Research Program in Finance
Graduate School of Business Administration

WORKING PAPER NO. 32

Tests of Capital Asset Pricing Hypotheses

By
BARR ROSENBERG
and
VINAY MARATHE
TESTS OF CAPITAL ASSET PRICING HYPOTHESES

by

Barr Rosenberg
and
Vinay Marathe

May 1975

Associate Professor and Graduate Student, respectively, Graduate School of Business Administration, University of California, Berkeley. The research for this paper was supported by NSF Grant GS 3306. Preparation of this manuscript was supported in part by the Dean Witter Foundation.
TESTS OF CAPITAL ASSET PRICING HYPOTHESES

Barr Rosenberg and Vinay Marathe*  

A B S T R A C T

Equilibrium models of capital asset markets imply relationships among the means, variances, and certain covariances of the expected returns on securities. A number of competing hypotheses, implying linear or nonlinear relationships, have been advanced. Inferences as to investors' unobservable expectations must be drawn from the distribution of realized returns. The problem of hypothesis testing is further complicated by heteroscedasticity in the residuals and by both sequential and serially uncorrelated variation over time in the crucial parameters of the problem. Immense data bases allow for thorough statistical treatment of these problems. This paper develops tests that rely on stochastic-parameter regression and examines the relationship with other tests that have been employed in the literature. The new procedure is applied to an extensive time series of cross sections of security returns, with some interesting results.

*The authors are associate professor and graduate student, respectively, at the School of Business Administration, University of California, Berkeley. This research was supported by NSF Grant GS 3306. We gratefully acknowledge the assistance of Brian Main and Michel Houglet and the helpful comments of colleagues at Berkeley and Stanford. Responsibility for any errors is, of course, our own.
1.1 Introduction

In equilibrium models of capital asset markets, investors are generally presumed to be risk-averse and to be investing in capital assets for the purpose of maximizing the expected utility of the stream of consumption and bequest opportunities provided by investment returns. The utility function over consumption opportunities may, under simplifying assumptions, be translated into an induced utility function over returns to the investor's portfolio. Either as the result of further simplifying assumptions, or by virtue of a second-order approximation of the utility function, the criteria for portfolio optimization may be expressed in terms of the first and second moments of portfolio return. It is after-tax returns that are relevant, so that attention must be paid to the different tax treatment of cash dividends or coupons, on the one hand, and of capital gains on the other.

Thus, the distribution of return for any portfolio is specified by the mean and variance of total return and the division of return between cash dividends and capital gains. (The variance of cash yield is so small, as a proportion of the variance of total return, that it can be effectively approximated as zero.) The space of available portfolio returns is determined by the first and second moments of the joint distribution of all individual security returns and by the first moments of cash yields on these securities. Equilibrium conditions on capital asset markets take the form of equilibrium conditions on these "return parameters."
These conditions apply to the expectations of investors—that is, to the parameters of the \textit{ex ante} distribution of returns. Since these expectations are unobservable, inference concerning them requires an assumption on the relationship between \textit{ex ante} expectations and \textit{ex post} realizations of returns. The form of this assumption depends on the degree of rationality accorded to expectations.

With this step determined, the problem is to test the consistency of the joint distribution of realized returns with the hypothesized equilibrium conditions. For each of a series of holding periods, the vector of realized returns on all securities in a cross section is treated as a realization of the joint distribution. An extensive data base allows 432 successive monthly holding periods to be studied, with the cross section of securities gradually increasing in size from 419 in 1931 to 874 in 1966.

The plan of this paper is as follows. In the balance of this first section, the notation for the joint distribution of security returns will be introduced, and the concept of the "market" model, an important ingredient in capital-asset-pricing hypotheses, will be developed. Then, some leading capital-asset-pricing hypotheses will be briefly reviewed, and the problem of inference concerning investors' expectations will be explored. The section concludes with the definition of a set of "descriptors" of security characteristics that will be employed as predictors of the parameters of the distribution of return.

The second section develops a multiple-factor model of security returns that provides an economical representation of covariances among
returns. This section concludes with a formal development of progressively more realistic stochastic models for the joint distribution of returns.

The third section develops the procedures to be used for statistical inference. In essence, the problem is one of stochastic-parameter regression, with some unusually interesting features. The problem is perhaps unique in econometrics in involving hypotheses that link means, variances, and covariances. There is intrinsic random parameter variation in each period because of the multiple factors underlying security returns. The long time series of data encourages the allowance for sequential variation or drifting in the parameters of the predictive relations for the return distribution. The huge sample permits sophisticated responses to these problems that demand a great deal from the data. An inference procedure is developed, which is then shown to be analogous to two other approaches that differ markedly at first sight.

In the fourth section, the data sample is defined and the results of the empirical application are presented. These results possess some important implications.

The fifth section reviews the previous empirical tests in this area and compares their results with the present study. This paper builds on the earlier work, and deferral of this review to the end is in no way intended to minimize the extent of this debt. However, the earlier studies did employ statistical methods that were somewhat opaque, and it is difficult to comment on the sampling properties of the tests without
reference to the methodological perspective of Section 3 and the results on parameter variation in Section 4. The paper concludes with a brief summary.

1.2 Security Returns: Notation

Throughout the paper, a few notational conventions will be followed: upper-case letters with a tilde beneath ($\tilde{A}$) will denote matrices; lower-case letters with a tilde beneath ($\tilde{a}$) will denote column vectors; letters with no tilde beneath will denote scalars. The addition of a presuper-script "e" ($e^a$) will denote the random component of the variable ($a$) that cannot be predicted on the basis of available information.

Let there be, $T$ time periods, subscripted $t = 1, \ldots, T$.

Let there be $N(t)$ securities in each period $t$, subscripted $n = 1, \ldots, N(t)$.

Let $i_{nt}$ be the rate of return on security $n$ in period $t$,

$$i_{nt} = \frac{P_{nt} + \text{DIV}_{nt} - P_{n,t-1}}{P_{n,t-1}},$$

where $P_{nt}$ is the end-of-period price, and $\text{DIV}_{nt}$ is dividends or other cash yield in the period.

Let $i_{Ft}$ be the rate of return on a risk-free asset in period $t$, the "risk-free rate."

