2002

Tests for Differing Sensitivity Among Asset Returns

Scott Gilbert
Southern Illinois University Carbondale

Petr Zemcik
Southern Illinois University Carbondale

Tests for Differing Sensitivity Among Asset Returns

Gilbert Scott* and Petr Zemčík†

November 2001

Abstract

When comparing assets, differences in sensitivity to economic variables are highly relevant, yet tests for such differences are absent from traditional studies. We therefore examine a range of tests for parameter differences across regression equations of asset returns. Simplistic approaches, which ignore conditional heteroskedasticity and/or serial correlation, suffer test distortions in simulations calibrated to asset returns, and more 'robust' methods also distort, for different reasons, but among these are the best methods. The tests suggest dramatic and time-varying differences between small and large firm sensitivity to market risk, the default premium and the term structure.

KEY WORDS: asset pricing; autocorrelation; cross-section restrictions; firm size; heteroskedasticity; simulation

*Corresponding author: Department of Economics, Mailcode 4515, Southern Illinois University at Carbondale, Carbondale, IL 62901-4515, Tel:(618) 453-5065, Fax: (618) 453-2717, E-mail: gilberts@siu.edu.

†We would like to thank John Cochrane, G. Wiiliam Schwert, Kris Jacobs and Richard C. Green for helpful suggestions and comments.

A main purpose of asset pricing models is to explain the cross-sectional variation in returns among stocks and other risky assets. In the tradition of the Sharpe (1964) and Lintner (1965) capital asset pricing model (CAPM) and its generalizations via the Merton (1973) and Breeden (1979) intertemporal equilibrium models and the Ross (1976) arbitrage pricing theory (APT), cross-sectional variation arises due to differing sensitivities to some economic variables, either some explicit source(s) of risk or some underlying state variable(s). The hypotheses of equal versus differing sensitivities have so far escaped formal testing in the literature on traditional asset pricing models, a missed opportunity with the potential to substantively alter the understanding of financial data.[1]

The existence of differing sensitivities among assets is a tacit and untested tenant of traditional empirical methods. In particular, for the highly influential Fama and MacBeth (1973) method[2] of testing the impact of sensitivity differences on mean values, if sensitivity exists but is the same across assets then the Fama-MacBeth test is unuseful since there are no sensitivity differences to impact mean values. Clearly there will be sample variation in estimated sensitivities (betas), but the notion of equality among true sensitivities, of various asset returns to economic variables, is entirely plausible, particularly for popular collections of assets such as stocks sorted by firm size. For the interpretation of the existing empirical literature, it is therefore very important to gauge the extent of sensitivity differences across assets.

The standard modelling framework for defining and discussing sensitivities is the regression of excess returns on covariates, e.g. risk factors and/or state variables.[3] With sensitivities defined via the slope parameters in the model, differences in sensitivity are clearly in the domain of formal, testable hypotheses. In the early research on asset pricing, emphasis was placed on models with zero intercepts, but the poor performance of such models has lead to inclusion of non-zero intercepts.[4] In this setting, the notion of sensitivity is linked to the intercept's value, this being the expected return conditional on zero values for the covariates, and for model interpretation it is important to compare both intercepts ('fixed effects') and slopes across assets.

The present work studies the problem of testing equality of parameters across assets in regressions of returns on economic variables. A related but distinct problem, which has attracted great attention, is the test for zero intercepts (Gibbons, Ross and Shanken (1989), Fama and French (1993, 1996), Cochrane (2001)), and the classical F test is standard for this problem. The F test is simplistic, as it ignores possible residual heteroskedasticity and autocorrelation, and it suffers distortions when such effects are present, but more 'robust' methods (discussed below) also suffer distortions, for other reasons, and the prevailing view is that the simplistic approach is best for testing the zero intercepts hypothesis. Such a view is specific to the hypothesis in question, and it is unknown whether any simplistic or 'robust' methods are reliable for testing equality of intercept or slope parameters across equations, when applied to typical financial data.

We examine the performance of classical F tests, as well as tests which are heteroskedasticity and autocorrelation consistent (HAC). The HAC methods, due to Newey and West (1987, 1994), Andrews (1991) and Andrews and Monahan (1992), and further studied by den Haan and Levin (1997), build on earlier work by White (1980) on heteroskedasticity-robust tests. Previous successes in financial applications of heteroskedasticity-robust test methods include MacKinlay and Richardson (1991), Ferson and Foerster (1994), Ferson and Korajczyk (1995) and He, Kan, Ng and Zhang (1996). The presence of autocorrelation in asset returns is well-documented[5], and even in cases where assets returns have little autocorrelation themselves, when multiplied by covariates the resulting series may be more strongly autocorrelated, and this particular phenomenon is important for testing restrictions on the regression parameters.

We simulate the behavior of the various tests for models calibrated to monthly financial data for the period January 1959 - December 1999. For the F test, we observe noticeable distortions due heteroskedasticity and autocorrelation built into the model, while for the 'robust' tests distortions arise due to the test rule, which relies on the asymptotic (chi square) distribution of test statistics, and which generally differs from rules based on the exact (but unknown) finite-sample distribution. MacKinnon and White (1985) acknowledge the distortions problem for heteroscedasticity-robust tests, proposing corrective methods, and Ferson and Foerster (1994) examine the importance of distortions for heteroskedasticity-robust tests of some financial models. Den Haan and Levin (1996) report on test distortions for a variety of HAC tests in a single equation context[6] and Cochrane (2001, Ch. 15) studies the zero intercepts hypothesis in the multi-equation context.

