Testing the Momentum Anomaly

Abstract

The consensus view in asset pricing, shaped by the results of Fama and French (1996), is that the three-factor model fails to account for stock return momentum while the Carhart (1995, 1997) four-factor model 'explains' the returns of momentum-sorted portfolios. According to Cochrane (2006), "The three-factor is worse than useless…" and "the returns of … momentum-sorted portfolios can be explained by an additional momentum factor…." Using momentum-sorted portfolios and a variety of tests over five-year and longer sub-periods during 1965-2004, we show that the consensus is not strongly supported by the data. The performance of the three-factor model is qualitatively similar to that of the four-factor model both on statistical and economic grounds. Hence, attaching a greater weight to the results from the four-factor model relative to the three-factor model in empirical applications may not be justified.

I. Introduction

Stock return momentum (Jegadeesh and Titman (1993, 2001)) is widely regarded as one of the most robust anomalies in the empirical asset pricing literature. In an influential paper Fama and French (1996) conclude that the three-factor Fama-French model fails to account for the cross-sectional differences in momentum-sorted stock portfolios during the period July 1963-December 1993. This conclusion is based on a *F*-test of the null hypothesis that the intercepts are zero for the three-factor model. By contrast, Carhart (1995, p. 51) presents evidence using 27 size, book-to-market and momentum-sorted portfolios for the Fama-French period that the four-factor model "noticeably reduces the average pricing errors relative to the CAPM and the [three]-factor model."

Since then, a consensus view has emerged that (a) the three-factor model is rejected using momentum-sorted stock portfolios as test assets, and (b) the four-factor model is superior to the three-factor model in the context of the momentum anomaly. For example, according to Cochrane (2006, p. 19) "The three-factor model is worse than useless at capturing the expected returns of this "momentum" strategy…." He goes on to note that "…the returns of these 10 momentum-sorted portfolios can be explained by an additional "momentum factor" umd of winner stocks less loser stocks." Contrary to the consensus, this paper shows that the data for five-year sub-periods tend to favor the zero intercept null for the three-factor model, and, more generally, the data do not provide strong support for the four-factor model relative to the three-factor model. These contrary results are due to the application of recently developed robust tests that suffer from substantially less size distortion relative to the conventional tests. This highlights

that debates in financial economics can often be shaped by the interaction between evolving statistical methodology and data.

This study reexamines the momentum anomaly over the period 1965-2004. The performance of the three-factor and four-factor models is analyzed using monthly returns on ten equally-weighted and value-weighted momentum-sorted portfolios. We focus on model performance during eight five-year and four ten-year sub-periods, in addition to longer time periods. A key question is whether the three-factor model is in fact rejected by the data. The consensus view regarding the failure of the three-factor model is shaped by the results of the *F*-tests reported by Fama and French (1996) for the approximately thirty-year period July 1963 – December 1993. As noted above, we find that the evidence from five-year sub-periods is generally favorable to the three-factor model. Turning to the four-factor model, its performance is qualitatively similar to that of the three-factor model for both the five-year sub-periods as well as the longer sub-periods; the zero intercept null is often accepted for the shorter periods and generally rejected for the longer periods.

In light of this evidence, a natural question is how much credence should be given to the thirty-year periods compared to shorter periods. Long-standing concerns about parameter stability in empirical research in finance (see, for example, Fama and Macbeth (1973)) would argue in favor of shorter sub-periods. We construct plausible scenarios that suggest inferences based on long periods of monthly data may be problematic due to structural breaks. For this purpose, we employ a simulation design that incorporates structural breaks at five-year intervals. With this design, the zero-intercept null is falsely rejected when the test uses the entire thirty-year sample period. These results suggest

that concerns about parameter stability over long periods in tests of asset pricing models are well founded and hence that the long period results are best interpreted with caution.

Our analysis begins by examining the evidence based on the classic *F*-test. Our test results show that the consensus is not strongly supported by the data for the five-year sub-periods. The *F*-test does not reject the null that the intercept vector for the three-factor model is zero at the 1% level for the majority of the five-year sub-periods. In light of the evidence on violations of the assumptions of the classic *F*-test, we extend the analysis using the conventional heteroskedasticity and autocorrelation robust (HAR) Wald test. This test employs a heteroskedastic and autocorrelation consistent (HAC) estimator of the covariance matrix. We use the well-known HAC estimator proposed by Newey and West (1987, 1994) for the conventional HAR Wald test. The conventional HAR test with asymptotic *P*-values rejects the three-factor model for all the five-year as well as the longer sub-periods. These rejections require further examination because it is well known that the conventional test suffers from size distortions that occur with asymptotic *P*-values, that is, error in the rejection probability (ERP) under the null hypothesis.

To reduce the ERP, Keifer, Vogelsang and Bunzel (2000, hereafter KVB) and Keifer and Vogelsang (2005, hereafter KV) proposed the use of kernel-based covariance estimators in which the bandwidth parameter *M* is set proportional to the sample size *T*, that is, $M = bT$. In this case, when the parameter *b* is fixed as *T* goes to infinity, the kernel-based estimators have a random limiting distribution, which implies that they are inconsistent. In turn, the associated test statistics have nonstandard limit distributions.

The nonstandard or new HAR tests are carried out in practice by approximating the finite sample distribution of the test statistic by its nonstandard limit distribution.

In the Gaussian location model, Sun, Phillips and Jin (2008) have analyzed the ERP for tests where b is fixed as T goes to infinity and where the critical values are obtained from the nonstandard limit distribution. This ERP is compared to that for conventional tests with critical values obtained from the standard approximation. They show that the ERP of the nonstandard approximation is smaller than that of the standard approximation by an order of magnitude. This result is an extension of an earlier finding by Jansson (2004). These analytical findings support the earlier simulation results by KVB, KV (2002a, 20002b) and Phillips, Sun and Jin (2006, 2007, hereafter PSJ). The conclusion from this analysis is that the nonstandard approximation provides a more accurate approximation to the finite sample distribution of the test statistic. Consequently, the nonstandard test has less size distortion than the conventional test.

Ray and Savin (2008) investigated the performance of the new HAR tests using size-sorted portfolios, namely the case where stocks are assigned to portfolios based on market equity. Their study illustrates that the new HAR tests can change the inferences drawn from the data. Consistent with these results, our study shows that when the new HAR tests are applied to momentum data they deliver results that are in stark contrast to those of the conventional HAR tests.

In this paper, the new HAR tests fail to reject the three-factor model in at least four of the eight five-year sub-periods we examine. We show that the conflict between the results of the conventional HAR test and the new HAR tests is resolved when inferences are based on simulated finite-sample P-values. The finite- sample P-values

favor the three-factor model for a majority of the five-year sub-periods considered. In contrast, the finite-sample evidence is less favorable for the model over thirty-year and longer time periods.

We next investigate whether the intercepts, which represent the model pricing error, are consistently smaller in economic terms in the four-factor model compared to the three-factor model. We analyze the intercepts for the two models in each of the eight five-year sub-periods for the ten equally-weighted momentum-sorted portfolio returns. Our results indicate that there is little difference between the models in terms of the pricing errors for the five-year sub-periods. The median absolute value of the model intercepts across the ten portfolios in each five-year sub-period is smaller for the three-factor model in three out of the eight five-year sub-periods. In nearly all cases the differences in the median values of the intercepts are small in economic terms, less than or equal to 10 basis points per month in absolute value. In fact, the three-factor model yields a slightly higher proportion of intercepts smaller than 10 basis points per month in absolute value, 29 percent compared to 25 percent for the four-factor model.

In summary, this paper demonstrates that the consensus view regarding the merits of the Fama-French three-factor model, both on its own as well as relative to the four-factor model, should be reconsidered. While the momentum anomaly continues to be a challenge for asset pricing, it poses a challenge for both the three-factor as well as the four-factor model. Our analysis suggests that the four-factor model does not dominate the three-factor model in either statistical or economic terms. Hence, attaching greater weight to the results from the four-factor model may not be justified. The latter issue is

important in applications, for example, when interpreting tests of abnormal performance in event studies that employ the four-factor model.

The organization of the paper is the following: Section II reviews the classical Wald test for the vector of intercepts while Section III presents empirical results based on the classic test statistic. Section IV motivates the conventional and the new HAR tests in the case of a simple location model and provides the intuition behind the superior performance of the new tests. Section V describes the conventional and new HAR tests for testing the intercept in the three-factor model. HAR test results using asymptotic *P*-values are reported in Section VI, and results using simulated finite-sample *P*-values are presented in Section VII. The power of the new HAR tests for the five-year sub-periods is investigated in Section VIII. The impact of structural change on test results is considered in Section IX. Section X considers the economic significance of the intercepts from the three-factor and the four-factor models. Section XI contains the concluding comments.

II. Classic Wald Test

In this section, the three- and four-factor models used by Fama and French (1996) and Carhart (1997) are formulated as multivariate linear regression models with random regressors. The classic Wald test for the intercept vector is reviewed.

Define the variables $y_1,...,y_N$, where y_i is the excess return for the ith momentum portfolio, and the variables x_1, x_2, x_3, x_4, where x_1 is the market factor (the excess return on the market portfolio), x_2 is the size factor (the difference between the return on a portfolio of small capitalization stocks and the return on a portfolio of large capitalization stocks (SMB, small minus big)), x_3 is the book-to-market factor (the difference between

the return on a portfolio of high-book-to-market stocks and the return to a portfolio of low-book-to-market stocks (HML, high minus low)) and x_4 is the momentum factor (the average of the returns on two (big and small) high prior return portfolios minus the average of returns on two low prior return portfolios (MOM).

The classic Wald test is developed for the three-factor model. Suppose that the conditional expectation function is linear,

$$E(y \mid x) = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3, \tag{1}$$

and the conditional variances are constant,

$$V(y \mid x) = \Sigma, \tag{2}$$

where $y = (y_1, ..., y_N)'$, $\beta_0 = (\beta_{01}, ..., \beta_{0N})'$, $\beta_1 = (\beta_{11}, ..., \beta_{1N})'$, $\beta_2 = (\beta_{21}, ..., \beta_{2N})'$, and $\beta_3 = (\beta_{31}, ..., \beta_{3N})'$. A nonzero value of the intercept is interpreted as saying that the model leaves an unexplained return, a mean excess return that is unexplained by the three factors.

Denote the t-th observation on *y* by $y_{\bullet t} = (y_{1t}, ..., y_{Nt})'$ and on *x* by $x_{\bullet t} = (x_{1t}, x_{2t}, x_{3t})'$. In addition, suppose the population of y and x is randomly sampled, that is, the pairs ($y_{\bullet t}, x_{\bullet t}$) are independently and identically distributed (iid). Then random sampling from the above multivariate population supports the classical multivariate linear regression model with random regressors.