Let $r_{nt}$ denote a return in excess of the risk-free rate, or an "excess return." Thus, in the "rate-of-return" model:

$$r_{nt} = i_{nt} - i_{Ft}.$$
For convenience in exposition, when an alternative "logarithmic-return model" is under examination, \( r_{nt} \) will denote the excess logarithmic return:

\[
r_{nt} = \log(1+i_{nt}) - \log(1+i_{Ft}).
\]

Let \( \mathbf{r}_t = (r_{1t}, \ldots, r_{N(t),t})' \) be the column vector of excess returns in period \( t \). As will be confirmed below, it is reasonable to assume that these returns obey a joint probability distribution with finite first and second moments:

(A1) Finite Moments

\[
E(\mathbf{r}_t) = (\mu_{1t}; \ldots; \mu_{N(t),t})' \equiv \mu_t
\]

\[
E((\mathbf{r}_t - \mu_t)(\mathbf{r}_t - \mu_t)') = \{\phi_{mnt}\} \equiv \phi_t.
\]

For each time period \( t \), let there be a set of investment proportions, \( w_{1t}, \ldots, w_{Nt} \), summing to one, that define the "market portfolio" in that period. Thus, the rate of return on the market portfolio is

\[
\sum_{n=1}^{N(t)} w_{nt} \cdot i_{nt} \equiv i_{Mt}.
\]

Let \( r_{Mt} \) denote the excess market return:

\[
r_{Mt} \equiv i_{Mt} - i_{Ft}
\]

in the rate-of-return model

\[
r_{Mt} \equiv \ln(1+i_{Mt}) - \ln(1+i_{Ft}),
\]

in the logarithmic-return model.

The return distribution implies a conditional probability distribution for individual security returns, conditional on the realized market
return. If the return distribution is normal, the following assumption will be satisfied:

**(A2): Linear Conditional Expectation**

\[ E(r_t | r_{Mt}) = \alpha_t + \beta_t r_{Mt}, \]

where \( \alpha_t = (\alpha_{1t}, \ldots, \alpha_{N(t)} \r_t) \)' is the vector of mean returns, conditional on a market return equal to zero, where \( \beta_t = (\beta_{1t}, \ldots, \beta_{N(t)} \r_t) \)' is the vector of regression coefficients on the market, and where both \( \alpha_t \) and \( \beta_t \) are independent of \( r_{Mt} \). If (A2) holds, \( \mu_t = \alpha_t + \beta_t r_{Mt} \), where \( r_{Mt} \) is the mean excess-market return. When \( r_{Mt} \) does not obey a multivariate normal distribution, this condition will not necessarily be satisfied. Instead, \( \beta_t \) (or, equivalently, \( \alpha_t \)) may be a function of \( r_{Mt} \). From (A2), the individual security returns may be written as:

\[ r_t = \alpha_t + \beta_t r_{Mt} + \gamma_t. \]

Here, \( \gamma_t = (\gamma_1, \ldots, \gamma_N) \)' is the vector of "residual returns," with \( E(\gamma_t) = 0, \) and \( COV(\gamma_t, r_{Mt}) = 0 \) by construction, \( VAR(\gamma_t) = \gamma_t \). The residual returns will not be independent of the market return unless the joint distribution is normal.

### I.3 Capital-Asset-Pricing Hypotheses

Jensen (1972) provides an excellent review of a number of hypotheses, derived under progressively more realistic assumptions, concerning the equilibrium values of expected return. The capital-asset-pricing hypotheses to be considered below concern the dependence of expected return
upon three other parameters of the *ex ante* distribution of returns:

\[ \beta_{nt}, \text{ the regression coefficient upon the market return;} \]

\[ \delta_{nt} = E[DIV_{nt}/P_{n,t-1}], \text{ the expected cash yield;} \]

\[ \phi_{nt} = VAR[r_{nt}], \text{ the variance of return.} \]

Some of these hypotheses are most naturally expressed in terms of the rate-of-return model; others in terms of the logarithmic-return model. All of them may be expressed as special cases of the linear equation,

\[ (2) \quad \alpha_{nt} = c_{1,t} + c_{\beta,t}\beta_{nt} + c_{\delta,t}\delta_{nt} + c_{\phi,t}\phi_{nt} = c_{t}'\gamma_{nt}, \]

where \( \beta_{nt} = \beta_{nt} - \bar{\beta}.t, \delta_{nt} = \delta_{nt} - \bar{\delta}.t, \phi_{nt} = \phi_{nt} - \bar{\phi}.t \) are differences from the average values in the cross section, and

\[ \gamma_{nt} \equiv (1:\beta_{nt}:\delta_{nt}:\phi_{nt})' \quad \text{and} \quad c_{t} \equiv (c_{1,t}:c_{\beta,t}:c_{\delta,t}:c_{\phi,t}). \]

The basic hypothesis is the Capital Asset Pricing Model (CAPM), first derived by Sharpe (1964) and Lintner (1965). Under this model, the market portfolio is the unique optimal portfolio of risky assets, and the expected return on securities must be such as to compensate for their contribution to the risk of the market portfolio. The equilibrium condition is:

\[ E(i_{n}) = i_{p} + \beta_{n}(E(i_{M}) - i_{p}), \]

or

\[ E(r_{n}) = \beta_{n}E(r_{M}). \]
In terms of the market model, the hypothesis becomes

\[ \alpha_{nt} = 0 \]

or

\[ c_1 = c_\beta = c_\delta = c_\phi = 0. \]

Next, it is possible that the constant term is not \( i_p \), but some other rate \( i_Z \). This may occur because the risk-free borrowing rate is incorrectly measured, in which case \( i_Z \) represents the true borrowing rate, or because the opportunity to borrow is entirely absent, in which case \( i_Z \) becomes the rate of return on the optimal portfolio of stocks having beta equal to zero, i.e., the "rate of return on the zero-beta portfolio." The equilibrium condition becomes:

\[ E(i_n) = i_Z + \beta_n (E(i_M) - i_Z). \]

This implies that, in the model with excess returns expressed relative to \( i_F \),

\[ E(i_n - i_F) = i_Z - i_F - \beta_n (i_Z - i_F) + \beta_n (E(i_M) - i_F) \]

or

\[ E(r_n) = i_Z - i_F - \beta_n (i_Z - i_F) + \beta_n E(r_M). \]

In terms of the market model, the hypothesis is

\[ \alpha_n = k(1 - \beta_n), \quad \text{where} \quad k = i_Z - i_F \]

or

\[ c_\beta = -k; \quad c_1 = c_\delta = c_\phi = 0. \]
Next, if investors hold inadequately diversified portfolios (because of transactions or information costs, capital gains taxation, or institutional constraints), then the specific risk of securities contributes to portfolio risk, and risk-averse investors would require compensation in the form of higher expected return for specific risk, or, as an implication of this, for total risk:

CAP:S: \[ c_\phi > 0. \]

If investors respond to the higher tax rate on cash dividends than on capital gains and, therefore, require higher expected return as a larger percentage of that return comes in the form of dividends, Brennan (1970) has derived a hypothesis, dependent on the tax rates, that:

\[
E(r_n) = (1-T)i_F + T\delta_n + \beta_n (E(r_M) - i_F + T(i_F - \delta_M)),
\]

where \( T \) is an indicator of the excess of the tax rate on cash yield over the tax rate on capital gains, \( T > 0 \), and where \( \delta_M \) is the expected yield on the market portfolio. For the model in terms of expected returns, this implies:

\[
E(r_n) = -T i_F + T\delta_n + \beta_n T(i_F - \delta_M) + \beta_n E(r_M).
\]

Therefore,

\[
\alpha_n = -T i_F + T\delta_n + \beta_n T(i_F - \delta_M), \text{ or}
\]

CAP:T: \[ c_\delta = T; \quad c_\beta = T(i_F - \delta_M); \quad c_1 = c_\phi = 0. \]
Finally, attention to the compounding of returns over holding periods longer than the observation interval suggests that the equilibrium condition should be expressed in terms of the expectations of the logarithmic returns, Merton (1973) orLintner (1973), and that the equilibrium expected logarithmic return obeys

\[ \alpha_n = \frac{V_M}{2} \beta_n - 1/2 \phi_{nt}, \] or

\[ c_\beta = \frac{V_M}{2}; c_\phi = -1/2.\]

This equation is exact under the diagonal assumption (A6) to be introduced below, but is only an approximation for the multiple-factor model.