Our Monte Carlo results, for tests of equal parameters across equations in the regression model, suggest that the F test and the 'robust' HAC tests can work well when applied to models with relatively few equations, but both suffer major distortions when applied to larger models. The HAC tests include score and Wald type tests, with or without a parametric pre-whitening adjustment for serial correlation. In simulation the HAC score tests distort less than the F test and HAC Wald tests, and simple pre-whitening is as good or better than other methods of handling serial correlation. We find that (i) there are test methods that work well, (ii) the methods that work particulary well are very different in design than the traditional F test, and (iii) such tests are only appropriate when applied to sufficiently parsimonious models, models which exclude some interesting cases such as regressions of monthly post-war returns sorted by size decile (ten equations) and those which include many covariates.

We use the afore-mentioned methods to test for differences in the economic sensitivity of small and large firm excess returns. The potential for such differences has long been recognized[7], and as covariates we include standard economic risk factors (market return and consumption growth) as well as other standard economic variables (default premium, term structure, money growth, etc.) related to the economy. Some researchers use instead covariates consisting of size and book-to-market related portfolios (see Fama and French (1993) and Chan and Chen (1991)), or other statistical factors (Lehmann and Modest (1988) and Connor and Korajczyk (1988)), but since statistical and economic covariates appear to have similar predictive power with respect to stock returns[8], we use just the latter.

We find numerous dramatic differences in the economic sensitivity of returns on small and large firms. Stability over time is a concern in such regressions[9], and we find that differing sensitivities to the market return appeared quite important for small and large firms prior to 1980, while more recently this effect is more subdued and the effect of differing sensitivity to the default premium and term structure are more pronounced. The time-varying nature of such differences is very important for the interpretation of existing literature that reports on sensitivities of small and large firms.[10]

I. Model

For a collection of n risky assets, each earning a return during periods $t = 1, 2, ..., T$, let r_{it} denote the excess return to the i-th asset. The linear regression model of asset returns takes the form:

$$r_{it} = \alpha_i^* + \beta_i^* x_t + \varepsilon_{it}, \quad i = 1, ..., n, \quad t = 1, ..., T, \tag{1}$$

where x_t is a $K \times 1$ vector of covariates (risk factors and/or state variables), β_i^* is the true value of the i-th $1 \times K$ vector of slopes ('betas'), α_i^* is the true value of the i-th intercept, and the errors ε_{it} have conditional expectation $E[\varepsilon_{it}|x_t] = 0$. In this model, β_{ik} is the expected increase in the excess return r_{it}, given a one unit increase in the covariate x_{tk}, while $\alpha_i = E[r_{it}|x_t = (0, ..., 0)']$, e.g. α_i is the expected excess return when each covariate equals 0. The model is linear in the parameters α and β, but x_t itself may be non-linear in some underlying variables which themselves may be non-contemporaneous with r_t, hence the model may be both non-linear and dynamic in some underlying variables.[11] Estimation of (1), using time series data, is the first step in the Fama and MacBeth (1973) empirical method, and the possibilities in the model (1) itself have attracted wide interest.[12]

In the Sharpe-Lintner CAPM version of the model, x is the excess return on the market portfolio, and the betas measure sensitivity to market risk. Other candidates for x include consumption growth, as in the Breeden (1979) consumption-based CAPM, and other variables, possibly instruments for some latent factors.[13] Such models offer an explanation of differences among average returns for various assets, provided that sensitivities differ among assets. Informal comparisons of betas across assets are widespread in the industry, facilitated by point and interval estimates, but formal comparison via hypothesis tests has not received attention in the literature.

The hypotheses of present interest take the form of linear restrictions on β and/or α. To concisely express such hypotheses for the purpose of testing, for each equation i we denote by $\theta_{[i]}$ the $(K+1)\times 1$ vector $(\alpha_i, \beta_{i1}, ..., \beta_{iK})'$, and let θ be the $n(K+1)\times 1$ vector $(\theta'_{[1]}, \theta'_{[2]}, ..., \theta'_{[n]})'$. With 0_p the column vector consisting of p entries each equal to 0, and with A some user-specified $p \times n(K+1)$ matrix, each linear restriction on the model parameters takes the form:

H_0: $A\,\theta^* = 0_p$.

As a familiar example, when testing the zero intercepts or zero betas null hypothesis the restriction is of the form:

$$C\,\theta^*_{[i]} = 0_q, \quad i = 1, ..., n, \tag{2}$$

for some $q\times(K+1)$ matrix C and some number q of restrictions, in which case the appropriate form of the matrix A in H_0 is:

$$A = I_n \otimes C, \tag{3}$$

where I_n is the $n \times n$ identity matrix, and $\otimes$ is the Kronecker product operator. Hypothesis tests for asset return regressions have typically targeted restrictions of the form (2)-(3).

We are interested in testing for differences in slopes and/or intercepts across equations, and the relevant restriction is of the form:

$$D\,\theta^*_{[i]} = D\,\theta^*_{[j]}, \qquad i, j = 1, ..., n, \tag{4}$$

for some $r \times (K + 1)$ matrix D, some number r of restrictions, and all assets i, j. The appropriate form of the matrix A in H_0 is then:

$$A = J_n \otimes D, \tag{5}$$

where J_n is the $(n-1) \times n$ matrix with entries $J_{ni1} = 1$, $J_{n,i,i+1} = -1$, and $J_{nij} = 0$ otherwise. We are primarily interested in testing restrictions of the form (4)-(5).

II. Tests

In this section we describe methods of hypothesis testing based on HAC Wald and HAC score tests which we study later as alternatives to the F test.