Following Greene (2003), the multivariate regression model can be restated as a seemingly unrelated regressions (SUR) model with identical regressors for the purpose of presenting the classic and conventional robust Wald tests. The SUR model is formulated using the N regression equations $y_{i\bullet} = X\theta_i + u_{i\bullet}, (i = 1, ..., N),$ where $y_{i\bullet} = (y_{i1}, ..., y_{iT})'$,

$X = [\iota, x_{1\bullet}, x_{2\bullet}, x_{3\bullet}]$, $\iota = (1,...,1)'$, $x_{j\bullet} = (x_{j1},..., x_{jT})'$, $(j = 1,2,3)$, $\theta_{\bullet i} = (\beta_{0i}, \beta_{1i}, \beta_{2i}, \beta_{3i})'$, and $u_{i\bullet} = (u_{i1},..., u_{iT})'$. Stacking the *N* regressions,

$$y_{\bullet\bullet} = (I \otimes X)\theta + u_{\bullet\bullet} = Z\theta + u_{\bullet\bullet},$$

where I is an *N*×*N* identity matrix, $\theta = (\theta_1', ..., \theta_N')'$, and $u_{\bullet\bullet} = (u_{1\bullet}', ..., u_{N\bullet}')'$. The least squares estimator of θ is obtained by regressing $y_{\bullet\bullet}$ on Z. This produces the estimator $\hat{\theta} = (Z'Z)^{-1}Z'y_{\bullet\bullet} = \theta + (Z'Z)^{-1}Z'u_{\bullet\bullet}$.

The null hypothesis of interest is $H_0 : \beta_0 = 0$, and the alternative is $H_1 : \beta_0 \neq 0$. The classic Wald statistic for testing H_0 is based on

$$W = c^{-1}\hat{\beta}_0'\hat{\Sigma}^{-1}\hat{\beta}_0,$$

where $c = (X'X)^{11}$ is the 1,1-th element of the inverse of $X'X$, $\hat{\beta}_0 = (I \otimes (1,0,0,0))\hat{\theta} = R\hat{\theta}$, $\hat{\Sigma} = T^{-1}\sum_t \hat{u}_{\bullet t}\hat{u}_{\bullet t}'$ and $\hat{u}_{\bullet t} = (y_{\bullet t} - \hat{\beta}_0 - x_{1t}\hat{\beta}_1 - x_{2t}\hat{\beta}_2 - x_{3t}\hat{\beta}_3)$. Under suitable regularity conditions, the statistic W has a limiting chi-square distribution with N degrees of freedom when H_0 is true.

In the case where the $y_{\bullet t}$ are independently distributed as $N(\beta_0 + x_{1t}\beta_1 + x_{2t}\beta_2 + x_{3t}\beta_3, \Sigma)$, or equivalently, the $u_{\bullet t}$ are iid *N*(0, Σ), the statistic $F = ((T - N - 3)/NT)W$ is unconditionally distributed as central *F* with N degrees of freedom in the numerator and (*T*-*N*-3) degrees of freedom in the denominator when H_0 is true. This follows from the fact that (*T*-4)*W*/*T* is a generalized Hotelling's T^{-2} statistic where $[(X'X)^{11}]^{-1/2}\hat{\beta}_0$ is distributed as $N(0,\Sigma)$ under H_0, and $T\hat{\Sigma}$ is independently distributed as a Wishart with parameters (*T*-4) and Σ; see Anderson (1958, Theorem 5.2.2, p. 106). For further treatment of testing in the normal case, see Stewart

(1997). Both Greene (2003) and Campbell et al. (1997) report the *F*-statistic for the one-factor model. However, in Greene, the denominator is missing the term *T*.

In the case of the four-factor model, X, θ and $u_{\bullet\bullet}$ are modified to incorporate x_4. If the $y_{\bullet t}$ are independently distributed as $N(\beta_0 + x_{1t}\beta_1 + x_{2t}\beta_2 + x_{3t}\beta_3, \Sigma)$, then the statistic $F = ((T - N - 4)/NT)W$ is distributed as a central F with N and (T-N-4) degrees of freedom when H_0 is true.

III. Empirical Results for the Classic Test

This section presents test results for the three- and four-factor models using the classic Wald test. These results are of interest because the qualitative conclusions are not essentially different from those produced by the new HAR tests with asymptotic *P*-values.

The data are obtained from Ken French's website (March 25, 2007). The return data consist of value-weighted and equally-weighted monthly returns for ten (N = 10) momentum-sorted portfolios. In this paper, the focus is on the sample from January 1965 through December 2006. The one-month Treasury bill as reported on the website is used as a measure of the risk-free return. The tests are performed for five-year, ten-year, thirty-year sub-periods and longer periods. The sub-periods include those used by Campbell et al. (1997) and Fama and French (1996).

Table 1 reports the results of the *F*-tests when the $u_{\bullet t}$ are iid $N(0, \Sigma)$. The consensus view is that the *F*-test rejects the zero intercept null for the three-factor model when it is estimated from equally-weighted returns. The test results show that this view is not strongly supported by the data for the five-year sub-periods: The null is not rejected for five out of eight sub-periods at the 1 percent level. Support for the consensus view is

primarily based on the results for the thirty-year and longer sub-periods. The null is rejected at the 1 percent level for all of the thirty-year and longer sub-periods. The story is essentially the same for the four-factor model. At the 1 percent level, the *P*-values do not reject the null for six out of the eight five-year sub-periods and one of the longer sub-periods. Notice that the four-factor model is not strictly superior to the three-factor model when comparing the magnitudes of the *P*-values.

Another aspect of the consensus is that the value-weighted returns favor the null. The *P*-values in Table 1 for the value-weighted returns tend to support this view, especially for the three-factor model. The null is not rejected at the 1 percent level for seven out of the eight five-year sub-periods for the three-factor model and five out of the eight five-year sub-periods for the four-factor model. The null is also not rejected at the 1 percent level for two out of the four of the ten-year periods for the three-factor model and three out of the four of the ten-year periods for the four-factor model. On closer examination, Table 1 reveals that the *P*-values for the value-weighted returns are not always larger than the *P*-values for the equally-weighted returns.

Although there are more rejections when the tests use the 5 percent level as the criterion, the number of rejections is the same for the three-factor and four-factor models for the equally-weighted portfolios, and there are fewer rejections for the three-factor models than the four-factor models for the value-weighted portfolios.

The *P*-value for a five-year sub-period is sensitive to the choice of dates and similarly for longer sub-periods. This is illustrated by using a different set of sub-periods for data analysis. Table 2 reports the *P*-values when all the sub-periods in Table I are shifted forward by two years. The results for the shifted sub-periods show that a shift by

two years can have a large impact on the *P*-value. For example, in the case of the three-factor model, shifting the five-year sub-period January 1965 - December 1969 to January 1967 - December 1971 increases the *P*-value from 0.24 percent to 8.55 percent for equally-weighted returns and reduces the *P*-value from 12.98 percent to 0.60 percent for value-weighted returns.

A comparison of Tables 1 and 2 for the equally-weighted returns shows that number of rejections for the five-year sub-periods is similar at the 1 percent level. However, at the 5 percent level, the shift in the five-year sub-periods considered here has a substantial effect. There are six rejections at the 5 percent level in Table 1 and only three in Table 2. In the case of value-weighted returns, the number of rejections in Table 1 is not much different from that in Table 2. Hence, the conclusion from Table I that the zero intercept null is often supported by the three-factor model in the case of five-year sub-periods is not overturned by the shifts in the sub-periods considered here.

The *P*-values in Table 1 are calculated on the assumption that the disturbances are normally distributed. The normality assumption may or may not be a good approximation to the actual distribution of the disturbances. Assuming that the iid assumption is correct, the relevant approximation to the actual distribution of the disturbances is the empirical distribution of the residual vectors. Table 3 reports the simulated *P*-values obtained by randomly resampling the empirical distribution of the residuals. The null hypothesis is imposed in the simulation experiments. Accordingly, the residual vectors are obtained by estimating the three- and four-factor models by constrained least squares where the constraint is that the intercept vector is zero. The constrained residual vectors are demeaned so that the mean of the empirical distribution

of the residual vectors is zero. In the experiments, the regressor vectors are randomly sampled with replacement from the empirical distribution of regressor vectors $x_{\bullet t}$ and, independently, the demeaned constrained residual vectors are randomly resampled with replacement. Further details of this resample-resample (RR) experiment are given in Appendix A.

The simulated *P*-values in Table 3 differ from the exact normal theory *P*-values in Table 1. They are larger than the exact normal theory *P*-values for all of the five-year and ten-year sub-periods for both the three-factor and four-factor models. This indicates that the empirical distribution of the regression errors have fatter tails than those of the normal distribution. The results are similar for the value-weighted returns. However, the qualitative results for Table 3 are the same as for Table 1. For example, for equally-weighted returns, the *P*-values do not reject the null at the 1 percent level for six out eight of the sub-periods for the three-factor and four-factor models. This suggests that the inferences based on Table 1 are relevant for the actual distributions of the disturbances. There are two complementary explanations for the similarity of the *P*-values. One is that the departures from normality are relatively small and the other is effects of the departures are washed out, at least partially, by the operation of the Central Limit Theorem.

The empirical results in this section show that the evidence from five-year sub-periods tends to be favorable to the three-factor model, while the evidence from the thirty-year and longer sub-periods is unfavorable to both the three-factor and four-factor models. The apparent contrast between the five-year sub-period results and those for the thirty-year and longer sub-periods is essentially accounted for by the smaller standard

deviations for the longer sub-period results, assuming no structural change over the extended time frames. This is because the estimates of the model pricing errors, that is, the intercepts, for the five-year sub-periods are roughly similar in magnitude to those for the thirty-year and longer sub-periods. As will be seen below, the rejection of the zero intercept null for the longer sub-periods is also a feature of the new HAR test results. However, it is plausible that the thirty-year and longer sub-periods may have structural breaks. In the presence of structural breaks the contrast may be one of appearance rather than substance. The potential impact of structural breaks on inference is investigated in more detail in Section 8 below.

IV. HAR Inference for the Mean

Campbell et al. (1997, p. 208) present evidence in finance on the failure of the classical assumptions to hold when tested with real data. The departures include nonnormality, heteroskedasticity and temporal dependence. These departures call for the use of robust tests. This section reviews HAR testing of the mean in the case of a simple location model. In the context of this model, the conventional and nonstandard or new tests reduce to *t*-tests. This simplification is an advantage when describing the motivation for and the construction of the tests. Another advantage is of the location model is that it permits an analytical investigation of the properties of the tests. In this section, results on the accuracy of the normal and the nonstandard approximations are reported, and the intuition behind the superior accuracy of the new HAR tests is discussed.