The existence of these several hypotheses lends interest to the problem of ascertaining the empirical relationship between \( \alpha \) and the other parameters of the distribution of returns.

1.4 Operational Measures of the Expectational Parameters

The capital asset pricing hypotheses state equilibrium relationships among parameters of the return distribution expected by investors. In practice, these parameters are unobservable, for no historical record of expectations is available. To operationalize these hypotheses, it is necessary to infer from the historically realized distribution of returns what sort of expectations might reasonably have been held by investors. Two distinct approaches have been taken in the literature.

The most common approach has been to obtain, for each date \( t \), historical estimators of the parameters in question, based on data from an
historical period just prior to date $t$, and to identify these estimates with the parameters of the \textit{ex ante} distribution for $t$. The conventional historical measures, which will also be used in subsequent empirical work in this study, follow:

$$
H_{nt}^2 = \sum_s (r_{ns} - \bar{r}_n)(r_{Ms} - \bar{r}_M)/\sum_s (r_{Ms} - \bar{r}_M)^2,$$
where the sum is taken over the sixty monthly holding periods in the five calendar years prior to $t$, and where $\bar{r}_n$ and $\bar{r}_M$ denote averages over this interval.

$H_{nt}^2$ = residual mean square from the above regression, and

$H_{nt}^2$ = dividends per share paid in the prior calendar year divided by the price at the end of that year.

The information contained in these historical measures was published as of January 1 of the year containing date $t$, so that it was certainly feasible for investors to form expectations by this "naive prediction rule" that sets the expected value for the parameter in period $t$ equal to the historical estimate. However, is it probable that this naive prediction rule was employed by investors? Each of these historical measures is only one among many possible descriptions of the characteristics of the firm that investors might have utilized in forecasting risk and yield. It will be helpful to represent the relevant information set by the values of a set of variables describing the firm. The historical measures just described could be three among this set.

For each security $n$, let $x_{In}, \ldots, x_{Jn}$ be a set of $J$ variables describing characteristics of that security, or "descriptors." For each
holding period \( t \), the values \( x_{1t}, \ldots, x_{Jt} \) are assumed to be based upon information published prior to the beginning of the holding period and therefore available to investors for the purposes of forming an investment strategy. Let \( x_{nt} \) be the \( J \)-element column vector of descriptors and assume, for convenience, that the first descriptor is a constant, \( x_{1t} = 1 \) for all \( n, t \).

The descriptors are assumed to be linearly independent in each period \( t \), since a dependent descriptor can be discarded with no loss of information. These descriptors are intended to serve as predictors for return parameters in period \( t \), and thus to be relevant for portfolio investment decisions.

These descriptors may provide predictive content in two ways: the descriptor may be a measure of the current state of the firm as it impacts future return, risk of return, and yield; the descriptor may be an estimate of characteristics of the firm evidenced in the recent past that are expected to persist into the future. In the second category fall the historical estimates of the return parameters defined above. Before discussing the likely relationship between the expectational return parameters and these descriptors, it will be helpful to formalize the dependence of the true return parameters upon the descriptors.

An important simplifying assumption is that the descriptors are so constructed that the conditional mean of the return parameters for any security \( n \), given the predetermined descriptors, is a linear function of the descriptors. We assume that the descriptors are constructed so that:
(A3): Linear Predictive Relations

Each of the return parameters is given as a function of the descriptors by a linear relation,

\[ \alpha_{nt} = a'_{-t}x_{-nt} + e^{\alpha}_{nt} \]

\[ \beta_{nt} = b'_{-t}x_{-nt} + e^{\beta}_{nt} \]

\[ \delta_{nt} = d'_{-t}x_{-nt} + e^{\delta}_{nt} \]

\[ \phi_{nt} = h'_{-t}x_{-nt} + e^{\phi}_{nt} \]

(4)

where the residual terms with presuperscript "e" are independent of the descriptors.

The relationship between the return parameters and the descriptors can be collected in the matrix equation:

\[
\begin{pmatrix}
\alpha_{nt} \\
\beta_{nt} \\
\delta_{nt} \\
\phi_{nt}
\end{pmatrix} =
\begin{pmatrix}
1 & \cdots & i \cdot 0 & \cdots & 0 \\
\end{pmatrix}
\begin{pmatrix}
a'_{-t} \\
b'_{-t} \\
d'_{-t} \\
h'_{-t}
\end{pmatrix}
\begin{pmatrix}
x_{-nt} \\
0 \\
0 \\
0
\end{pmatrix} +
\begin{pmatrix}
e^{\alpha}_{nt} \\
e^{\beta}_{nt} \\
e^{\delta}_{nt} \\
e^{\phi}_{nt}
\end{pmatrix} =
\begin{pmatrix}
a'_{-t} \\
b'_{-t} \\
d'_{-t} \\
h'_{-t}
\end{pmatrix}
\begin{pmatrix}
x_{-nt} \\
0 \\
0 \\
0
\end{pmatrix} +
\begin{pmatrix}
e^{\alpha}_{nt} \\
e^{\beta}_{nt} \\
e^{\delta}_{nt} \\
e^{\phi}_{nt}
\end{pmatrix}.
\]

(5)
This assumed relationship is the rational prediction rule for the return parameters, in the sense of providing the minimum-variance unbiased prediction that employs information contained in the descriptors. Notice that each descriptor possibly provides predictive content for all return parameters.

Now, what relationship would investors' expectational parameters have borne to these descriptors? Four possible approaches have been taken in the literature:

1. **Expectation Equal to Historical Estimator.** Under this approach, as mentioned previously, the expectational parameter is assumed to be identical to a historical estimator taken from the recent past.

2. **Expectation Equal to the Historical Value.** This approach remains a "naive" prediction rule, in that the prediction is equal to the average value over some historical interval. However, investors are assumed to have learned, perhaps through information sources not now available to us, the true historical average as opposed to the estimator of that average.