A. HAC Statistics

To conduct HAC Wald tests we let $\hat{\theta}$ denote the ordinary least squares (OLS) estimator, and we let $\hat{V}_{\hat{\theta}}$ denote an estimator, further described below, of the variance-covariance matrix for $\hat{\theta}$. For each given choice of $\hat{V}_{\hat{\theta}}$, the test statistic is:

$$W \;=\; \hat{\theta}' A' \left(A \hat{V}_{\hat{\theta}} A' \right)^{-1} A \, \hat{\theta}. \tag{6}$$

The statistic W measures the distance (in R^p, with norm $||v|| = v'(A\hat{V}_{\hat{\theta}}A')^{-1}\, v$) between the vector $A\,\hat{\theta}$ and the value 0_p hypothesized under H_0, hence larger values of W suggest larger departures of the data from H_0.

To conduct HAC score tests, for any parameter values α_i and β_i define the regression residuals for the model (1):

$$e_{it} = r_{it} - \alpha_i - \beta_i \, x_t, \quad i = 1, ..., n, \quad t = 1, \ldots, T.$$

The relevant sample moments comprise the $n(K+1) \times 1$ vector $m(\theta)$, given by:

$$m(\theta) = \frac{1}{T} \sum_{t=1}^{T} z_t \otimes e_t,$$

where z_t is the $(K+1) \times 1$ vector $(1, x_t')'$. Denoting by $\hat{V}_{m(\theta)}$ an estimator (specified below) of the variance-covariance matrix of $m(\theta)$, the score test statistic is:

$$S = \min_{\theta \in H_0} \;\; m(\theta)' \, \hat{V}_m^{-1} \, m(\theta). \tag{7}$$

The score test measures the distance (in $R^{n(K+1)}$, with the norm $||v|| = v'\hat{V}_m^{-1}\, v$) between the vector $m(\theta)$ of sample moments and the value $0_{n(K+1)}$ hypothesized under H_0, hence larger values of S suggest larger departures from H_0.

For econometric testing of linear restrictions H_0 on linear regression systems, score tests are seldom used while F and Wald tests are popular, whereas for nonlinear problems the score test is common, as in Hansen (1982) and Ferson and Foerster (1994). Yet our simulations (reported later) suggest a useful role for HAC score tests of parameter equality across equations in linear systems.

B. Computation

To compute the HAC test statistics we apply formulas (5), (6) and (7), with various specifications for the covariance matrix estimators $\hat{V}_{\hat{\theta}}$ and $\hat{V}_m$. For the HAC Wald and score tests, we use a variety of HAC covariance estimators. Among these are the Bartlett kernel and the data-dependent Newey and West (1994) bandwidth, with and without pre-whitening (denoted NW and NW-P, respectively), the quadratic spectral kernel with the Andrews (1991) data-dependent bandwidth (without prewhitening, denoted A), and the Andrews and Monahan (1992) method (denoted AM) with pre-whitening. Further, we include the simple pre-whitening method (denoted VARHAC) with parametric, vector autoregressive, adjustment for serial correlation, studied by den Haan and Levin (1996, 1997). Finally, for comparison purposes we include the White covariance estimator (WH) which is robust to heteroskedasticity but not serial correlation. Since the technical details of covariance estimators are neatly summarized in Campbell, et al. (1997) and Cushing and McGarvey (1999), we omit them for brevity.

To carry out the minimization (7) required for the score statistic S, we use the GMM (simultaneous-iteration) routine, which at each iteration stage simultaneously solves for updated parameter and covariance matrix estimates, as in Hansen, Heaton and Yaron (1996).[14]

C. Decision Rule

In the presence of conditional heteroskedasticity or serial correlation, each of the aforementioned HAC test statistics have unknown distributions in finite samples[15] and we take a standard approach[16] which is to base test decisions on the asymptotic chi square distribution (with p degrees of freedom) of such tests. For the HAC Wald tests, the asymptotic chi square distribution follows from asymptotic normality of OLS regression estimators and consistency of HAC covariance estimators[17], while for the HAC score tests the asymptotic distribution can also be shown chi square by invoking suitable assumptions, as we now briefly discuss.

To further justify the presumed asymptotic properties of the ‘robust’ HAC tests, we assume that the covariance estimators $\hat{V}_{\hat{\theta}}$ and $\hat{V}_m$, when multiplied by the number of time periods T, converge as follows:

$$T\,\hat{V}_{\hat{\theta}} \xrightarrow{p} \Omega_\theta, \quad T\,\hat{V}_m \xrightarrow{p} \Omega_m, \tag{8}$$

where $\xrightarrow{p}$ denotes convergence in probability, in which case Ω_θ and Ω_m are the large-T limits of T times the variance-covariance matrix for $\hat{\theta}$, and for $m(\theta^*)$, respectively. Simplified versions S^* and W^* of the Wald and score statistics are then:

$$W^* \;=\; \hat{\theta}' A' \left(A \frac{\Omega_\theta}{T} A' \right)^{-1} A\,\hat{\theta}, \tag{9}$$

and

$$S^* \;=\; \min_{\theta \in H_0} \; m(\theta)' \left(\frac{\Omega_m}{T} \right)^{-1} m(\theta), \tag{10}$$

in which case we assume that:

$$W = W^* + o_p(1), \quad S = S^* + o_p(1), \tag{11}$$

where $o_p(1)$ denotes a term converging to 0 in probability.

Under the null hypothesis and suitable regularity conditions[18], the statistics W^* and S^* are distributed asymptotically as chi square variables with p degrees of freedom[19], and hence W and S also have this property when (11) holds. Using these asymptotic distributions, the decision rule for testing H_0 is to reject if the test statistic exceeds the relevant critical value from the chi square distribution.

