares p_1 with p_2, any difference in the standard errors $\mathrm{se}(\widehat{\alpha}_1)$ and $\mathrm{se}(\widehat{\alpha}_2)$ is ignored. It may thus occur that, for example, $|\widehat{\alpha}_2| > |\widehat{\alpha}_1| > 0$ and $\mathrm{se}(\widehat{\alpha}_2) \gg \mathrm{se}(\widehat{\alpha}_1)$, resulting in $p_1 < p_2$. In this case, Model 1 delivers a smaller pricing error, but, according to the p-value, Model 2 does better at the zero-alpha test because its alpha estimate is much less precise. The situation may arise, for example, when Model 2 adds a not-so-relevant factor to Model 1 that is highly collinear with the other factors. Obviously, it would be questionable, then, to claim success of Model 2 over Model 1 based on the p-value comparison alone. The problematic aspect

Please cite this article in press as: De Moor, L., et al. On comparing zero-alpha tests across multifactor asset pricing models. J. Bank Finance (2015), http://dx.doi.org/10.1016/j.jbankfin.2015.08.032