Following KVB and Jansson (2004), consider inference about β in the case of the location model:

$$y_t = \beta + u_t, (t = 1, ..., T)$$

where u_t is a zero mean process with a nonparametric autocorrelation process. The least squares estimator of β gives $\hat{\beta} = \bar{Y} = T^{-1}\sum_{t=1}^{T} y_t$, and the scaled and centered estimation error is

$$T^{1/2}(\hat{\beta} - \beta) = T^{-1/2} S_T,$$

where $S_t = \sum_{\tau=1}^{t} u_\tau$. Let $\hat{u}_\tau = y_\tau - \hat{\beta}$ be the time series of residuals. Suppose that S_T satisfies assumptions such that the estimation error converges in distribution to a normal distribution:

$$\sqrt{T}(\hat{\beta} - \beta) \Rightarrow \omega W(1) = N(0, \omega^2).$$

This result provides the usual basis for robust testing about β. Here ω^2 is the long run variance of u_t and W(r) is standard Brownian motion.

The conventional approach is to estimate ω^2 using kernel-based nonparametric estimators that involve some smoothing and possibly truncation of the autocovariances. When u_t is stationary with spectral density function $f_{uu}(\lambda)$, the long run variance (LRV) of u_t is

$$\omega^2 = \gamma_0 + 2\sum_{j=1}^{\infty} \gamma(j) = 2\pi f_{uu}(0),$$

where $\gamma(j) = E(u_t u_{t-j})$. The HAC estimates of ω^2 typically have the following form

$$\hat{\omega}^2(M) = \sum_{j=-T+1}^{T-1} k(\frac{j}{M})\hat{\gamma}(j), \quad \hat{\gamma}(j) = \begin{cases} T^{-1}\sum_{t=1}^{T-j} \hat{u}_{t+j}\hat{u}_t & \text{for } j \geq 0, \\ T^{-1}\sum_{t=-j+1}^{T} \hat{u}_{t+j}\hat{u}_t & \text{for } j < 0, \end{cases}$$

involving the sample covariances $\hat{\gamma}(j)$. In this expression, $k(\cdot)$ is some kernel; M is a bandwidth parameter and consistency of $\hat{\omega}^2(M)$ requires $M \to \infty$ and $M/T \to 0$ as

$T \to \infty$; see, for example, Andrews (1991), Hansen (2002) and Newey and West (1987, 1994).

To test the null $H_0 : \beta = \beta_0$ against the alternative $H_1 : \beta \neq \beta_0$, the conventional approach relies on a nonparametrically studentized *t*-ratio statistic of the form

$$t_{\hat{\omega}(M)} = T^{1/2}(\hat{\beta} - \beta_0) / \hat{\omega}(M),$$

which is asymptotically $N(0,1)$. The use of this *t*-statistic is convenient empirically and is widespread in practice, in spite of well-known problems with size distortion in inference.

To reduce size distortion, that is, the error in the rejection probability (ERP) under the null, KVB and KV(2005) proposed the use of kernel-based estimators of ω^2 in which the *M* is set equal to or proportional to *T*, that is, $M = bT$ for some $b \in (0,1]$. In this case, the estimator becomes

$$\hat{\omega}_b^2 = \sum_{j=-T+1}^{T-1} k\left(\frac{j}{bT}\right)\hat{\gamma}(j),$$

and the associated *t*-statistic is given by

$$t_b = T^{1/2}(\hat{\beta} - \beta_0) / \hat{\omega}_b .$$

The estimate $\hat{\omega}_b$ is inconsistent and tends to a random quantity instead of ω with the result that so the t_b -statistic is no longer standard normal.

When the parameter *b* is fixed as $T \to \infty$, KV showed that under suitable assumptions $\hat{\omega}_b^2 \Rightarrow \omega^2 \Xi_b$, where the limit Ξ_b is random. Under the null hypothesis,

$$t_b \Rightarrow W(1)\Xi_b^{-1/2} .$$

Thus, the t_b -statistic has a nonstandard limit distribution arising from the random limit of the LRV estimate $\hat{\omega}_b$.

Sun, Phillips and Jin (2008) have obtained the properties of the tests analytically under the assumption of normality. The assumption employed is that u_t is a mean zero covariance stationary Gaussian process with $\sum_{h=-\infty}^{\infty} h^2 |\gamma(h)| < \infty$. The ERP of the nonstandard *t*-test with *b* fixed is compared to that of the conventional *t*-test. The nonstandard test is based on the t_b -statistic and uses critical values obtained from the nonstandard limit distribution of $W(1)\Xi_b^{-1/2}$, while the conventional test is based on the $t_{\hat{\omega}(M)}$ -statistic and uses critical values from the standard normal distribution. Sun et al. show that the ERP of the nonstandard test is $O(T^{-1}),$ while that of the conventional normal test is $O(1)$. Hence, when *b* is fixed, the error of the nonstandard approximation to the finite sample distribution of the t_b -statistic under the null is smaller than that of the standard normal approximation to the finite sample distribution of the $t_{\hat{\omega}(M)}$ -statistic, again under the null. Moreover, the error of the nonstandard approximation is smaller than that of the normal approximation by an order of magnitude.

This result is related to that of Jansson (2004), who showed that the ERP of the nonstandard test based on the Bartlett kernel with $b = 1$ is $O(\log T/T)$. The Sun et al. (2008) result generalizes Jansson's result in two ways. First, it shows that the log (T) factor can be dropped. Second, while Jansson's result applies only to the Barlett kernel with $b = 1$, the Sun et al. result applies to more general kernels than the Bartlett kernel and to kernels with both $b = 1$ and $b < 1$.

There are two reasons for the improved accuracy of the nonstandard approximation. One is that the nonstandard distribution mimics the randomness of the denominator of the *t*-statistic. In other words, the nonstandard test behaves in large

samples more like its finite sample analogue than the conventional asymptotic normal test. By contrast, the limit theory for the conventional test treats the denominator of the *t*-ratio as if it were non-random in finite samples. The other reason is that the nonstandard distribution accounts for the bias of the LRV estimator resulting from the unobservability of the regressors errors, that is, the inconsistency mimics the bias.

In related work, PSJ (2006, 2007) proposed an estimator of ω^2 of the form

$$\hat{\omega}_\rho^2 = \sum_{j=-T+1}^{T-1} \left[k(\frac{j}{T}) \right]^\rho \hat{\gamma}(j),$$

which involves setting *M* equal to *T* and taking an arbitrary power $\rho \geq 1$ of the traditional kernel. The associated *t*-statistic $t_\rho = T^{1/2}(\hat{\beta} - \beta_0) / \hat{\omega}_\rho$ has a nonstandard limiting distribution arising from the random limit of the estimator $\hat{\omega}_\rho$ when ρ is fixed as $T \to \infty$. Statistical tests based on $\hat{\omega}_b^2$ and $\hat{\omega}_\rho^2$ share many of the same properties, which is explained by the fact ρ and *b* play similar roles in the construction of the estimates. An analysis of tests based on t_ρ is reported by PSJ (2005a, 2005b)

V. HAR Tests of the Three-Factor Model

This section presents the conventional HAR test and the new HAR tests for the intercept vector in the three-factor model. The extension to the four-factor model is straightforward.

From Section 2, the scaled and centered estimator is

$$\sqrt{T}(\hat{\theta} - \theta) = (T^{-1}Z'Z)^{-1}(T^{-1/2}Z'u_{\bullet\bullet}) = (I \otimes (T^{-1}X'X))^{-1}T^{-1/2}\sum_{t=1}^{T} v_{\bullet t},$$

where $v_{\bullet t} = u_{\bullet t} \otimes (1, x_{1t}, x_{2t}, x_{3t})'$. Under general assumptions, for example, those given in KV and PSJ (2005), the estimator converges in distribution to a normal:

$$\sqrt{T}(\hat{\theta}-\theta) \Rightarrow N(0, Q^{-1}\Omega Q^{-1})$$

where $Q = (I \otimes (p\lim T^{-1}X'X))$ and Ω is the long run variance of $\nu_{\bullet t}$. In the case of the three-factor model, Ω is a $4N \times 4N$ matrix

The conventional HAR Wald statistic for testing the null hypothesis $H_0 : \beta_0 = 0$ is

$$W_M = T\hat{\beta}_0' \left[R\hat{Q}^{-1}\hat{\Omega}(M)\hat{Q}^{-1}R' \right]^{-1} \hat{\beta}_0,$$

where $\hat{\Omega}(M)$ is an HAC estimator of Ω and $\hat{\alpha} = (I \otimes (1,0))\hat{\theta} = R\hat{\theta}$. When $H_0 : \beta_0 = 0$ is true, is asymptotically distributed as a chi-square with *N* degrees of freedom; for details, see KV.

The conventional approach to HAR testing relies on consistent estimation of the sandwich variance matrix $Q^{-1}\Omega Q^{-1}$. The term Q can be consistently estimated by $\hat{Q} = (I \otimes (T^{-1}X'X))$. When $\nu_{\bullet t}$ is stationary with spectral density matrix $f_{\nu\nu}(\lambda)$, the LRV of $\nu_{\bullet t}$ is

$$\Omega = \Gamma_0 + \sum_{j=1}^{\infty} (\Gamma(j) + \Gamma(j)') = 2\pi f_{\nu\nu}(0),$$

where $\Gamma(j) = E(\nu_{\bullet t}\nu_{\bullet t-j}')$. Consistent kernel-based estimators of Ω are typically of the form

$$\hat{\Omega}(M) = \sum_{j=-T+1}^{T-1} k\left(\frac{j}{M}\right)\hat{\Gamma}(j), \quad \hat{\Gamma}(j) = \begin{cases} T^{-1}\sum_{t=1}^{T-j} \hat{\nu}_{\bullet t+j}\hat{\nu}_{\bullet t}' & \text{for } j \geq 0, \\ T^{-1}\sum_{t=-j+1}^{T} \hat{\nu}_{\bullet t+j}\hat{\nu}_{\bullet t}' & \text{for } j < 0, \end{cases}$$

which involves sample covariances $\hat{\Gamma}(j)$ based on estimates $\hat{v}_{\bullet} = \hat{u}_{\bullet t} \otimes (1, x_{1t}, x_{2t}, x_{3t})'$ of $\nu_{\bullet t}$ that are constructed from regression residuals $\hat{u}_{\bullet t} = (y_{\bullet t} - \hat{\beta}_0 - x_{1t}\hat{\beta}_1 - x_{2t}\hat{\beta}_2 - x_{3t}\hat{\beta}_3)$.

The method proposed by Newey and West (1987, 1994) is used to obtain the HAC estimator of Ω for the conventional HAR test in this paper.