3. **Expectations Given by a Rational Prediction Rule Fitted Over a Prior Interval.** Under this approach, the predicted return parameters are given by

\[
\begin{pmatrix}
\hat{\alpha} \\
\hat{\gamma}_{nt}
\end{pmatrix} = \begin{pmatrix}
\hat{a} \\
\hat{b}
\end{pmatrix} \hat{x}_{nt},
\]

where the hats denote estimators of the rational prediction rule (5), fitted by regression of the realized distribution of returns in previous periods on the predetermined descriptors referring to those periods.
(E4) Expectations Given by a Rational Prediction Rule Fitted to the Entire Time Series. Under this approach, the expectations are presumed to be identical to the predicted values for return parameters obtained from estimation of (5) over the entire sample.

The choice among these four approaches hinges on one's view of something unobservable, the process of expectations formation. The issue is an extremely critical one, for as will be seen in the empirical results, the different prediction rules include quite different coefficients for the descriptors and, hence, imply different conclusions as to the meaning of the contributions of the descriptors to subsequent returns.

We believe that, if the descriptors are so chosen as to provide rational prediction rules that are stationary over time, the fourth approach is the appropriate one. The grounds for this belief are that the information in possession of investors at the time the expectations were formed was vastly greater than anything now available to us in machine-readable data files. As a result, we conclude that investors must have captured all the information in the descriptors that we can reconstruct, and, moreover, that investors possessed greater knowledge about the predictive content of those descriptors, even at an early date in the sample history, than we now possess from an estimated prediction rule based on the entire sample history.

Is it true that investors possessed far greater information than we can now reconstruct? Considering first the information contained in historical price behavior, one might advance the contrary argument that
investors, most of whom lacked quantitative training, were bound to waste much of the information in their possession, with the result that systematic statistical methods would allow us to construct equal and possibly superior predictions of risk. This argument applies particularly to the years prior to 1950, when the understanding of stock price behavior and investment risk, and the computers available to implement that understanding, were relatively primitive. However, in our view, any possible inefficiency in investors' processing of information would have been more than offset by the vastly greater extent of information in their possession. Our information concerning returns prior to 1962 is effectively limited to returns over monthly holding periods. By contrast, investors could observe the continuous process of trading, and daily, weekly, and monthly returns and price ranges were published. Because of the virtual absence of serial correlation in returns, daily returns on securities offer the potential to estimate \( \beta \) and \( \phi \) with less than one-twentieth the variance of monthly returns, and daily ranges offer further information concerning \( \phi \). As a consequence, far less efficient processing of the available information would have led to estimates with as much predictive content as the best we can reconstruct from data now available to us.

Second, investors were possessed of information about the fundamental operations of the firm, which offered the potential to improve upon the best predictions based upon historical price behavior. Historical records of some of the most important items of fundamental information are available to us for the period since 1949 through the COMPSTAT database. Recent studies (Rosenberg and McKibben (1973) and Rosenberg et al.
(1973)) have confirmed two relevant points. First, in best linear prediction rules for risk parameters based upon fundamental data as well as historical price behavior, fundamental data plays an important part and allows major improvements over the predictive power of historical risk descriptors. Second, the best linear prediction rules respond to fundamental data in conformity with the conventional wisdom about the determinants of risk. Since the fundamental information employed in these studies constitutes only a small fraction, although perhaps the most important fraction, of information available to investors, it becomes difficult to reject the conclusion that investors could have employed that information to achieve risk predictions that were superior to those based on historical price behavior alone. This conclusion is indirectly supported by many empirical studies that have confirmed the high degree of efficiency with which market prices impound fundamental data in the prediction of subsequent expected returns.

In short, we view the historical descriptors that can now be constructed as a small window on a rich data set available to investors at the time. We conjecture that the expectations as to risk and yield would have incorporated all of the information visible through this window, and then some. Formally, let

\[
\begin{align*}
(\alpha) & = (\hat{\alpha}) + \xi_{nt} \\
(\gamma)_{nt} & = (\hat{\gamma})_{nt} + \zeta_{nt} \\
(\hat{\alpha}) & = (\hat{\alpha}) + \xi_{nt} \\
(\hat{\gamma})_{nt} & = (\hat{\gamma})_{nt} + \zeta_{nt} + \eta_{nt},
\end{align*}
\]
where the superscript "ω" denotes the expectations of investors, ζ introduces the surprises relative to expectations, and η is that component of expectations that is unpredictable by the information in the descriptors. If all information in the descriptors was incorporated in the predictions, as we are assuming, and if the criterion for prediction was a minimum mean square error prediction rule, then ζ and η will be independent of \( x_{nt} \), as we have already assumed to be the case for their sum; as a result, any procedure that estimates the regression of the true parameters on the descriptors unbiasedly (or consistently) will also estimate the regression of the expectational parameters on the descriptors unbiasedly (or consistently).

However, it is certainly possible that the process of expectation formation either did not employ some of the information in the descriptors or employed it inefficiently. Then the estimated prediction rule attributes to some descriptor predictive content for the return parameter that was not reflected in the expectation, because \( d_{nt} \) is negatively correlated with the descriptors. This possibility must be recognized as one of the potential explanations for a rejection of the CAP hypotheses. If investors failed to incorporate the information in a descriptor in forecasting the parameters of return, then prices would not have adjusted in order to set expected returns to reflect that contribution.

1.5 Hypothesis Tests and the Choice of Descriptors

The CAP hypotheses imply constraints on the vector \( (\alpha_{nt}, \gamma'_{nt},) \) of expectational parameters that should hold across all \( n \) in period \( t \).
Under the assumption that a prediction rule for this vector of the form (5) has been correctly identified, the same constraints must apply to the predicted values for all \( n \), provided that the residual or unpredicted component is independent of the predicted one. To explore this point further, note that equation (5) can be rewritten as:

\[
\begin{bmatrix}
\alpha_{nt} \\
\gamma_{nt}
\end{bmatrix} = \sum_j x_{jnt} \begin{bmatrix}
\alpha_{jt} \\
\gamma_{jt}
\end{bmatrix} + \begin{bmatrix}
e_{\alpha_{nt}} \\
e_{\gamma_{nt}}
\end{bmatrix},
\]

where \( B_{jt} \), the \( j \)th column of \( B \), is the vector of coefficients for \( x_j \).

Since the residual term is independent of the descriptors, variation in any descriptor \( x_j \) implies an increment in the vector of return parameters that is proportional to the vector \( (\alpha_{jt} : B_{jt} \gamma_{jt})' \). Therefore, if any exact linear constraint is to apply to the vector \( (\alpha_{jt} : \gamma_{jt})' \) for all \( n \), that constraint must be satisfied for all \( x_{jnt} \). Since the descriptors are not linearly dependent, the constraint must therefore be satisfied by \( (a_j : B_{j} \gamma_j)' \) for each \( j \). In other words, the relative predictive content of each descriptor for the various return parameters must obey the same constraint as the return parameters themselves. It is this fact that will allow tests of the relationships among the return parameters, by means of tests of the predictive relationships.