Despite asymptotic equivalence of the chi square tests S and W, in finite samples they can behave very differently. In the classical single-equation regression model, with regression parameters and covariance parameters estimated via maximum likelihood, the inequality of score (Lagrange multiplier) and Wald test statistics (Engle (1984)) is:

$$S \leq W, \tag{12}$$

and in practice it is possible that S is far smaller than W. Hence, by using the same (chi square) decision rule for both statistics it is possible to reach different conclusions from the two tests. The problem arises due to test *distortions*, caused by use of inaccurate chi square approximations to the true sampling distribution. For 'robust' HAC Wald and score tests of equality between parameters in a regression system, the bound (12) generally fails, and test distortions are more complex. To describe the magnitude of this problem, in Section 5 we report the results of simulations for testing differences in sensitivity among stocks.

III. Data

We examine excess returns on stocks of firms ranked by capitalization. To compute these returns we use CRSP NYSE Cap-Based Portfolio Indices, monthly time series based on portfolios rebalanced quarterly. Frequently, cross-sectional differences among stock returns are investigated using decile indices; however, to limit the number of dependent variables (and the potential for test distortions, reported later), we use one return for portfolios combining Deciles 1 through 5, and a second return for deciles 6 through 10, where the largest companies are in Decile 1 portfolio and the smallest in portfolio 10. These returns are produced by CRSP, and we calculate excess returns using the 30-Day Treasury Bill return, also provided by CRSP. We denote the excess returns as r_{LARGE} and r_{SMALL}, respectively.

Summary statistics, for monthly excess returns in the period 1959:02 - 1999:12, are in Table I. The starting period of the data series is determined by availability of the consumption series (defined below). We further split the sample in two sub-samples, 1959:02-1979:12 and

1980:01-1999:12, enabling us to examine stability of regression parameters. A comparison of the sample means for excess returns, for the sample period from 1959 to 1979, reveals that the excess return on the large-cap portfolio (3.02% annually) is far less then the excess return on the small-cap portfolio (8.08% annually), but the gap in average excess returns changes sign in the second sub-sample (9.98% for the large-caps vs. 8.34% for the small-caps, respectively), consistent with Fama and French (1993) and Horowitz, Loughran and Savin (2000). In all considered sample periods, the excess return on small caps tends to be more volatile, in accord with Malkiel and Xu (1997).

As covariates in the model[20], we choose ones likely to affect the stochastic discount rate and/or the expected stream of cash flows. We follow Chen, Roll and Ross (1986) and use data on the stock market, bond market, the business cycle and inflation, and we augment the dataset by the growth of monetary base to address the issue of asymmetric reaction of firms of different capitalization to restrictive monetary policy.[21]

To describe the stock market we use the CRSP NYSE value-weighted index. Again, we use returns in excess of the 30-Day Treasury Bill, denoting the results by r_{VW}. The correlation with the large-cap return is close to one (see Table II), and since the large-cap firms account for most of the market value, this is not surprising.[22]

We consider two bond market variables. The effect of unanticipated changes in bond risk premia is measured by the difference (denoted r_{DEF}) between interest rates on the low grade bonds and long-term government securities. The low grade bond interest rate is measured by the Seasoned Baa Corporate Bond Yield, collected by Moody's Investors Service and available at the St. Louis Federal Reserve bank's website. The long-term government bond return-to-maturity is from the 5-year Treasury Bonds, also obtained from the St. Louis Fed website. To describe the term structure we use the difference between the one-period holding return on the 5-year Treasury Bond, collected by CRSP, and the first lag of the return on a 30-Day Treasury Bill. This term premium (r_{TERM}) proxies for the influence of changes in the term structure on equity returns.

As measures of real economic activity, we include the growth rates of industrial production (g_{IP}) and real per capita consumption (g_{CONS}). We obtain industrial production data (Market Groups, series b50001, seasonally adjusted) from the Federal Reserve Board's website, and we obtain consumption data (series PCEND, non-durables, series PCES, services, POP, population, series CPIAUCSL, Consumer Price Index For All Urban Consumers, All Items 1982-84=100, all series seasonally adjusted), from the St. Louis Fed's website.

To get a one-period forecast of the inflation rate, we run a regression of the inflation rate, measured by the above consumer price index, on a constant, its lagged value, the lagged value of a Treasury Bill rate and a moving average term.[23] The unexpected inflation (π_{UI}), in the style of Chen et al. (1986), is defined as the difference between actual inflation and forecasted inflation.[24] For money growth, we use the growth rate of the seasonally adjusted monetary base (g_{MON}), obtained from the St. Louis Fed's website (series AMBSL, seasonally adjusted).

IV. Simulation

We use computer simulation, based on a calibrated model of asset returns, to assess test performance. Of interest are rejection rates under the null hypothesis and under the alternative. If the nominal distribution (F distribution for the F test, chi square distribution for the HAC tests) is an accurate approximation then the tests should reject under H_0 at a rate near the theoretical test size; otherwise, the tests will exhibit noticeable distortions.

To set up the simulation, we define a first-order vector autoregressive (VAR) process for covariates x_t:

$$x_t = c + \Phi \, x_{t-1} + u_t, \tag{13}$$

where c is a $K \times 1$ vector of constants, Φ is an $K \times K$ matrix of coefficients, and u_t is a $K \times 1$ vector of random variables which are independent over time and normally distributed with zero mean and cross-sectional variance-covariance matrix Λ.