The new HAR Wald statistics used to test $H_0 : \beta_0 = 0$ are generalizations of the new *t*-statistics for testing the mean, namely t_b and t_ρ. When $M = bT$, the kernel-based estimator of Ω becomes

$$\hat{\Omega}_b = \sum_{j=-T+1}^{T-1} k\left(\frac{j}{bT}\right)\hat{\Gamma}(j),$$

and the associated test statistic is given by

$$W_b = T\hat{\alpha}'[R\hat{Q}^{-1}\hat{\Omega}_b\hat{Q}^{-1}R']^{-1}\hat{\alpha}.$$

In the case of exponentiated or power kernels, the estimator of Ω is

$$\hat{\Omega}_\rho = \sum_{j=-T+1}^{T-1} \left[k\left(\frac{j}{T}\right)\right]^\rho \hat{\Gamma}(j),$$

and the associated test statistic is given by

$$W_\rho = T\hat{\alpha}'[R\hat{Q}^{-1}\hat{\Omega}_\rho\hat{Q}^{-1}R']^{-1}\hat{\alpha}.$$

In this paper, two kernel functions are considered, both of which are commonly used in practice. One is the Bartlett kernel,

$$k(x) = \begin{cases} (1-|x|) & |x| \le 1, \\ 0 & |x| > 1, \end{cases}$$

and the other is the Parzen kernel,

$$k(x) = \begin{cases} (1-6x^2+6|x|^3) & \text{for } 0 \le |x| \le 1/2, \\ (2(1-|x|)^3) & \text{for } 1/2 \le |x| \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Taking an arbitrary power $\rho \ge 1$ of these kernels gives

$$[k(x)]^\rho = \begin{cases} (1-|x|)^\rho & |x| \le 1, \\ 0 & |x| > 1, \end{cases}$$

and

$$[k(x)]^{\rho} = \begin{cases} (1-6x^2+6|x|^3)^{\rho} & \text{for } 0 \le |x| \le 1/2, \\ (2(1-|x|)^3)^{\rho} & \text{for } 1/2 \le |x| \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

The properties of the kernels are discussed in PSJ (2006, 2007).

VI. Asymptotic Test Results

This section reports test results for equally-weighted returns for the three- and four-factor models using the conventional HAR test and the new HAR tests when the tests are based on asymptotic *P*-values. The asymptotic *P*-values are obtained from asymptotic chi-square distribution for the conventional test and the simulated nonstandard asymptotic distributions for the fixed-b and fixed-ρ tests.

The asymptotic P-values for the conventional HAR tests for the three-factor model are presented in Table 4. The asymptotic *P*-values reject the null at the 1 percent significance level for all sub-periods. A truncated Bartlett kernel (Newey and West (1987, 1994)) is used to calculate the HAC estimator. The bandwidth for the tabled results is $M = 6$. The results for the four-factor model are not reported since they are qualitatively the same as those for the three-factor model.

Regarding the bandwidth choice, in conventional approaches (Andrew (1991); Newey and West (1987, 1994)) the value of M is chosen to minimize the asymptotic mean squared error of the asymptotic standard error. Following this approach, the optimal M is approximated by the square root of T when the Bartlett kernel is employed. As Sun, Phillips and Jin (2008) note, this approach is not necessarily best suited for hypothesis testing. They advocate choosing a value of M that minimizes a loss function that involves a weighted average of Type I and Type II errors. Using this approach, the optimal M is approximated by the cube root of T when using the Bartlett kernel. We computed W_{M}

using both the square and cube root rules for M. For all sub-periods, W_M is smaller using the cube square root, and the asymptotic P-values are essentially zero for this choice. Hence, the null is rejected at the 1 percent level using both approaches. For simplicity, we report the P-values for $M = 6$ in Table 4 and subsequent tables. The value of W_M for $M = 6$ is between the value of W_M for the square root and cube root rules for the five-year sub-periods and is close to the value of W_M for the cube root rule for the ten-year and longer sub-periods.

In contrast to the results of the conventional test, the null is often not rejected by the asymptotic P-values for the new HAR tests, especially for the five-year and ten-year sub-periods. The asymptotic P-values for the new HAR tests are shown in Table 4 for the three-factor model. The asymptotic P-values for the fixed b-test do not reject the null at the 1 percent significance level for six out of the eight five-year sub-periods, for all of the four ten-year sub-periods, but do reject for all but one of the thirty-year and longer sub-periods. In the case of the fixed-ρ tests, the results are even more favorable for the three-factor model. The null is not rejected at the 5 percent level by the asymptotic P-values for the fixed-ρ tests for six out of eight-five sub-periods, for all of the ten-year sub-periods, and for three out of the six thirty-year and longer sub-periods. The null is rejected for the Fama-French period by all the tests. The asymptotic P-values for the fixed-b tests are calculated using the Bartlett kernel and b =1 and those for the fixed-ρ tests use the Parzen kernel and ρ = 32. The results are qualitatively similar for values of b = 0.5 and for ρ = 16.

The difference between the results for the conventional test and the new HAR tests suggests that the chi-square distribution provides a poor approximation to the finite

sample distribution of the conventional test statistic for the shorter sub-periods. The next section reports evidence that shows the chi-square distribution is a poor approximation for the five-year and ten-year sub-periods.

VII. Finite Sample Test Results

This section reports simulated finite-sample *P*-values for the conventional and the new HAR tests for the three-factor model. The simulated *P*-values are calculated for the three forms of the HAR test in four different experiments. The four experiments are conducted for each of the sub-periods.

A description of the experiments for the January 1965 -1969 sub-period follows. The value of $y_{\bullet t}$ is simulated using the constrained least squares estimate of the conditional expectation function under the null:

$$y^*_{\bullet t} = x^*_{1t}\tilde{\beta}_1 + x^*_{2t}\tilde{\beta}_2 + x^*_{3t}\tilde{\beta}_3 + u^*_{\bullet t} \quad (t = 1, ..., T),$$

where $y^*_{\bullet t}, x^*_{\bullet t}, u^*_{\bullet t}$ are the simulated values of $y_{\bullet t}$, $x_{\bullet t}$, $u_{\bullet t}$ and $\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3$, the constrained least squares estimates of the slope vectors calculated from the sample data for the sub-period. The simulated finite-sample *P*-values are conditional on the values of the parameters not specified by the null, that is, the nuisance parameters. The nuisance parameters include not only the slope parameters but also those that specify the process generating the factors and the errors. The values of the nuisance parameters are set equal to estimates based on the sample data. The level of the tests refers to the probability of a Type I error, not the size where the latter is defined as the maximum level over all admissible values of the nuisance parameters.

A brief statement of the purpose of each of the *P*-value experiments is the following:

Normal-Normal (NN) Experiment. This experiment produces data that satisfy the assumptions of the classical normal SUR model with normally distributed regressors.

Resample-Resample (RR) Experiment. This experiment captures the non normality present in the data.

Normal-VAR (NV) Experiment. This experiment introduces serial correlation in the errors.

Resample-Block (RB) Experiment. This experiment allows for volatility clustering of the returns.

Details of the steps in the simulation procedure for the each of the four experiments are given in Appendix A.

The NN and RR experiments provide evidence on how the tests perform when the multivariate iid assumption holds with and without normality. If the tests exhibit poor performance under this assumption, it is unlikely that they will perform well in the presence of autocorrelation or volatility clustering. The rationale for the NV and RB experiments is a substantial body of evidence in finance documenting departures from the iid assumption for stock portfolio returns. A preliminary check for heteroskedasticity and autocorrelation for momentum based portfolios was carried out using the Breusch and Pagan (1979) test for heteroskedasticity as modified by Cook and Weisberg (1983) and the Breusch-Godfrey test for autocorrelation (Godfrey (1988)). These tests were performed equation-by-equation for each sub-period. The test of heteroskedasticity tended to reject the null of constant variance at the usual levels. The test for zero

autocorrelation did not reject for many of the five-year and ten-year sub-periods, especially for the higher momentum portfolios.

There is considerable evidence that asset return volatility is both time-varying and predictable, again for portfolios sorted by market equity; for example, see Bollerslev (1986) and Bollerslev et al. (1994). As a preliminary check for autoregressive conditional heteroskedasticity, a GARCH (1, 1) model for the errors was estimated equation-by-equation for each sub-period. The maximization of the pseudo-log likelihood tended to fail for the five-year sub-periods and for some of the ten-year sub-periods. Estimates of the ARCH and GARCH coefficients were obtained for the thirty-year and longer sub-periods, and these were often significantly different from zero.

In the simulation experiments, it was not feasible to generate the errors for each period using an estimated multivariate GARCH model. Instead, we use a procedure that is employed in bootstrap sampling with dependent data. The procedure is to divide the residual vectors for each sub-period into blocks, and then randomly resample the blocks with replacement. In the RB experiments, six-month length blocks were chosen because this is approximately the half-life of an estimated univariate GARCH process for monthly stock returns; for example, see French et al. (1987) for estimates for the period 1928-1984.

More generally, the RB experiments capture dependence in the errors. There are other processes that may be generating dependence in addition to autoregressive conditional heteroskedasticity. These include ARMA models and also models that produce non-martingale difference sequences such as nonlinear moving average and

bilinear models. Consequently, the results of the RB experiments cannot be interpreted as only due to volatility clustering, although this may be the dominant effect.

Table 5 reports the simulated finite-sample *P*-values for the conventional and new HAR tests for the four experiments. The simulated finite-sample *P*-values tend to be larger for the RR, NV and RB experiments than for the NN experiments. As a consequence, in the RR, NV and RB experiments, the null is not rejected at the 5 percent level as well as the 1 percent level for most of the five-year and ten-year sub-periods. The results are qualitatively similar for values of M = 4, $b = 0.5$ and for $\rho = 16$. In summary, the finite-sample evidence favors the null for most of the shorter sub-periods, even when using the conventional test statistic.

Table 5 illustrates that the asymptotic chi-square distribution may be a poor approximation to the finite-sample distribution of the conventional test statistic for samples of T = 60 and T =120. A comparison of Tables 4 and 5 shows that the difference between the asymptotic and finite-sample *P*-values is much smaller for the new HAR tests than for the conventional test. This is consistent with the simulation results of KVB, KV (2002a, 20002b) and PSJ (2006, 2007) and the theoretical results obtained by Jansson (2004) and PSJ (2008). Our results are also consistent with previous studies in business and financial economics that have documented finite-sample size distortions in the conventional HAR tests; see, for example, Ferson and Foerster (1994), Burnside and Eichenbaum (1996) and Altonji and Segal (1996).

VIII. Power of New HAR Tests

This section reports simulated level-corrected powers of the conventional and the new HAR tests. The main motivation for calculating the powers is the frequent non-

rejections of the null by the new HAR tests. The level-corrected powers are calculated for the three forms of the HAR test in four different experiments. The four experiments are conducted for each of the sub-periods.

The simulated powers are estimates of the true level-corrected powers conditional on the experimental design. The design specifies the vector of intercepts under the alternative, the nuisance parameters including the slope vectors and the long run variance matrix and the process generating the factors as well the errors.