Notice that the same constraint must be obeyed by the residual vector \( (e_{\alpha} : e_{\gamma}) \) for each \( n \). This implies a singularity in the distribution of these random terms.
The constraints implied by CAP:M and CAP:Z illustrate this approach to hypothesis testing. The basic CAP:M hypothesis requires that \( a_n = 0 \) for all \( n \). Stated in terms of the coefficients of the prediction rules, this requirement becomes:

\[
\text{CAP:M} \quad \quad a = 0.
\]

Notice that this requirement holds regardless of the descriptors being used. Under the CAP:M, the only permissible contribution of a descriptor to expected return arises as a compensation for the contribution of that descriptor to systematic risk and will be attributed to that source through the estimate of \( b \).

The CAP:Z requires that \( a_n = k(1-\beta_n) \) for all \( n \). Since the contribution of the descriptors to \( \beta \) is given by \( b'x_n \), we must have \( a'x_n = k(1-b'x_n) \) for all \( n \). Since \( x_{1n} = 1 \) for all \( n \), this can hold iff:

\[
\begin{pmatrix}
a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_J
\end{pmatrix}
= k
\begin{pmatrix}
1 - b_1 \\ - b_2 \\ - b_3 \\ \vdots \\ - b_J
\end{pmatrix}.
\]

Thus, the \( 2J \) coefficients \( a \) and \( b \) obey \( J \) constraints, in which the variable \( k \) appears multiplicatively.
The justification behind this testing procedure is that equilibrium requires that it be impossible for any investor, by exploiting the predictive content of these descriptors, to construct or alter his portfolio to increase the expected utility of the distribution of returns on his portfolio. The predictive relationships determining the parameters of the return distribution as functions of the descriptors must, in equilibrium, be such that for all values of the descriptors the marginal contribution of security return to portfolio return satisfies the Kuhn-Tucker conditions for the optimality of the investor's portfolio. If this were not the case, then a security could be found such that, as a consequence of the value of its descriptors, the predicted return parameters made it a desirable addition to (subtraction from) the portfolio and, as a result, the price of the security would be bid up (down), thereby adjusting expected return in the opposite direction until the predicted return parameters conformed to the equilibrium condition.

In the next section, the problem of estimating the various prediction rules will be considered. For the moment, notice that the framework for the hypothesis tests imposes no restrictions on the kind of information contained in the descriptors. Any variables may be included as descriptors, provided that:

(1) the descriptors were predetermined and published prior to the holding period \( t \) to which they refer;

(2) the relationship between the descriptors and return parameters was stationary over time;
(3) the relationship between the descriptors and parameters was independent of the random components of returns in period $t$; and

(4) the descriptors were such that their information content would have been reflected in investors' expectations.

To insure comparability of our results with earlier studies, we have chosen a set of descriptors that adhere to these conditions to a limited degree, but which leave much room for improvement. The six descriptors that will be used are the constant term, the three defined previously—$H\beta$, $H\delta$, and $H\sigma^2$—, and two that introduce nonlinearities, $H\sigma \equiv (H\sigma^2)^{1/2}$, and $H\beta^2 \equiv (H\beta)^2$. The latter is a necessary component of the prediction for total risk, $\phi_{nt} = \sigma_{nt}^2 + \beta_{nt}^2 V_{mt}$.

It would have been beneficial to include many other descriptors, indicative of industry group and other fundamental characteristics, that have been shown to improve prediction of the risk parameters. Fundamental variables other than industry groupings were not available for the first twenty-five years of the sample. Industry groups were available and were excluded only to simplify the analysis. Additional transformations on the descriptors that might potentially have improved explanatory power, and stationarity and linearity of the predictive relationships—in particular, the formulation of a "Bayesian adjusted" descriptor of historical beta of the form $(H\beta_{nt} - 1)/(\text{estimation error variance in } H\beta_{nt})$—were excluded, again to keep the problem manageable.
II. A FORMULATION OF THE JOINT DISTRIBUTION OF RETURNS

In this section, a model for the joint distribution of returns in one time period, \( t \), will be developed. Since all constructs refer to period \( t \), the time subscript will be deleted to simplify the notation. It should be understood that the parameters and variables refer to time period \( t \) only. The time subscript will return at the conclusion of this section where a model of the variation of these parameters over time is proposed.

II.1 A Multiple-Factor Model, with Known Loadings and Predictable Heteroscedasticity

Consider the following model for security returns:

\[
\mathbf{r}_n = \sum_{j=1}^{J} \mathbf{x}_j \mathbf{f}_j + \mathbf{\rho}_n = \mathbf{x}_n' \mathbf{f} + \mathbf{\rho}_n, \quad n=1, \ldots, N(t)
\]

where the \( \mathbf{f}_j, j=1, \ldots, J \) are random "factors" influencing returns in period \( t \), and \( \mathbf{\rho}_n \) is the component of return that is truly specific to security \( n \), in that it is unrelated to returns on other securities.

The \( J \) factors introduce the correlation between returns. The \( \mathbf{x}_j \) are the partial derivatives of return on security \( n \) with respect to these factors, or the factor loadings of security \( n \). An important simplifying assumption is that the \( J \) factors having these known loadings account for all covariance between security returns:

(A4): Multiple Factor Model with Known Loadings

The individual component of return, \( \mathbf{\rho}_n \), is specific to security \( n \), in that \( \text{COV}(\mathbf{\rho}_n, \mathbf{\rho}_m) = 0 \) for \( m \neq n \), \( \text{COV}(\mathbf{\rho}_n, \mathbf{f}_j) = 0 \).
This model was developed in Rosenberg (1974), based upon the more intuitively appealing model in which factors are associated with events in the macroeconomy. Under the assumption that the loadings of a firm's return on these macroeconomic factors are determined by microeconomic characteristics of the firm that are captured in the descriptors, a transformation of the original macroeconomic factors yields a set of factors of the form assumed here, each associated with a descriptor.

Notice that, since the factor loadings equal the descriptors, the values of the factors enter in the model as if they were random parameters multiplying the descriptors. The existence of such factors, multiplying descriptors such as Hβ and H0², was demonstrated in Black, Jensen, and Scholes (1972) and Fama and MacBeth (1973). In Rosenberg, Houglet, Marathe, and Mckibben (1973), the model was applied with an extensive list of descriptors capturing income-statement and balance-sheet characteristics, industry groupings, and descriptors of historical price behavior such as those defined above. The results indicated that a high degree of covariance among security returns could be attributed to factors with these loadings. In the empirical application developed below, a middle ground will be chosen, in that the number of factors will be small, but random factors will be introduced for the primary descriptors in an attempt to capture the largest part of the covariance among residuals.