To see what range of values might be realistic for the parameters of the x_t process, we estimate (13) for $K = 4$ by OLS using $x_t = \{r_{VWNY}, r_{TERM}, g_{CONS}, g_{MON}\}$. Estimates of elements the matrix Φ range from -0.25 to 0.32, and for our simulation we set $\Phi_{ij} = 0.10$ for $i = j$ and $\Phi_{ij} = 0$ for $i \neq j$. Estimates of the constant term tend to be small relative to elements of Φ, and we set $c = 0.002$ in our simulation exercise. The diagonal elements of the estimated residual covariance matrix $\hat{\Lambda}$ are typically of order 0.0001, and the off-diagonal elements are typically much smaller, hence we let Λ be a diagonal matrix with each diagonal entry equal to 0.0001.

For the regression errors ε_{it} in (1), we posit a dynamic model with serial correlation and generalized autoregressive conditional heteroskedasticity (GARCH), as follows:

$$\varepsilon_{it} = \psi_1 \varepsilon_{i,t-1} + \psi_2 \sqrt{1 + \psi_3\, \varepsilon_{i,t-1}^2}\, \eta_{it}, \qquad i = 1, \ldots, n,$$

with η standard normal noise. Parameter ψ_1 specifies the autocorrelation, and parameters ψ_2 and ψ_3 specify the conditional heteroskedasticity. We choose ψ so that the autocorrelation of the error term ε_{it}, as well as its variance relative to that of x's, corresponds to what we observe in historical data series, with r_{1t} and r_{2t} excess returns on portfolios of small and large firms, respectively. In this case, we set $\psi_1 = .1$, $\psi_2 = .003$ and $\psi_3 = .2$. The cross-sectional empirical covariance of η_{it} is sometimes positive and sometimes negative, and we specify the population covariance between η_{1t} and η_{2t} to be 0.

To get a sense for the behavior of the F test and 'robust' HAC tests, we first generate results for the case $n = 2$, with $K = 2$ and, alternately, $K = 4$, using 500 simulated time series for r_{it}, $i = 1, 2$, with 240 and 492 observations, corresponding to 20 and 41 years of our historical monthly data, respectively. We conduct a Monte-Carlo experiment based on a calibrated model, rather than a bootstrap method as in Ferson and Foerster (1994), for two reasons: First, the calibrated model allows us to identify the source of test success or failure; second, the regression errors have posited dynamics which would not be replicated by standard bootstrap sampling.[25] We record the number of rejections of the null hypothesis using the chi square critical values at the 5% level of significance.

Table III reports rejection rates under the null hypothesis of cross-equation equality for all regression parameters, e.g. the case where the restriction defining matrix D in Section 2 equals the $p \times p$ identity matrix. Results are of similar nature when testing equality of intercepts only, or slopes only (simulation results available on request). We calibrate all β values to equal to 1, and all α values to equal 0. Our simulations show a serious tendency for distortion in most but not all tests. Specifically, the F test and the HAC Wald tests over-reject[26], and the two of the score tests (Newey-West and Newey-West with pre-whitening) under-reject the null hypothesis. On the other hand, three of the score tests (VARHAC, Andrews and Andrews-Monahan) show minimal distortion, and of these three the VARHAC test is by far the simplest to compute and interpret. We have examined the score VARHAC test in numerous other simulations exercises: For $n = 2$, we gradually increase the number of covariates K by two up to $K = 8$, and rejection rates fall toward 0.03 and 0.04 for sample sizes 240 and 492. Since many studies consider decile indices, we also look at $n = 10$ and increase the number of covariates from two to eight, in which case the rejection rates for the VARHAC score tests are respectively 0.01 and 0.02 for the two sample sizes. For no other test method do we find less distortion than for the VARHAC score test, and our results suggest that a researcher attempting to investigate the relationship between various variables and asset returns is 'safer' when the number of assets is smaller since the asymptotic and finite sample distributions of the test statistic are closer.

To describe performance under the alternative hypothesis, we generate simulated time series for excess returns via:

$$\begin{aligned}
r_{1t} &= x_{1t} + x_{2t} + \ldots + x_{Kt} + \varepsilon_{1t}, \\
r_{2t} &= x_{1t} + x_{2t} + \ldots + x_{\frac{K}{2},t} + (1 + \tfrac{0.2}{K})\big(x_{(\frac{K}{2}+1),t} + x_{(\frac{K}{2}+2),t} + \ldots + x_{Kt}\big) + \varepsilon_{2t}.
\end{aligned}$$

Table IV reports rejection rates under the alternative hypothesis for $K = 2$ and $K = 4$, with relatively high rejection rates for the F test, and with higher rejection rates for the HAC Wald test than for the corresponding HAC score test. Among the HAC score tests, the Andrews, Andrews-Monahan and VARHAC methods reject more frequently than than the others. These results describe the frequency with which an economist would correctly reject the null hypothesis, using the nominal (F or chi square) distribution of the relevant statistic. A related, but different, issue is the frequency of correct rejection for an economist who knows and uses the exact test distribution. The latter power calculations are not interesting here because the economist does not know the exact distribution, and it is impossible to concisely report on this distribution in a way that would be broadly useful for asset return regression. We have nevertheless done such power calculations, with the same rankings described above, for the various tests.

Overall, the simulations reveal some serious problems with the F test and with the 'robust' HAC Wald tests, in terms of over-rejection under the null hypothesis, whereas three of the HAC score tests avoided serious distortions and were also best among score tests under the alternative hypothesis. Among these favored three we recommend the VARHAC score test, with it's simple, parametric pre-whitening approach to serial correlation adjustment. For the range of sample sizes under study, the VARHAC score test performance under null and alternative hypotheses suggests that for a small number of assets, $n = 2$, we can have as many as 8 covariates and still avoid major test distortions. In cases of $n = 10$ asset returns, the number of covariates in a restricted econometric model should be kept small, perhaps no more than 4 or 6. In cases where larger models and a greater number of restrictions are desired, larger sample sizes (weekly rather than monthly data, for example) may be necessary for satisfactory results.