The powers are calculated for a test of H_0 against the alternative $H_1 : \beta_0 = c\iota(0.001), |c| > 0$. Here the alternative intercept vector β_0 is proportional to a vector of ones, ι, where *c* is a scalar. With this setup, a unit increase in *c* translates into an increase of 10 basis points per month in the intercept. In Section 9, we present the portfolio intercepts for the three-factor and four-factor models for each of the ten equally-weighted momentum sorted portfolios for the eight five-year sub-periods. The median absolute values of the intercepts for the three-factor model range between 8.43 and 68.18 basis points per month. The median values are similar for the four-factor model. This suggests that the empirically relevant range for alternative values of the intercepts, and hence *c* is from *c* = 1 to *c* = 6. This setup provides a natural metric for interpreting the power, which is often absent in power studies.

A description of the power experiments for the January 1965 to 1969 sub-period follows. The value of $y_{\bullet t}$ is simulated using

$$y^*_{\bullet t} = \beta_0 + x^*_{1t}\tilde{\beta}_1 + x^*_{2t}\tilde{\beta}_2 + x^*_{3t}\tilde{\beta}_3 + \tilde{u}^*_{\bullet t} \quad (t = 1, ..., T),$$

where $y^*_{\bullet t}, x^*_{\bullet t}, u^*_{\bullet t}$ are the simulated values of $y_{\bullet t}, x_{\bullet t}, u_{\bullet t}$. The intercept vector β_0 is known constant given by the alternative H_1. The slopes $\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3$ are obtained by running a constrained least squares regression of $y_{\bullet t}$ on $x_{\bullet t}$ for the sample data where the constraint is $\beta_0 = 0$. Further details of the power experiments are given in Appendix B.

The powers for the RR experiments are reported in Table 6. The powers are reported only for positive values of *c* since the power curves are symmetric in *c*. The results show that the tests tend to have high level-corrected power against empirically relevant departures from the null. The level-corrected powers tend to be high at *c* = 5 (monthly pricing error of 50 basis points), and, although not reported, close to one at *c* = 6. The power results support the conclusion that the non-rejections by the fixed-*b* tests and fixed-ρ tests are not due to low power. The same conclusion is supported by the results from the NN, NV and RB power experiments.

IX. Structural Change

As noted in the introduction, support for the consensus view about the three-factor model rests primarily on the results for the thirty-year and longer sub-periods. This section examines the interpretation of the results for the longer sub-periods when the five-year sub-periods represent structural breaks.

The results reported in this paper for the ten-year and longer sub-periods implicitly assume that the data generation process is constant over the length of the sub-period. This assumption implies that the longer the period, the higher the precision of the estimator, and hence the higher the power of the test. In turn, the higher power is reflected in a smaller *P*-value. This interpretation is the basis for attaching greater weight to the results for the longer sub-periods. The problem with this approach is that it

can produce misleading conclusions if there are structural breaks in the longer sub-periods. A more plausible assumption is that the data generation process is changing over time. A well known procedure in finance for coping with structural breaks is to divide the time series into five-year sub-periods. As in Fama and Macbeth (1973), the motivation for using a five-year sub-period is the presumption that the change is small over a period of this length.

Table 7 illustrates what can go wrong in the presence of structural breaks. The table reports simulated *P*-values for the conventional test for the thirty-year sub-periods where the data is generated by a sequence of six models, one for each five-year sub-period. The RR design is used to simulate the ten portfolio returns in each sub-period. The simulated finite-sample *P*-values for the January 65-December 1994 sub-period are presented in Panels A and B.

In Panel A the zero intercept null is imposed in each sub-period. Despite the fact that the null is true in each of the six five-year sub-periods that constitute the thirty-year period, the simulated finite-sample *P*-values for the thirty-year sub-periods are less than 1 percent implying a Type I error. In Panel B the null is not imposed in each five-year sub-period. Instead, the values of the intercepts are set equal to their empirical counterparts. In this exercise, the finite-sample *P*-values in percentage terms turned out to greater than 5 percent, and hence the null was not rejected for the thirty-year sub-periods even though the null was false for some of the five-year sub-periods. With respect to the two other thirty-year sub-periods in Table 7, the null is rejected in some cases in Panel A when it is in fact true, and the null is not rejected in Panel B when it is in fact false for some five-year sub-periods.

These results illustrate that using the entire thirty-year sample period for testing can lead to incorrect rejections of the three-factor model. This suggests that the thirty-year evidence is problematic at best, and that long-standing concerns about parameter stability in empirical tests are well-founded. In other words, the seeming contrast between the results for the five-year sub-period and the thirty-year periods may be one of appearance rather than substance.

X. Economic Significance

The three-and four-factor models can be evaluated in terms of economic significance as well as by using tests of statistical significance. Following Fama and French (1996), this section considers the economic significance of the model pricing errors as measured by the absolute value of the intercepts translated into basis points. Of course, the model pricing errors are subject to sampling variation. In the spirit of Cochrane (1996, p. 17), our objective is to provide the reader with a sense of the economic magnitude of estimated pricing errors to supplement the evidence from the formal statistical tests presented earlier.

In the finance literature, a monthly excess return of 10 basis points is considered small in terms of the model pricing error; see Fama and French (1996, p.57). Table 8 shows that the momentum-sorted portfolio intercepts are generally larger than 10 basis points. This table displays the portfolio intercepts from the Fama-French model (Panel A), and the Carhart model (Panel B) for each of the ten equally-weighted momentum-sorted portfolios for the eight five-year sub-periods. The intercepts are reported in terms of absolute values and expressed in basis points per month. The median values of the

intercepts for the three-factor model range between 8.43 and 68.18. The median values are similar for the four-factor model.

For the eight five-year sub-periods in our sample, the median absolute value of the model intercept across the ten portfolios is smaller for the three-factor model in three cases. It can be readily calculated from Table 8 that the proportion of the portfolio intercepts smaller than 10 basis points per month, is slightly higher for the three-factor model (29 percent) compared to the four-factor model (25 percent). As expected, the four-factor model does a better job of explaining returns for the high momentum portfolio (portfolio 10). This is evidenced by the fact that the four-factor model intercepts are almost always smaller for the high momentum portfolio. However, even for this portfolio, the four-factor model intercepts are higher than 10 basis points in absolute value, in seven out of the eight five-year sub-periods.

To underscore that the four-factor model does not consistently dominate the three-factor model, Panel C presents the ratios of the absolute value of the Fama-French model intercepts to the absolute value of the corresponding Carhart model intercepts. In many cases, these ratios are less than one. In the same spirit, Figure 1 displays the actual (not absolute) values of the intercepts for each portfolio for all eight five-year sub-periods for the three-factor model (top panel) and the four-factor model (bottom panel). The figure illustrates that the intercepts for the both the three-factor and four-factor models are often quite large in magnitude and that the four-factor model intercepts are not noticeably smaller. In summary, it is clear that the four-factor model does not dominate the three-factor model when judged on the basis of the economic significance of the model pricing errors.

XI. Concluding Comments

The empirical evidence of momentum in stock returns has proven to be a challenge for rational asset pricing. In the face of this anomaly, a consensus view has emerged in the finance literature during the past decade regarding the relative merits of some well known asset pricing models. According to the consensus, the Fama-French (1993) three-factor model is rejected by the data, and the Carhart (1995, 1997) four-factor model dominates the three-factor model in both statistical and economic terms. In this study we examine the momentum anomaly over the period 1965-2004 using monthly returns on ten momentum-sorted stock portfolios. Our analysis indicates that the performance of the three-factor and four-factor models is qualitatively similar for both the five-year sub-periods as well as the longer sub-periods; the zero intercept null is often accepted for the shorter periods and generally rejected for the longer periods.

As noted earlier, Cochrane (2006) considers the three-factor model to be "worse than useless" in the context of the momentum anomaly. If this were true, the logical implications of our findings would be that the Carhart four-factor model also does not merit serious consideration. Fortunately, both the three-factor and four-factor models perform better than claimed by Cochrane, although they suffer from limitations

We provide evidence on the performance of the models using a variety of test procedures. These include the classic *F*-test, the conventional HAR Wald test and the new HAR tests developed by KVB, KV (2005), and PSJ (2006, 2007). The test results, when based on finite sample *P*-values, produce a consistent message which can be readily summarized. The zero-intercept null hypothesis for the three-factor model is not rejected for most five year sub-periods during 1965-2004. In contrast to the five-year sub-

periods, the evidence for the thirty-year or longer sub-periods is less favorable for both the three-factor and four-factor models.

We identify a major concern with regards to the interpretation of the evidence for the longer periods: How much weight should be attached to the evidence from thirty-year periods relative to the shorter periods, given the potential for structural breaks over long periods? To investigate this issue we employ a simulation design in which portfolio returns are generated according to the three-factor model, subject to structural breaks at five-year intervals. We confirm that using the entire thirty-year sample period for testing can lead to incorrect rejections of the three-factor model. This suggests that the thirty-year evidence is problematic at best, and that long-standing concerns about parameter stability in empirical tests are well-founded.

We next show that the three-factor model performs no worse than the four-factor model when judged by the magnitude of the model intercepts, which provide an economic measure of the model pricing error. Hence, the four-factor model fails to dominate the three-factor model in both statistical and economic terms.

The central message of this paper is that the consensus view regarding the appropriate benchmark model to be used in a range of empirical applications needs to be revised. As noted in the introduction, one has to be cautious when interpreting measures of abnormal returns in event studies that rely on the four-factor model. Similar caution is advisable when relying on the four-factor model intercept or alpha as the measure of managerial skill in the performance evaluation of managed portfolios. The four-factor model is clearly useful in applications where the primary objective is to control for the momentum investing style favored by many managers. However, in light of our results,

the four-factor alpha may not necessarily be a better measure of managerial skill than the three-factor alpha. In this context, the framework suggested by Pástor and Stambaugh (1999) and Pástor (2000) that explicitly allows for the possibility of less-than-perfect model pricing ability in various applications, represents an important advance. Our analysis provides further evidence that the search for a satisfactory factor model for asset pricing remains unfinished business.

Appendix A: *P*-Value Experiments

Normal-Normal (NN) *P*-value Experiment. The *P*-value simulation procedure consists of five steps:

S1. Generate a sample of T = 60 $x^*_{\bullet t}$ vectors by randomly sampling the $N(\bar{x}, S)$ distribution where $\bar{x} = T^{-1}\sum_t x_{\bullet t}$ and $S = T^{-1}\sum_t (x_{\bullet t} - \bar{x})(x_{\bullet t} - \bar{x})'$ are calculated from sample data for the sub-period.

S2. Generate a sample of $T = 60$ $u^*_{\bullet t}$ vectors independently of $x^*_{\bullet t}$ by randomly sampling the $N(0, \tilde{\Sigma})$ distribution where $\tilde{\Sigma} = T^{-1}\sum_t (\tilde{u}_{\bullet t} - \bar{\tilde{u}}_{\bullet})(\tilde{u}_{\bullet t} - \bar{\tilde{u}}_{\bullet})'$ and $\bar{\tilde{u}}_{\bullet} = T^{-1}\sum_t \tilde{u}_{\bullet t}$ are calculated from the constrained residual vectors $\tilde{u}_{\bullet t} = (y_{\bullet t} - x_{1t}\tilde{\beta}_1 - x_{2t}\tilde{\beta}_2 - x_{3t}\tilde{\beta}_3)$ for the sub-period.