Under assumption (A4), from the finite-moments assumption (A1), the variance-covariance matrix of returns becomes:
\[ \phi_t = XFX' + \Sigma, \]

where: \( X = (x_1, \ldots, x_{N(t)})' \) is the \( N(t) \times J \) matrix of predetermined descriptors

\[ F \equiv \text{VAR}(\xi), \]

\[ \Sigma = \text{DIAG}(\sigma_n^2), \text{ with } \sigma_n^2 \equiv \text{VAR}(\rho_n). \]

The mean vector of returns is given by:

\[ \mu = X\bar{\xi} + \bar{\rho}, \]

where \( \bar{\xi} \) is the vector of means of the factors, and \( \bar{\rho} \) is the vector of means of the specific returns.

Three possible simplifying assumptions on the joint distribution are:

(A5): Predictable Heteroscedasticity

The variance of specific return differs across firms in a predictable fashion:

\[ \sigma_n^2 = \sigma^2 \left( \sum_{j=1}^J s_j x_{jn} + \sigma_n^2 \right) = \sigma^2 \left( \bar{s}' \bar{x}_n + \sigma_n^2 \right), \]

where the \( s_j \) are fixed regression coefficients, and \( \sigma_n^2 \) is an unpredictable component that is uncorrelated with the descriptors across the sample and that is hoped to be of small importance.

(A6): The Diagonal Model

There is only one factor, i.e., \( J = 1 \).
(A7): Homoscedasticity

The variance of specific return is identical for all firms, or
\[ \sigma_n^2 = \sigma^2 \quad \text{for all} \quad n. \]

Previous studies have shown that the diagonal model may be rejected in favor of the multiple-factor model with known loadings (King (1966) and Rosenberg et al. (1973)), and that the homoscedasticity assumption may be rejected in favor of predictable heteroscedasticity (Rosenberg and McKibben (1973)). In the application developed in the present article, the exact nature of the variance-covariance matrix of returns is of secondary interest. What will be essential is to capture those aspects that influence hypothesis tests on the values of coefficients in a regression of the returns upon the descriptors. Efficient prosecution of these tests requires that the approximation to \( \tilde{\Phi} \), \( \tilde{\Phi} \), be such that:

\[
X\tilde{\Phi}^{-1}X \approx \tilde{X}'\tilde{\Phi}^{-1}\tilde{X},
\]

since the accuracy of this approximated information matrix determines the efficiency of the generalized least squares estimator and the validity of the sampling theory for this estimator.

The two simplifying assumptions (multiple-factor model with known loadings equal to \( \tilde{x} \)) and (heteroscedasticity linear in \( \tilde{x} \)) amount to taking:

\[
(10) \quad \tilde{\Phi} = \tilde{\Phi} + \tilde{R} = \tilde{X}\tilde{F}\tilde{X}' + \text{DIAG}(\tilde{s}'\tilde{x}_n) + \tilde{R},
\]

where \( \tilde{R} \) is the matrix of unexplained covariances not captured in the simplified model. \( \tilde{R} \) is intrinsically small relative to the explained
variances and covariances in \( \Phi \), as has been confirmed in empirical studies. As a result, the step from the simplified homoscedastic-diagonal model (A6, A7) to the multiple-factor, predictably heteroscedastic model (A4, A5) is relatively more important than the improvement possible by further relaxation of the simplifying assumptions.

II.2 The Implied Model of Market and Residual Returns

Let \( \mathbf{w}_M = (w_{M1} : w_{M2} : \ldots : w_{MN})' \) be the vector of weights in the market portfolio. Then, from (7):

\[
\mathbf{r}_M = \mathbf{w}_M' \mathbf{r} = \mathbf{w}_M' \mathbf{Xf} + \mathbf{w}_M' \mathbf{o} = \mathbf{x}_M' \mathbf{f} + \mathbf{w}_M' \mathbf{o},
\]

where \( \mathbf{x}_M = \mathbf{E} \mathbf{w}_M \mathbf{x}_N \) is the vector of loadings of the market onto the factors. Therefore, the moments of the market return are:

\[
\mathbf{r}_M = \mathbf{E}[\mathbf{r}_M] = \mathbf{x}_M' \mathbf{f} + \mathbf{w}_M' \mathbf{o}, \quad \mathbf{V}_M = \mathbf{VAR}[\mathbf{r}_M] = \mathbf{x}_M' \mathbf{F} \mathbf{x}_M + \sum_{n=1}^{N} w_{Mn}^2 \sigma_n^2.
\]

The factors and the specific risk constitute the components of risk in the multiple-factor model. The regression coefficients of these risk components onto the market return, analogous to beta coefficients for these items, are:
\[
\begin{align*}
\theta = \begin{pmatrix}
\theta_1 \\
\vdots \\
\vdots \\
\theta_J
\end{pmatrix} = \frac{\text{COV}(z, r_M)}{\text{VAR}(r_M)} = \frac{F_{z M}}{V_M} \\

\psi = \begin{pmatrix}
\psi_1 \\
\vdots \\
\vdots \\
\psi_N
\end{pmatrix} = \frac{\text{COV}(-, r_M)}{\text{VAR}(r_M)} = \frac{\Sigma_{z M}}{V_M} = \frac{1}{V_M} \begin{pmatrix}
\sigma^2_{W M L} \\
\vdots \\
\vdots \\
\sigma^2_{W M N}
\end{pmatrix}
\end{align*}
\]

(13)

It is useful to express each underlying risk component as the sum of its regression onto the market plus a residual component which is uncorrelated with the market:

\[
f - \bar{f} = \theta (r - \bar{r}_M) + g, \quad \psi = \psi + \psi (r - \bar{r}_M) + u,
\]

where \(g\) and \(u\) are the "residual factors" and "residual risk," respectively. The residuals are linear functions of the underlying risk components.

\[
\begin{pmatrix}
g' \\
u'
\end{pmatrix} = \begin{pmatrix}
f - \bar{f} \\
\psi - \bar{\psi}
\end{pmatrix} = \begin{pmatrix}
\theta \\
\psi
\end{pmatrix} \begin{pmatrix}
x'_M \\
\psi'_M
\end{pmatrix} = \begin{pmatrix}
f - \bar{f} \\
\psi - \bar{\psi}
\end{pmatrix}
\]

(14)

From (14), the moments of the residual factors are found to be:

\[
\bar{g} = 0 \quad , \quad \bar{u} = 0 \quad , \quad \text{and}
\]
\[
\begin{pmatrix}
\text{VAR}(\tilde{z}) : \text{COV}(\tilde{z}, u) \\
\text{COV}(u, \tilde{z}) : \text{VAR}(u)
\end{pmatrix}
= \begin{pmatrix}
\bar{\Sigma} & \Sigma w_M \\
0 & \Sigma w_M
\end{pmatrix}
- \begin{pmatrix}
\Sigma w_M \\
\Sigma w_M
\end{pmatrix}
\begin{pmatrix}
\Sigma w_M \\
\Sigma w_M
\end{pmatrix}^{-1}
\begin{pmatrix}
\Sigma w_M \\
\Sigma w_M
\end{pmatrix}.
\]

Thus, the variance of the residual components is equal to the variance of the original components less a positive semidefinite matrix of rank 1, which removes that part of variance attributable to the market return.