V. Empirical results

Having scrutinized a variety of test methods, we turn now to the problem of testing for differences in sensitivity among firms of different size. We conduct tests of equality of parameters, equality of slopes for a specified covariate, equality of intercepts, and intercepts being equal to zero. The tests are formulated by defining matrices C and D in Section 2 accordingly. As our simulations warn against the use of overly large models, we take the rather unconventional empirical approach of modelling just two excess returns, for broad portfolios of small and large firms, defined earlier. To save space we report only the score-type tests with parametric VARHAC adjustment for residual serial correlation and heteroskedasticity, as these tests showed relatively little distortion in simulation, and are generally in agreement with the other tests for the models we analyze.

We report first on univariate models of excess returns, each with a covariate given by one of the economic variables defined earlier. Table V gives results for the full monthly sample (1959-1999) and two sub-samples. The first covariate in Table V is the market excess return. When testing the joint equality of intercepts and slopes for both assets, at the 5% significance level we reject in the first sub-sample but not in the second sub-sample, consistent with descriptive evidence in Table I. The test of equality of market betas suggests significant difference in risk exposure, for large and small firms, in the first but not second sub-sample, while the test of equal intercepts fails to find a significant difference during either sub-sample. We also test for whether the intercepts are each 0, with mixed results (rejection on the second sub-sample but not the first one). The test of zero intercepts for the market model is essentially a version of the standard F-test commonly applied in testing the CAPM.[27] The performance of the market model could be viewed as evidence against the CAPM since only p-values for the period from 1959 to 1979 are larger than standard significance values. Tests of equal parameters for covariates other than the market reject the null hypothesis for consumption growth and inflation in the full sample. For consumption, the null is rejected in the first sub-sample, while for the default premium and inflation, it is rejected in the second sub-sample, but not the first. Again, the apparent source of this parameter heterogeneity is in betas, in the full sample and each sub-sample.[28] Tests for a zero intercept tend to reject the null during the second sub-sample, but not the first sub-sample.

We next examine bivariate models, with covariates given by the market return and one of the remaining five economic variables. Table VI reports results, and as in the univariate models we see pervasive temporal instability in the sensitivity of excess returns. For each version of the model, tests reject parameter equality, for small and large firms, in the whole sample and first sub-sample. The apparent source of parameter heterogeneity is mostly difference in betas rather than alphas. In the first sub-sample, differences between firms of different sizes are a result of different market betas. Sensitivity to the second covariate shows in each case no significant differences in the first sub-sample, but frequently shows such differences in the second sub-sample (when the second factor is the default premium, consumption growth, or unexpected inflation) or overall (term premium). Tests for zero intercepts reject the null occasionally in the whole sample and always in the second sub-sample, but not in the first sub-sample.[29]

To investigate whether test results are robust to the number of explanatory variables, we examine the model in which all seven covariates are included. Table VII reports parameter estimates and their standard errors.[30] To further describe the model we report in Table VIII residual diagnostic tests. As indicated, there is strong evidence of both residual heteroskedasticity and autocorrelation, in which case our use of HAC test methods is highly appropriate.[31] Finally, Table IX reports results of tests of cross-section restrictions.

Tables VII and IX reveal a pattern consistent with our results for models of a smallerscale (see Tables V and VI). The market betas change through time, significantly different in the first period but both close to unity in the second period.[32] While the betas on the market variable become indistinguishable from the statistical point of view, the betas for all the other variables with the exception of money supply differ for firms of different size, with the difference being more pronounced in the second period. For instance, the absolute difference between slopes from the first sub-sample is 2.679 for the default premium and 0.12 for the term premium (see Table VI), respectively. This difference increases substantially in the second period, to 13.871 for the default premium and 0.348 for the term premium, respectively.[33] Thus, for these two variables we can now reject the null hypothesis of equal slopes across firms of different size, and we get closer to rejection for the other covariates, again with the exception of the money supply.[34]

Overall, our empirical analysis suggests the presence of interesting and often dramatic differences in the sensitivity of small and large firms to economic variables, in spite of the fact that time-averaged excess returns on small and large firms' portfolios have become closer (see Table I). While the gap in market betas for small and large firms has narrowed, differences in sensitivities to other economic variables - notably to the default and term premiums - have grown over time.[35]

VI. Conclusion

This study brings into the foreground an assumption which is fundamental to asset pricing theory, but has escaped formal testing in traditional empirical work. The going assumption is that cross-sectional variation in asset returns is due to cross-sectional differences in the sensitivity of assets to economic variables. Differences in sensitivity, of great interest theoretically, are also critical for the success of standard empirical methods such as the Fama and MacBeth (1973) procedure.

A variety of methods can conceivably deliver satisfactory tests for differing sensitivities, but a modern understanding of asset return data brings special testing challenges. In particular, evidence of residual heteroskedasticity and serial correlation calls for an attempt of 'robust' methods more complex than the classical F test. A simulation study calibrated to asset returns shows distortions in the F test due to the presence of heteroskedasticity and autocorrelation, and also shows distortions in 'robust' methods due to failure of the chi square decision rule. Overall, we recommend the robust score test with a simple pre-whitening adjustment for serial correlation.

In application to stocks sorted by firm size, using a variety of models we frequently find significant differences in sensitivity to economic variables. The extent of such differences varies substantially over time, for each variable studied. The gap in market betas has narrowed, while the differences in sensitivities to other variables, such as the default and the term structure premiums, the growth rates of consumption and industrial production, and the unexpected component of inflation, have become more pronounced.

REFERENCES

Andrews, Donald W. K., 1991, Heteroskedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica* 59, 817-858.