S3. Generate a sample of $T = 60$ $y^*_{\bullet t}$ vectors from (9) using the $x^*_{\bullet t}$ vectors from S1, the $u^*_{\bullet t}$ vectors from S2 and the constrained least squares estimates as the values for the slope parameters.

S4. Compute the three forms of the HAR test statistic from the simulated dataset of size $T = 60$.

S5. Repeat steps S1, S2, S3 and S4 10,000 times. Compute the *P*-value for each form of the HAR test statistic from the empirical distribution of the test statistic.

Resample-Resample (RR) *P*-value Experiment. In this and the remaining experiments, only one or both of the first two steps differ from those in the NN experiment.

S1′. Generate a sample of $T = 60$ $x^*_{\bullet t}$ vectors by randomly sampling with replacement the observations $x_{\bullet t}$.

S2′. Generate a sample of $T = 60$ $u^*_{\bullet t}$ vectors independently of $x^*_{\bullet t}$ by randomly sampling with replacement the demeaned constrained least squares residuals $\tilde{u}_{\bullet t} - \bar{\tilde{u}}_{\bullet}$.

Normal-VAR (NV) *P*-value Experiment. The first step is the same as in the NN experiment.

S2″. Generate a sample of $T = 60$ $u^*_{\bullet t}$ vectors independently of $x^*_{\bullet t}$ using a Gaussian VAR(1) process

$$u^*_{\bullet t} = \tilde{\Phi} u^*_{\bullet t-1} + \eta^*_{\bullet t},$$

where $\tilde{\Phi}$ is a 10×10 matrix of autoregressive coefficients. The autoregressive matrix $\tilde{\Phi}$ is obtained by a least squares regression of $\tilde{u}_{\bullet t}$ on $\tilde{u}_{\bullet t-1}$ using the constrained least square residuals for the sub-period. The vector $\eta^*_{\bullet t}$ is randomly sampled from the $N(0, \tilde{\Sigma}_\eta)$ distribution, where $\tilde{\Sigma}_\eta = T^{-1}\sum_{t=1}^{T}(\tilde{\eta}_{\bullet t} - \bar{\tilde{\eta}}_{\bullet})(\tilde{\eta}_{\bullet t} - \bar{\tilde{\eta}}_{\bullet})'$ and $\bar{\tilde{\eta}}_{\bullet} = T^{-1}\sum_t \tilde{\eta}_{\bullet t}$ are calculated from the VAR residuals. The conditions for covariance-stationarity are checked by calculating the roots of the $\tilde{\Phi}$ matrix. In each replication, the initial values of $u^*_{\bullet t-1}$ in the VAR (1) are set equal to zero, and the first 200 draws are discarded in order make the results independent of the initial values.

Resample-Block (RB) *P*-value Experiment. The first step is the same as in the RR experiment.

S2‴. Generate a sample of $T = 60$ $u^*_{\bullet t}$ vectors independently of $x^*_{\bullet t}$ by randomly sampling with replacement the demeaned constrained least squares residuals

$\tilde{u}_{\bullet_t} - \bar{\tilde{u}}_{\bullet}$ in consecutive fixed-length non-overlapping blocks where the block length is six months.

Appendix B: Power Experiments

Normal-Normal (NN) Power Experiment. The power simulation procedure consists of five steps for each value of c. For $c = 0$, steps S1, S2, S3 and S4 are the same as in the *P*-value simulation procedure. The fifth step is:

S5. Repeat steps S1, S2 S3 10, 000 times. Compute the 5 percent critical value for each form of the HAR test statistic from the empirical distribution of the test statistic under H_0 ($c = 0$).

For $c = 1$, steps S1, S2, S3, S4 are the same as the *P*-value simulation procedure. The fifth step is:

S5. Repeat steps S1, S2, S3 and S4 10,000 times. Compute the power for each form of the HAR test statistic from the empirical distribution of the test statistic using the simulated five percent critical value obtained from the $c = 0$ experiment.

For $c > 1$, the power experiments are similar to those for $c = 1$.

The steps in the RR, NV and RB power simulation experiments are obtained by making the analogous changes to the RR, NV and RB *P*-value simulation experiments.

REFERENCES

Altonji, J., Segal, L., 1996. Small-sample bias in GMM estimation of covariance structures. *Journal of Business and Economic Statistics* 14: 353-366.

Anderson, T. W., 1958. An Introduction to Multivariate Statistical Analysis. John Wiley and Sons: New York.

Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59: 817-854.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31: 307-327.

Bollerslev, T., Engle, R. R., Nelson, D. B., 1994. Arch Models. In: Handbook of Econometrics, vol. 4, Engle, R. R., McFadden, D. L. (eds). North-Holland: Amsterdam.

Breusch, T., Pagan, A., 1979. A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47: 1287-1294.

Burnside, C., Eichenbaum, M., 1996. Small-sample properties of GMM-based Wald tests.
Journal of Business and Economic Statistics 14: 294-308.

Campbell, J.Y., Lo, A. W., MacKinlay, A. C., 1997. The Econometrics of Financial Markets. Princeton University Press: Princeton, New Jersey.

Carhart, M.,1995. Survivor Bias and Persistence in Mutual Fund Performance. Unpublished dissertation, University of Chicago.

Carhart, M.,1997. On Persistence in Mutual Fund Performance. *Journal of Finance* 52: 57-82.

Cochrane, J., 2006. Financial markets and the real economy. Working paper. University of Chicago.

Cook, R.D., Weisberg, S., 1983. Diagnostics for heteroscedasticity in regression. *Biometrika* 70: 159-178.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33: 3-56.

Fama, E. F., French, K. R., 1996. Multifactor explanations of asset pricing anomalies. *The Journal of Finance* 51: 55-83.

Fama, E.F., Macbeth, J., 1973. Risk, return and equilibrium: empirical tests. *Journal of Political Economy* 81: 607-36

Ferson, W.E., Foerster, S.R., 1994. Finite sample properties of the generalized method of moments in tests of conditional asset pricing models. *Journal of Financial Economics* 36, 29– 55.

French, K.R., G. W. Schwert, Stambaugh, R.F., 1987. Expected stock returns and volatility. *Journal of Financial Economics* 19: 3-30.

Godfrey, L. G., 1988. Misspecification Tests in Econometrics. Econometric Society Monographs, No. 16. Cambridge University Press: Cambridge

Greene, W. H., 2003. Econometric Analysis, fifth edition. Prentice Hall: New Jersey.

Jansson, M., 2004. The error in rejection probability of simple autocorrelation robust tests. *Econometrica* 72: 937-946.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: implications for stock market efficiency, *Journal of Finance* 48, 65-91.

Jegadeesh, N., Titman, S., 2001. Profitability of momentum strategies: An evaluation of alternative explanations, *Journal of Finance* 56, 699-720.

Kiefer, N. M., Vogelsang, T.J., 2002a. Heteroskedasticity –autocorrelation robust testing using bandwidth equal to sample size," *Econometric Theory* 18, 1350-1366.

Kiefer, N. M., Vogelsang, T.J., 2002b. Heteroskedasticity –autocorrelation robust standard errors uinsing the Bartlett kernel without truncation," *Econometrica* 70, 2093-2095.

Kiefer, N. M., Vogelsang, T.J., 2005. A new asymptotic theory for heteroskedasticity-autocorrelation robust tests," *Econometric Theory* 21, 1130-1164.

Kiefer, N. M., Vogelsang, T.J., Bunzel, H., 2000. Simple robust testing of regression hypotheses, *Econometrica* 68, 695-714.

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

Newey, W.K, West, K.D., 1994. Automatic lag selection in covariance estimation. *Review of Economics Studies* 61: 631-654.

Pástor, L., 2000. Portfolio selection and asset pricing models. *Journal of Finance* 55, 179- 223.

Pástor, L., Stambaugh, R. F., 1999. Costs of equity capital and model mispricing. *Journal of Finance* 54, 67-121.

Phillips, P.C.B., Sun,Y., Jin, S., 2005a. Improved HAR inference using power kernels without truncation," Mimeographed, Yale University.

Phillips, P.C.B., Sun, Y., Jin, S., 2005b. Balancing size and power in nonparametric studentized testing with quadratic power kernels without truncation, Working paper, Department of Economics, UCSD.

Phillips, P.C.B., Sun, Y., Jin, S., 2006. Spectral density estimation and robust hypothesis testing using steep origin kernels without truncation," *International Economic Review* 21, 837-894.

Phillips, P.C.B., Sun, Y., Jin, S.,2007. Long-run variance estimation and robust regression using sharp origin kernels with no truncation," *Journal of Statistical Planning and Inference* 137, 985-1023.

Ray, S., Savin, N. E., 2008. The performance of heteroskedasticity and autocorrelation robust tests: a monte-carlo study with an application to the three-factor Fama-French asset-pricing model. *Journal of Applied Econometrics* 23: 91-109.

Stewart, K. G., 1997. Exact testing in multivariate regression. *Econometric Reviews* 16: 321-352.

Sun, Y., Phillips, P.C.B., Jin, S., 2008. Optimal bandwidth selection in heterokedasticity –autocorrelation robust testing. *Econometrica* 76: 175-194.