The covariance between any security return and the market return is:

\[
\text{COV}(r_m, r_M) = \text{COV}(x'_m \bar{f} + \rho_m, x'_m \bar{f} + \rho_m) = x'_m \Sigma \bar{f} + \rho_m \sigma_m^2.
\]

Therefore, the regression coefficient of the security return onto the market return is:

\[
\beta_m = \frac{(x'_m \Sigma \bar{f} + \rho_m \sigma_m^2)}{(x'_m \Sigma \bar{f} + \rho_m \sigma_m^2 + \sum_n \frac{w_{Mn} \sigma_n^2}{Mn}}.
\]

\[
= \bar{x}' \Sigma \bar{f} + \rho_m = \Sigma \bar{x}_j \bar{f}_j + \frac{w_{Mm} \sigma_m^2}{Mm}.
\]

Thus, the security beta is a weighted sum of the "betas" of the underlying factors, weighted by the security's factor loadings, plus the beta of the specific risk of the security. This latter is nonzero only because the specific risk of the security enters into the market return due to the security's inclusion in the market index.

The mean return on the security is:

\[
u_m = E[r_m] = x'_m \bar{f} + \rho_m = x'_m \bar{f} + \rho_m - \beta_m \bar{f} + \beta_m \bar{f}_M
\]
\[ \bar{r} - \bar{r}_m = x'f + \bar{\rho}_m - (x'\bar{\theta} + \bar{\psi})'\bar{r}_M + \beta \bar{r}_M. \]

Therefore, since \( \mu_m = \alpha_m + \beta \bar{r}_M \),

\[ \alpha_m = \mathbb{E}[r_m | r_M = 0] = x'\bar{f} - \bar{\theta}'\bar{r}_M + (\bar{\rho}_m - \bar{\psi})' \bar{r}_M. \]  \( \text{(18)} \)

From (1), the individual security return may be written as:

\[ r_m = \alpha_m + \beta r_M + \psi_m, \]

where substitution of (16) and (18) yields:

\[ \psi_m = r_m - \alpha_m - \beta r_M = (r_m - \bar{r}_m) - \beta (r_m - \bar{r}_M) \]

\[ = x'(r - f) + \bar{\rho}_m - \bar{\psi}_m - (x'\theta + \psi_m)(x'f - \bar{f}) + \psi_m(\bar{\rho} - \bar{\psi}) \]

\[ = x'(r - f) + \psi_m. \]  \( \text{(19)} \)

Finally, (15) and (19) yield the familiar formula for the variance of a conditional distribution,

\[ \psi \equiv \text{VAR}(\psi) = \text{VAR}(r | r_M) = \text{VAR}(\bar{r}) - \beta V_M \beta' = \Sigma - \beta V_M \beta'. \]

The constructs of the market-return and residual-return model \((r, r_m, \psi_m, \alpha, \beta, \psi, \text{ and } \psi)\) have now been expressed in terms of the underlying multiple-factor model. However, it will be convenient to make a few approximations in the model for residual returns, so that it assumes the same simple multiple factor form as the raw returns. To accomplish this,
we need to find a set of transformed factors, factor loadings, and specific returns, such that:

\[ v_n = \tilde{z}_n g^* + u_n^* \]

where \( \tilde{z}_n \) denotes the transformed factor loadings and the superscript "*" indicates a transformed factor, such that:

(a) \( \text{VAR}(g^*) = G^* \) is nonsingular

(b) \( \text{VAR}(u^*) = \Sigma^* \) is diagonal and nonsingular and satisfies (A5) or (A7) if the underlying model satisfies these

(c) \( \text{COV}(g^*, u^*) = 0 \) so that (A4) is satisfied.

Now, conditions (b) and (c) are satisfied to a very close approximation by the untransformed residual factors and specific returns, as inspection of (15) will verify.\(^1\) Thus, there is little cost in imposing these conditions on the residual returns. However, a transformation of the residual factors is suggested by the requirement that \( G^* \) be nonsingular.

---

\(^1\)The magnitude of the off-diagonal entries in \( \text{VAR}(u) \) may be expressed in terms of the implied correlations between the residual specific returns, \( \text{CORR}(u_m, u_n) \). With an equal weighted market index, these will be on the order of \( 10/N^2 \), or miniscule. In a capitalization-weighted market, with capitalization typical of the U.S. economy, the correlation might be as great as \(-.02\) among the five or so largest firms. The violation of assumptions (A5) or (A7) would be of the same order of magnitude.

Similarly, the magnitude of the covariance between the residual factors and the residual specific returns may be measured by the correlation between an eigenvector of the residual factors and an individual residual return. From the empirical results to date (Rosenberg et al. (1973)), the typical value of the correlation is no more than \( (1/N) \), and even in a capitalization-weighted market, the correlation between the largest firm's specific-residual return and the most important residual factor would be, at most, .04.
In reality, the contribution of specific risk to the market variance given in (12) is small, because of the small weights accorded to individual stocks. Consequently, \( V_M = x_M^t F x_M \). When this is substituted into expression (15) for \( \zeta \equiv \text{VAR}(\tilde{y}) \), this matrix is found to approach singularity, since the variance of the linear combination of residual factors, with weights equal to the market weights, is close to zero:

\[
(20) \quad \text{VAR}(x_M^t \tilde{y}) = x_M^t G x_M = x_M^t F x_M - \frac{(x_M^t F x_M)^2}{V_M} \approx V_M - V_M = 0.
\]

Since it may be inconvenient to have a singular variance matrix for the residual factors, and since reduction of the number of residual factors facilitates computation, it is helpful to transform the \( J \) residual factors into \( J-1 \) factors with a nonsingular variance matrix. This is easily accomplished when \( x_{1n} = 1 \) for all firms, for then the residual factors may be defined as the \( J-1 \) factors:

\[
\begin{pmatrix}
\tilde{\gamma}_1^* \\
\vdots \\
\tilde{\gamma}_{J-1}^* \\
\tilde{\gamma}_J^*
\end{pmatrix} = \begin{pmatrix}
\tilde{\gamma}_1 \\
\vdots \\
\tilde{\gamma}_{J-1} \\
\tilde{\gamma}_J
\end{pmatrix},
\]

with loadings defined by:

\[
(21) \quad z_n = \begin{pmatrix}
z_{1n} \\
\vdots \\
z_{J-1,n} \\
z_{J,n}
\end{pmatrix} = \begin{pmatrix}
x_{1n} - x_{2n} \\
x_{2n} - x_{2M} \\
\vdots \\
x_{Jn} - x_{JM}
\end{pmatrix} \quad \text{for securities } n=1,\ldots,N.
\]
The variance matrix of $g^*$, defined as $G^*$, is the principal submatrix made up of the last $J-1$ rows and columns of $G$. That this is the appropriate transformation follows because, for any $m$ and $n$:

$$\text{COV}(x_n', g_m, x_n' g_m) = x_n' G x_n = (x_n' - x_M')' G (x_n' - x_M')$$ from (20)

$$= \begin{pmatrix} 1 & 0 \\ z_n' & z_m' \end{pmatrix} G \begin{pmatrix} 1 \\ z_m' \end{pmatrix} = z_n' G z_m'. $$

Thus, we have the simplifying assumption:

(A8): The Transformed Residual Returns Obey the Multiple Factor Model

For simplicity of exposition, the superscript (*) will be deleted below, and the transformed factors will replace the originals.