Andrews, Donald W. K. and J. Christopher Monahan, 1992, An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator, *Econometrica* 60, 953-966.

Barber, Brad M. and John D. Lyon, 1997, Firm size, book-to-market ratio, and security returns: A holdout sample of financial firms, *The Journal of Finance* 52, 875-883.

Breeden, Douglas, 1979, An intertemporal asset pricing model with stochastic consumption and investment opportunities, *Journal of Financial Economics* 7, 256-296.

Brennan, Michael J., Tarun Chordia and Avanidhar Subrahmanyam, 1998, Alternative factor specifications, security characteristics, and the cross-section of expected stock returns, *Journal of Financial Economics* 49, 345-373.

Campbell, John Y., Andrew W. Lo and A. Craig MacKinlay, 1997, *The Econometrics of Financial Markets* (Princeton University Press: Princeton, NJ).

Chan, K. C. and Nai-Fu Chen, 1991, Structural and return characteristics of small and large firms, *The Journal of Finance* 46, 1467-1484.

Chan, K. C., Nai-Fu Chen and David A. Hsieh, 1985, An exploratory investigation of the firm size effect, *Journal of Financial Economics* 14, 451-471.

Chen, Nai-Fu, Richard Roll and Stephen A. Ross, 1986, Economic forces and the stock market, *Journal of Business* 59, 383-403.

Cochrane, John H., 2001, *Asset Pricing* (Princeton University Press: Princeton, NJ).

Connor, Gregory and Robert A. Korajczyk, 1988, Risk and return in an equilibrium apt: Application of a new test methodology, *Journal of Financial Economics* 21, 255-289.

Cushing, Matthew J. and Mary G. McGarvey, 1999, Covariance matrix estimation, in L. Mátyás, ed., *Generalized Method of Moments Estimation* (Cambridge University Press: Cambridge, MA).

Davidson, James, 1994, *Stochastic limit theory* (Oxford University Press: Oxford, United Kingdom).

Davidson, James, 2000, *Econometric Theory* (Blackwell: New York, NY).

den Haan, Wouter and Andrew Levin, 1996, Inferences from parametric and non-parametric covariance matrix estimation procedures, Working paper, National Bureau of Economic Research.

den Haan, Wouter and Andrew Levin, 1997, A practitioner's guide to robust covariance matrix estimation, in G. Maddala and C. Rao (eds.) *Handbook of Statistics*, Volume 15 (Elsevier Science, North-Holland: Amsterdam, Holland).

Engle, Robert, 1984, Wald, likelihood ratio and Lagrange multiplier tests in econometrics, in Z. Griliches and M. Intriligator (eds.) *Handbook of Econometrics* (North-Holland, Amsterdam, Holland).

Fama, Eugene F. and Kenneth R. French, 1988, Dividend yields and expected stock returns, *Journal of Financial Economics* 22, 3-27.

Fama, Eugene F. and Kenneth R. French, 1992, The cross section of expected stock returns, *The Journal of Finance* 47, 427-465.

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

Fama, Eugene F. and Kenneth R. French, 1996, Multifactor explanations of asset pricing anomalies, *The Journal of Finance* 51, 55-84.

Fama, Eugene F. and Michael R. Gibbons, 1984, A comparison of inflation forecasts, *Journal of Monetary Economics* 13, 327-348.

Fama, Eugene F. and James D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy* 81, 607-636.

Ferson, Wayne E. and Stephen Foerster, 1994, Finite sample properties of the generalized method of moments in tests of conditional asset pricing models, *Journal of Financial Economics* 36, 29-55.

Ferson, Wayne E. and Campbell Harvey, 1999, Conditioning variables and the cross section of stock returns, *The Journal of Finance* 54, 1325-1357.

Ferson, Wayne and Robert A. Korajczyk, 1995, Do arbitrage pricing models explain the predictability of stock returns? *Journal of Business* 68, 309-349.

Gertler, Mark and Simon Gilchrist, 1994, Monetary policy, business cycles, and the behavior of small manufacturing firms, *The Quarterly Journal of Economics* 109, 309-340.

Gibbons, Michael R., Stephen A. Ross and Jay Shanken, 1989, A test of the efficiency of a given portfolio, *Econometrica* 57, 1121-1152.

Hansen, Lars P., 1982, Large sample properties of the generalized method of moments estimator, *Econometrica* 50, 1029-1054.

Hansen, Lars P., John Heaton and Amir Yaron, 1996, Finite-sample properties of some alternative gmm estimators, *Journal of Business and Economic Statistics* 14, 262-280.

Harris, David and László Mátyás, 1999, Introduction to the Generalized Method of Moments, in L. Mátyás (ed) *Generalized Method of Moments Estimation* (Cambridge University Press: Cambridge, United Kingdom).

He, Jia, Raymond Kan, Lillian Ng and Chu Zhang, 1996, Tests of the relations among marketwide factors, firm-specific variables, and stock returns using a conditional asset pricing model, *The Journal of Finance* 51, 1891-1908.

Horowitz, Joel L., Tim Loughran and N. E. Savin, 2000, Three analyses of the firm size premium, *Journal of Empirical Finance* 7, 143-153.

Kim, Dongcheol, 1997, A reexamination of firm size, book-to-market, and earnings price in the cross-section of expected stock returns, *Journal of Financial and Quantitative Analysis* 32, 463-489.

Lehmann, Bruce N. and David M. Modest, 1988, The empirical foundations of the arbitrage pricing theory, *Journal of Financial Economics* 21, 213-254.

Li, Li and Zuliu F. Hu, 1998, Responses of the stock market to macroeconomic announcements across economic states, Working paper, International Monetary Fund.