Table 1. *P*-values (%) for intercept *F*-tests for ten equally-weighted and value-weighted momentum-sorted portfolios

	Equally-Weighted Portfolios				Value-Weighted Portfolios			
	Three-factor Model		Four-factor Model		Three-factor Model		Four-factor Model	
Sub-Period	*F*-Statistic	*P*-value	*F*-Statistic	*P*-value	*F*-Statistic	*P*-value	*F*-Statistic	*P*-value
Five-Year								
1/65-12/69	3.33	0.24	3.62	0.13	1.62	12.98	1.60	13.61
1/70-12/74	1.76	9.48	1.77	9.46	1.34	23.92	1.28	27.11
1/75-12/79	2.48	1.78	2.37	2.34	1.14	35.41	0.86	57.63
1/80-12/84	3.71	0.10	3.58	0.14	1.37	22.20	3.68	0.11
1/85-12/89	2.38	2.25	2.43	2.02	5.72	0.00	6.18	0.00
1/90-12/94	2.42	2.07	1.81	8.56	2.69	1.09	2.96	0.60
1/95-12/99	3.36	0.23	2.30	2.75	2.14	3.91	2.28	2.87
1/00-12/04	1.95	6.21	2.68	1.13	1.11	37.24	2.24	3.17
Ten-Year								
1/65-12/74	3.84	0.02	2.82	0.38	2.72	0.51	1.92	5.07
1/75-12/84	5.10	0.00	4.05	0.01	1.78	7.20	2.25	1.98
1/85-12/94	2.56	0.82	2.36	1.46	4.79	0.00	4.81	0.00
1/95-12/04	3.45	0.06	2.75	0.47	1.32	22.89	1.38	19.98
Thirty-Year								
1/65-12/94	8.40	0.00	5.10	0.00	7.62	0.00	5.99	0.00
1/70-12/99	7.34	0.00	3.90	0.00	6.59	0.00	5.36	0.00
1/75-12/04	4.91	0.00	3.32	0.04	4.04	0.00	3.50	0.02
More Years								
1/65-12/99	9.59	0.00	5.34	0.00	8.24	0.00	6.22	0.00
1/70-12/04	5.63	0.00	3.19	0.06	4.62	0.00	3.36	0.03
1/65-12/04	7.03	0.00	4.16	0.00	5.65	0.00	3.95	0.00
Fama-French								
7/63-12/93	8.83	0.00	5.40	0.00	8.02	0.00	6.13	0.00

Table 2. *P*-values (%) for *F*-tests for ten equally-weighted and value-weighted momentum-sorted portfolios for shifted sub-periods

	Equally-Weighted Portfolios				Value-Weighted Portfolios			
	Three-factor Model		Four-factor Model		Three-factor Model		Four-factor Model	
Sub-Period	*F*-Statistic	*P*-value	*F*-Statistic	*P*-value	*F*-Statistic	*P*-value	*F*-Statistic	*P*-value
Five-Year								
1/67-12/71	1.81	8.55	1.52	16.43	2.94	0.60	2.76	0.93
1/72-12/76	2.02	5.26	2.04	5.09	1.12	36.48	0.93	51.50
1/77-12/81	4.12	0.04	3.50	0.17	2.34	2.48	3.37	0.23
1/82-12/86	2.40	2.14	2.71	1.07	2.02	5.21	2.69	1.10
1/87-12/91	1.12	36.86	1.05	41.74	4.06	0.05	4.20	0.04
1/92-12/96	5.82	0.00	5.40	0.00	1.06	41.36	2.30	2.76
1/97-12/01	1.93	6.37	1.62	13.14	2.07	4.60	1.74	10.02
1/02-12/06	1.98	5.78	3.64	0.13	1.02	43.97	1.87	7.38
Ten-Year								
1/67-12/76	3.27	0.10	2.54	0.88	2.47	1.05	1.74	8.04
1/77-12/86	4.63	0.00	3.70	0.03	2.94	0.27	3.76	0.02
1/87-12/96	3.57	0.04	2.89	0.32	4.06	0.01	4.98	0.00
1/97-12/06	2.40	1.32	2.09	3.13	0.88	55.47	1.18	31.20
Thirty-Year								
1/67-12/96	8.45	0.00	5.10	0.00	7.21	0.00	5.99	0.00
1/72-12/01	6.05	0.00	3.34	0.04	5.21	0.00	3.35	0.03
1/77-12/06	4.91	0.00	3.51	0.02	3.99	0.00	3.91	0.01
More Years								
1/67-12/01	6.66	0.00	3.73	0.01	6.58	0.00	4.45	0.00
1/72-12/06	6.20	0.00	3.70	0.01	4.14	0.00	2.99	0.12
1/67-12/06	6.78	0.00	4.05	0.00	5.25	0.00	3.83	0.01
Fama-French								
7/63-12/93	8.83	0.00	5.40	0.00	8.02	0.00	6.13	0.00

Table 3. *P*-values (%) for *F*-tests for equally-weighted and value-weighted momentum-sorted portfolios for the RR experiments

	Equally-Weighted Portfolios				Value-Weighted Portfolios			
	Three-factor Model		Four-factor Model		Three-factor Model		Four-factor Model	
Sub-Period	*F*-Statistic	*P*-value	*F*-Statistic	*P*-value	*F*-Statistic	*P*-value	*F*-Statistic	*P*-value
Five-Year								
1/65-12/69	3.33	0.58	3.62	0.37	1.62	16.52	1.60	16.19
1/70-12/74	1.76	12.58	1.77	12.34	1.34	27.00	1.28	29.92
1/75-12/79	2.48	3.46	2.37	3.65	1.14	38.85	0.86	59.38
1/80-12/84	3.71	0.37	3.58	0.48	1.37	25.60	3.68	0.37
1/85-12/89	2.38	4.12	2.43	3.44	5.72	0.00	6.18	0.00
1/90-12/94	2.42	3.75	1.81	11.65	2.69	2.41	2.96	1.26
1/95-12/99	3.36	0.47	2.30	4.01	2.14	5.75	2.28	4.07
1/00-12/04	1.95	9.82	2.68	1.79	1.11	44.55	2.24	5.78
Ten-Year								
1/65-12/74	3.84	0.03	2.82	0.49	2.72	0.71	1.92	5.91
1/75-12/84	5.10	0.02	4.05	0.04	1.78	8.45	2.25	2.04
1/85-12/94	2.56	1.08	2.36	1.73	4.79	0.00	4.81	0.00
1/95-12/04	3.45	0.14	2.75	0.64	1.32	26.11	1.38	22.63
Thirty-Year								
1/65-12/94	8.40	0.00	5.10	0.00	7.62	0.00	5.99	0.00
1/70-12/99	7.34	0.00	3.90	0.00	6.59	0.00	5.36	0.00
1/75-12/04	4.91	0.01	3.32	0.04	4.04	0.00	3.50	0.03
More Years								
1/65-12/99	9.59	0.00	5.34	0.00	8.24	0.00	6.22	0.00
1/70-12/04	5.63	0.00	3.19	0.02	4.62	0.00	3.36	0.04
1/65-12/04	7.03	0.00	4.16	0.00	5.65	0.00	3.95	0.00
Fama-French								
7/63-12/93	8.83	0.00	5.40	0.00	8.02	0.00	6.13	0.00

The resample-resample (RR) simulation experiment is described in detail in Section VII and Appendix A of the text.

Table 4. Asymptotic *P*-values (%) for HAR tests of the three-factor model with ten equally-weighted momentum-sorted portfolios

Sub-Period	Conventional: Bartlett		Fixed-b: Bartlett		Fixed-ρ: Parzen	
	W_M $M = 6$	*P*-value	W_b $b = 1$	*P*-value	W_ρ $\rho = 32$	*P*-value
Five-Year						
1/65-12/69	75.26	0.00	702.20	1.14	83.41	15.88
1/70-12/74	26.99	0.00	245.25	36.31	36.75	53.37
1/75-12/79	38.27	0.00	299.14	24.54	43.81	43.88
1/80-12/84	216.84	0.00	1570.50	0.00	303.15	0.42
1/85-12/89	49.81	0.00	416.51	10.21	59.21	29.35
1/90-12/94	63.38	0.00	653.61	1.70	88.76	13.97
1/95-12/99	230.68	0.00	1670.30	0.00	337.78	0.27
1/00-12/04	50.39	0.00	418.95	10.04	65.47	24.78
Ten-Year						
1/65-12/74	51.11	0.00	545.88	3.96	82.75	16.11
1/75-12/84	59.38	0.00	542.27	4.08	69.59	22.69
1/85-12/94	29.48	0.00	488.60	5.92	64.60	25.36
1/95-12/04	49.94	0.00	324.50	20.44	77.18	18.38
Thirty-Year						
1/65-12/94	88.09	0.00	1011.40	0.16	174.45	2.34
1/70-12/99	67.91	0.00	921.25	0.29	132.49	5.37
1/75-12/04	41.67	0.00	584.90	2.92	110.15	8.47
More Years						
1/65-12/99	97.68	0.00	1484.30	0.01	282.90	0.55
1/70-12/04	51.30	0.00	877.25	0.37	124.82	6.23
1/65-12/04	71.55	0.00	1442.3	0.01	254.16	0.74
Fama-French						
7/63-12/93	92.91	0.00	1205.18	0.00	192.16	0.02

The tabled asymptotic *P*-values for the fixed-b and fixed-ρ tests are computed by simulation using 10,000 replications of each experiment. The *P*-values for the conventional HAR test are calculated from the chi-square distribution with ten degrees of freedom.

Table 5. Simulated finite-sample *P*-values (%) for HAR tests of the three-factor model with equally-weighted momentum-sorted portfolios

	Conventional test: Bartlett				Fixed-*b*: Bartlett				Fixed-ρ : Parzen			
Sub-Period	$M = 6$				$b = \rho = 1$				$\rho = 32$			
Five-Year	NN	RR	NV	RB	NN	RR	NV	RB	NN	RR	NV	RB
1/65-12/69	6.88	8.93	5.17	10.11	4.24	5.75	2.90	7.04	22.47	24.28	19.66	29.84
1/70-12/74	54.91	58.58	47.63	66.74	46.69	51.20	40.28	60.72	60.11	63.63	56.37	76.60
1/75-12/79	36.64	39.01	34.42	40.80	38.39	41.21	35.76	41.07	53.82	56.19	51.72	61.33
1/80-12/84	0.02	0.17	0.09	0.94	0.09	0.26	0.12	1.14	0.81	1.39	0.60	3.94
1/85-12/89	20.36	24.89	14.31	20.99	18.76	23.16	12.52	18.59	36.47	40.40	31.77	42.18
1/90-12/94	9.16	11.87	6.56	25.03	3.71	5.67	2.79	15.86	16.91	19.20	15.30	39.36
1/95-12/99	0.02	0.13	0.03	0.56	0.04	0.27	0.04	0.74	0.60	1.12	0.40	3.24
1/00-12/04	23.29	28.6	22.17	33.87	21.76	27.3	20.74	31.70	34.84	39.28	33.63	48.45
Ten-Year												
1/65-12/74	1.27	1.36	0.74	3.16	4.86	5.59	4.28	7.73	17.59	18.79	16.92	24.58
1/75-12/84	0.63	0.79	0.66	1.33	5.81	6.96	5.23	7.21	25.23	26.28	24.98	29.91
1/85-12/94	11.87	14.27	11.55	19.42	7.54	8.72	6.93	12.52	27.23	29.10	27.07	36.39
1/95-12/04	1.36	2.34	1.59	4.19	24.25	27.41	25.08	28.59	20.71	22.70	20.93	28.58
Thirty-Year												
1/65-12/94	0.00	0.00	0.00	0.00	0.18	0.17	0.18	0.28	2.72	2.43	2.35	3.14
1/70-12/99	0.00	0.00	0.00	0.00	0.43	0.46	0.28	0.49	5.74	5.77	5.59	6.43
1/75-12/04	0.06	0.12	0.10	0.17	3.00	3.36	3.13	3.76	8.90	9.21	8.71	10.01
More Years												
1/65-12/99	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.45	0.50	0.53	0.51
1/70-12/04	0.00	0.00	0.01	0.02	0.41	0.42	0.44	0.65	6.29	6.39	6.09	7.40
1/65-12/04	0.00	0.00	0.00	0.00	0.10	0.10	0.02	0.02	0.71	0.74	0.71	0.74

The tabled asymptotic *P*-values are computed by simulation using 10,000 replications of each experiment. The simulation experiments are described in Section VII and Appendix A of the text.

Table 6. Simulated Power (%) of level-corrected 5 percent new HAR tests of the three-factor model for the RR experiments.

	Fixed-b					Fixed-ρ				
Sub-Period	Bartlett Kernel, $b = \rho = 1$					Parzen Kernel, $\rho = 32$, $b = 1$				
Five-Year	$c = 1$	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 1$	$c = 2$	$c = 3$	$c = 4$	$c = 5$
1/65-12/69	8.36	23.91	52.83	78.97	93.35	7.48	17.54	38.19	62.09	81.63
1/70-12/74	7.74	23.40	53.43	80.94	94.60	7.22	18.55	40.50	65.49	84.26
1/75-12/79	8.01	19.66	41.95	67.93	85.82	7.55	16.55	32.21	53.05	72.06
1/80-12/84	5.67	13.36	31.24	56.09	79.26	5.70	10.87	23.63	42.14	62.28
1/85-12/89	7.31	20.13	47.60	75.01	91.19	6.62	15.22	33.92	57.78	77.56
1/90-12/94	7.72	20.12	43.93	72.13	89.32	7.07	15.82	33.20	55.77	75.43
1/95-12/99	7.29	15.86	31.00	51.94	72.54	6.64	12.02	22.28	36.95	54.41
1/00-12/04	5.92	7.90	11.72	17.35	24.81	5.66	7.28	9.91	14.03	19.23

The tabled finite-sample powers are computed by simulation using 10,000 replications of each experiment. A unit increase in c translates into an increase of 10 basis points per month in the model intercept.

Table 7. Finite sample *P*-values for the conventional HAR test: sensitivity to structural breaks

Sub-Period	Three-factor Model			Four-factor Model		
	Statistic	*P*-Value		Statistic	*P*-Value	
		Asymptotic	Finite sample		Asymptotic	Finite sample
Panel A: Zero-Intercept Null Hypothesis is True						
1/65-12/94	75.2585	0.0000	0.0001	100.8660	0.0000	0.0000
1/70-12/99	26.9914	0.0026	0.0629	35.1390	0.0001	0.0045
1/75-12/04	41.6708	0.0000	0.0026	26.9130	0.0027	0.0646
Panel B: Zero-Intercept Null Hypothesis is False						
1/65-12/94	75.2585	0.0000	0.8563	100.8660	0.0000	0.0653
1/70-12/99	26.9914	0.0026	1.0000	35.1390	0.0001	0.9331
1/75-12/04	41.6708	0.0000	0.9988	26.9130	0.0027	0.9867

This table presents the finite sample *P*-values corresponding to the conventional Newey-West test statistic based on the Bartlett kernel with bandwidth parameter $M = 6$. The finite sample *P*-values are based on the RR simulation design described in Section VII and Appendix A. For each 30-year period we generate simulated returns for 10 momentum-based portfolios using the RR design described in the text. Panel A reports results for the case when the simulated portfolio returns conform to the zero-intercept null while Panel B reports results for the case when the portfolio returns are generated under the alternative of the non-zero intercept. Within each 30-year period, portfolio returns are generated for six five-year sub-periods based on the model parameter estimates for each individual sub-period. In each replication we compute the value of the test statistic based on the full 30-year simulated sample. The reported finite sample *P*-values are based on 10,000 replications each.

Table 8. Absolute values of ten equally-weighted momentum-sorted portfolio intercepts in basis points per month

Sub-Period	Portfolio										
	1 (low)	2	3	4	5	6	7	8	9	10 (high)	Median
A: Three-factor Model Intercepts											
1/65-12/69	77.97	19.78	3.46	18.97	7.51	7.82	20.76	17.76	43.48	53.46	19.37
1/70-12/74	41.37	0.54	3.69	14.32	8.89	7.97	4.62	10.42	5.20	41.41	8.43
1/75-12/79	17.17	3.67	16.49	2.51	9.96	11.52	21.17	30.52	41.10	54.30	16.83
1/80-12/84	126.67	32.21	10.74	18.56	4.33	10.77	28.90	33.60	72.14	81.57	30.55
1/85-12/89	119.09	40.11	19.48	2.71	1.04	0.26	1.68	8.52	16.25	21.02	12.39
1/90-12/94	7.37	12.69	14.84	5.46	16.30	21.42	33.10	37.76	64.01	78.32	18.86
1/95-12/99	48.75	28.64	20.80	9.80	9.12	2.42	18.19	36.11	36.17	99.62	24.72
1/00-12/04	165.29	58.46	40.67	60.21	54.97	71.11	65.25	78.86	110.35	105.52	68.18
B: Four-factor Model Intercepts											
1/65-12/69	33.47	13.12	35.62	4.85	20.71	3.05	15.79	7.82	26.14	19.54	17.66
1/70-12/74	24.59	14.16	12.80	19.80	9.26	9.43	9.61	21.80	23.57	16.77	15.46
1/75-12/79	39.30	43.04	53.04	22.01	25.67	15.10	14.40	20.23	25.60	16.37	23.81
1/80-12/84	85.41	6.86	24.70	5.09	15.46	14.76	24.69	9.86	40.87	30.53	20.07
1/85-12/89	95.23	26.47	8.33	9.75	2.20	0.59	1.54	1.91	5.38	8.34	6.85
1/90-12/94	53.93	22.52	7.05	12.67	25.63	23.25	21.22	21.90	52.40	45.92	22.88
1/95-12/99	49.95	22.34	19.36	42.82	0.10	5.68	21.94	31.17	20.82	75.55	22.14
1/00-12/04	123.96	39.00	27.44	51.30	48.11	67.82	64.06	80.88	116.53	120.38	65.94
C: Ratio of Model Intercepts											
1/65-12/69	2.33	1.51	0.10	3.91	0.36	2.56	1.31	2.27	1.66	2.74	1.97
1/70-12/74	1.68	0.04	0.29	0.72	0.96	0.85	0.48	0.48	0.22	2.47	0.60
1/75-12/79	0.44	0.09	0.31	0.11	0.39	0.76	1.47	1.51	1.61	3.32	0.60
1/80-12/84	1.48	4.70	0.43	3.65	0.28	0.73	1.17	3.41	1.77	2.67	1.62
1/85-12/89	1.25	1.52	2.34	0.28	0.47	0.44	1.09	4.47	3.02	2.52	1.38
1/90-12/94	0.14	0.56	2.10	0.43	0.64	0.92	1.56	1.72	1.22	1.71	1.07
1/95-12/99	0.98	1.28	1.07	0.23	87.92	0.43	0.83	1.16	1.74	1.32	1.12
1/00-12/04	1.33	1.50	1.48	1.17	1.14	1.05	1.02	0.97	0.95	0.88	1.10

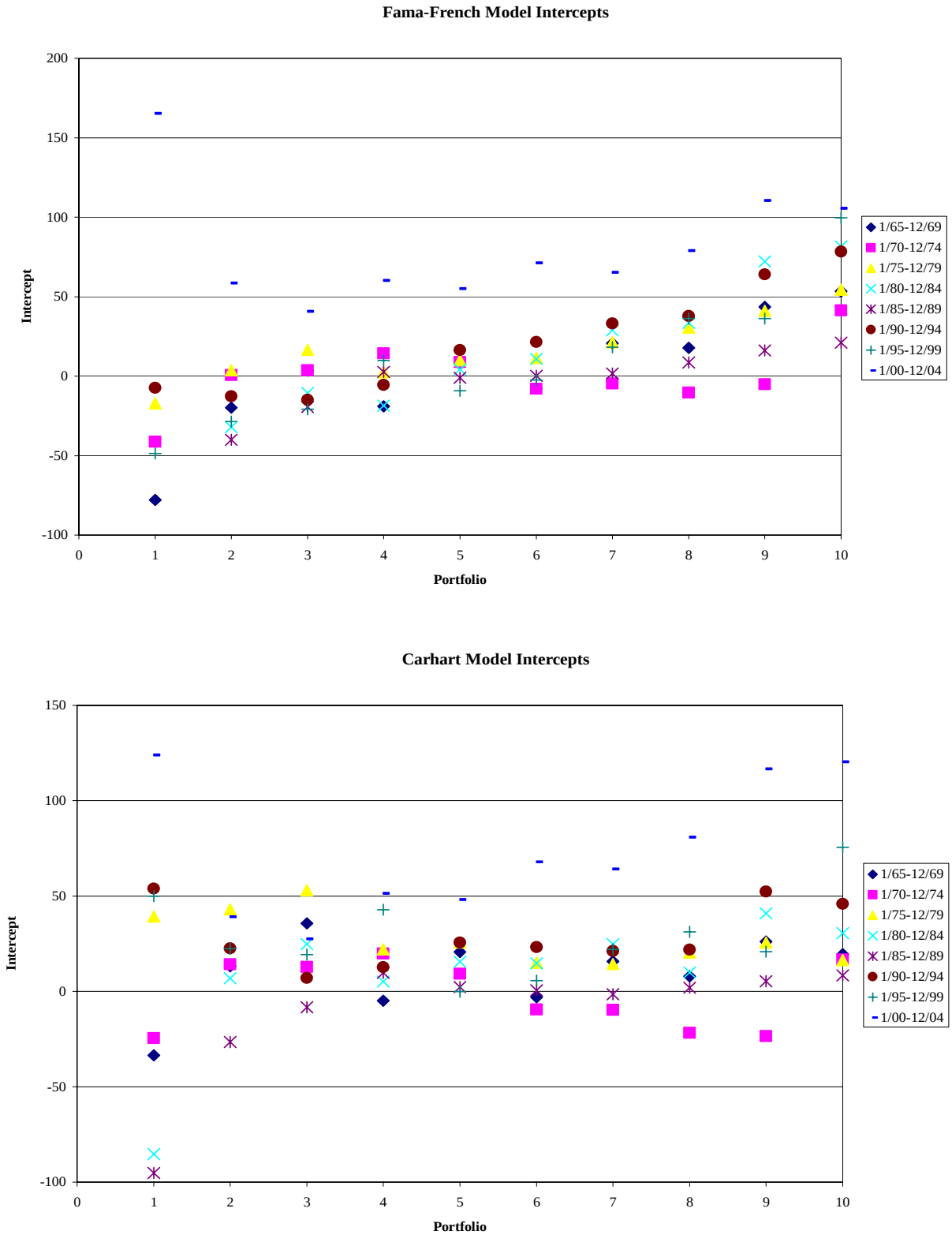

Figure 1. –Intercepts for the equally weighted momentum-sorted portfolios for five-year sub-periods in basis points per month. The top panel shows the Fama-French three-factor model intercepts while the bottom panel shows the intercepts for the Carhart four-factor model.