Lintner, John, 1965, Security prices, risk and maximal gains of diversification, *The Journal of Finance* 20, 587-615.

MacKinlay, A. Craig and Matthew P. Richardson, 1991, Using generalized method of moments to test mean-variance efficiency, *The Journal of Finance* 46, 511-527.

MacKinnon, James G. and Halbert White, 1985, Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties, *Journal of Econometrics* 29, 305-325.

Malkiel, Burton G. and Yexiao Xu, 1997, Risk and return revisited, *The Journal of Portfolio Management* 23, 9-14.

Merton, Robert C., 1973, An intertemporal capital asset pricing model, *Econometrica* 41, 867-887.

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

Newey, Whitney K. and Kenneth D. West, 1994, Automatic lag selection in covariance matrix estimation, *Review of Economic Studies* 61, 631-653.

Perez-Quiros, Gabriel and Allan Timmermann, 2000, Firm size and cyclical variations in stock returns, *The Journal of Finance* 55, 1229-1262.

Ross, Stephen A., 1976, The arbitrage pricing theory of capital asset pricing, *Journal of Economic Theory* 13, 341-360.

Schwert, G. William, 1983, Size and stock returns, and other empirical anomalies, *Journal of Financial Economics* 12, 3-12.

Sharpe, William, 1964, Capital asset prices: A theory of market equilibrium under conditions of risk, *The Journal of Finance* 19, 425-442.

White, Halbert, 1980, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* 48, 817-838.

White, Halbert, 2000, *Asymptotic Theory for Econometricians*, second edition (Academic Press: New York, NY).

Notes:

1. On an informal basis, Fama and French (1992, 1993, 1996) and Li and Hu (1998) compare the sensitivities of small and large firm returns to market risk, dividend yield, the default premium and the term structure. Perez-Quiros and Timmerman (2000) conduct formal tests for differing sensitivities in some non-linear dynamic models of asset returns, but do not carry out such testing in more traditional models.

2. This method, and variations on it, has been used in dozens of papers, for recent applications see Barber and Lyon (1997), Brennan, Chordia, and Subrahmanyam (1998), Ferson and Harvey (1999), Horowitz, Loughran, and Savin (2000) and Kim (1997). For a textbook treatment, see Cochrane (2001, Ch. 12).

3. Estimation of this model, using time series data, is the main focus of Fama and French (1993, 1996).

4. See, for example, Fama and French (1993, 1996).

5. See, for example, Campbell, Lo and MacKinlay (1997, Ch. 2), and Cochrane (2001, Ch. 20).

6. See also Cushing and McGarvey (1999).

7. See Schwert (1983) for a review of early theories and Fama and French (1992, 1993) for further discussion.

8. See Ferson and Korajczyk (1995).

9. Discussed in Ghysels (1998), for example.

10. Included here are Fama and French (1992, 1993, 1996) and Li and Hu (1998).

11. See Ferson and Harvey (1999) for a recent example

12. See Fama and French (1993, 1996), Li and Hu (1998) and Cochrane (2001).

13. See Section 4 for a detailed discussion.

14. We use the econometrics software Eviews 3.1 for all computations of score and other tests.

15. The F statistic also has an unknown distribution in this setting, but we use the F distribution as proxy when computing critical values.

16. See Ferson and Foerster (1994) and Campbell et al. (1997).

17. See, for example, Cushing and McGarvey (1999).

18. Included here are stationarity, finite moments, mixing, etc., as in White (2000) and Davidson (1994, 2000), for example.

19. See, for example Harris and Mátyás (1999).

20. See Table I for summary statistics of covariates, and Table II for correlations with dependent variables.

21. See Gertler and Gilchrist (1994), Li and Hu (1998) and Perez-Quiros and Timmermann (2000).

22. Fama and French (1996) report a similar correlation.

23. See Fama and Gibbons (1984) for a similar procedure.

24. Chen et al. (1986) also use the change in the expected inflation; since this variable is highly correlated with the unexpected inflation and implied test results are practically indistinguishable, we omit them for brevity.

25. Alternatively, one could employ a block-bootstrap method, as in Cochrane (2001, Ch. 15).

26. For similar results see Cushing and McGarvey (1999) and Cochrane (2001, Ch. 15).

27. As pointed out in Gibbons et al. (1989), the test of the CAPM is equivalent to the test of *ex-ante* mean-variance efficiency of a particular portfolio and the test statistic (either F, S or W) can then be interpreted as a measure of distance from the mean-variance frontier.

28. These results correspond to findings of Chan, Chen and Hsieh 1985) who compare similar explanatory variables to portfolios ranked by size for the sample period 1958-1977, roughly our first sub-sample.

29. The test of zero-intercepts is comparable to the F-test of Gibbons et. al (1989) conducted

in Fama and French (1996), with size-related and book-to-market related portfolios added to the excess return on the market proxy. Fama and French (1996) results are also mixed, sometimes rejecting and sometimes accepting the null hypothesis, depending on the type of used dependent portfolios.

30. Standard errors are computed via the VARHAC method.

31. We found similarly strong evidence of conditional heteroskedasticity and serial correlation in a majority of the univariate and bivariate models we studied.

32. Chan et al. (1985) and Fama and French (1993) find different market betas for different size for samples 1958-1977 and 1963-1991, respectively.

33. Fama and French (1993) also report estimates significantly different from zero for both variables in expected returns regressions. However, their estimates do not seem to differ across firms of various sizes.

34. Consistent with findings of Li and Hu (1998).

35. Instability of betas questions the common practice of attaching betas to characteristics such as firm size. Firm betas vary over time: Why should we expect firms grouped by size to have stable betas? We thank John Cochrane for this observation.