In summary, under assumptions (A1), (A2), (A4), (A5), and (A8), we have derived a model for market and residual returns with the following properties:

$$y = \alpha + \beta r_M + \upsilon$$
$$\alpha = X(\bar{c} - \Theta \bar{r}_M) + (\bar{p} - \gamma \bar{r}_M)$$ from (18)
$$\beta = X\bar{b} + \upsilon$$ from (16)
$$\upsilon = Z\bar{g} + u,$$

where $E(\bar{g}) = E(\bar{u}) = E(\upsilon) = 0$; $\text{COV}(\bar{g}, \bar{r}_M) = \text{COV}(\bar{u}, \bar{r}_M) = 0 = \text{COV}(\bar{g}, \bar{u})$; $\text{VAR}(\bar{g}) = \Sigma; \text{VAR}(\bar{u}) = \Sigma = \text{DIAG}(\sigma_n^2); \text{VAR}(\upsilon) = ZGZ' + \Sigma$; and

$$\sigma_u^2 = \sigma^2(x_i' s)$$ from (A5).

Under the further simplifying assumptions (A6) and (A7), the residual factors disappear and $\sigma_u^2$ is constant.
II.3 The Statistical Model for a Time Series of Cross Sections

This section resolves a number of difficulties that arise in combining numerous cross sections into a pooled time-series, cross-section model. One problem that is far more easily resolved here than in most econometric studies is that of serial correlation in the random variables. It may be assumed, to an excellent approximation, that there is none:

(A9): Absence of Serial Correlation in Specific Returns

For any \( s \neq t \), and for any \( m \) and \( n \), \( \text{COV}(u_{ms}, u_{nt}) = 0 \).

(A10): Absence of Serial Correlation in Factors

For any \( s \neq t \), and for any \( j, k, m \), \( \text{COV}(g_{js}, g_{kt}) = 0 = \text{COV}(g_{js}, u_{mt}) \).

The justification for these assumptions, of course, lies in the fact that serial correlation would create the opportunity to forecast returns and hence to earn a speculative profit. The absence of serial correlation in monthly residual returns, essentially equivalent to (A9), has been confirmed in a number of empirical studies. One reason for this is that the standard deviation of return on common stocks is remarkably large, relative to the mean. For a typical stock, the standard deviation of monthly return has been about .10 (10 percent of the original value of the holding), but mean returns have been on the order of .01 per month. As a consequence, serial correlation in returns as small as .02 would create an opportunity for significant speculative profit. In the sample of NYSE common stocks to be studied, where relatively active trading occurs and where floor traders pay negligible transaction costs, it is to be expected that speculative activity will adjust current prices so as to
anticipate any forecastable component of future return, thereby eliminating serial correlation.

Assumption (A10) on serial correlation in factors will be checked in the empirical study. The argument developed in the previous paragraph applies here as well, but it is weakened by two considerations: (1) the variance of factors is much smaller, so that the same degree of serial correlation would imply a smaller profit; and (2) speculation on a factor requires investment in a portfolio, rather than a single stock, and consequently involves greater transaction costs.

A more difficult problem arises from possible nonconstancy in the parameters of the cross-sectional model over time. Consider first the matter of constancy in the predictive relationships for the return parameters. For simplicity, the predictive relations for specific variance, total variance, and yield, will be assumed to exhibit constant parameters:

**(A11): Constant Parameters**

\[
\sigma_{nt}^2 = \sigma_t^2(s'x_{nt}) + \epsilon_{nt}^2 \\
\phi_{nt} = \phi_t (h'x_{nt}) + \epsilon_{\phi_{nt}} \quad n=1,\ldots,N(t) \\
\delta_{nt} = \delta_{t} (d'x_{nt}) + \epsilon_{\delta_{nt}},
\]

where the predictions are taken relative to the market average of the descriptor in period \( t \), with \( \sigma_t^2 = \sum_n w_{Mt} \sigma_{nt}^2 ; \phi_t = \sum_n w_{Mt} \phi_{nt} ; \delta_{Mt} = \sum_n w_{Mt} \delta_{nt} \). Note that these average values are contemporaneous
with the variables being predicted, so that the prediction rule is really a conditional prediction rule, conditional on the market average. This appears to be the appropriate context in which to analyze the predictive content of the descriptors, but it begs the question of forecasting the average values, an issue that will be taken up later. For the coefficients of the rule to be stationary over time, it is necessary that each descriptor also be taken as the difference between the historical value for the individual firm and the historical market average, and the descriptors will be defined in this way. These devices will reduce the variation over time in the optimal coefficients \( (s_t, d_t, h_t) \) but are certainly not enough to insure that the coefficients are actually constant. Thus, \( (A1) \) must be regarded as an undesirable assumption.

In the predictive relations for \( \alpha \) and \( \beta \), greater care will be taken to capture nonconstancy in the parameters. As a preliminary, note that these parameters are, by definition, relative to market averages, for, if \( (A2) \) is satisfied, then, from \( (1) \),

\[
\sum_n w_{Mt} \alpha_{nt} = 0, \quad \sum_n w_{Mt} \beta_{nt} = 1 \quad \text{for all } t.
\]

Therefore, as in the previous cases, it is appropriate that each descriptor be taken relative to the market average. When this is done, the constant terms in the prediction rules will not vary over time. To see this, note that:

\[
\sum_n w_{Mt} \hat{\alpha}_{nt} = \sum_n w_{Mt} \hat{\alpha}_{1t} + \sum_{j=2}^J \sum_n w_{Mt} (x_{jnt} - x_{Mnt}) \hat{\alpha}_{jt} = \hat{\alpha}_{1t}
\]
and, similarly, \[ \sum_{n} w_{\text{mt}} \hat{\beta}_{\text{nt}} = \hat{b}_{\text{lt}}. \]

Thus, in each period \( t \), the average predicted value for the parameter is identically equal to the estimated constant term. Therefore, if (A2) holds, and if the weights in the regression coincide with the weights in the market index, we should have that \( a_{\text{lt}} = 0, \quad b_{\text{lt}} = 1 \).

When the least squares regressions become weighted, as they will be to deal with heteroscedasticity, and when these weights do not coincide with the market weights, this property no longer holds exactly, and optimal estimates of these coefficients may vary slightly over time. We will assume, however, that the discrepancies are negligible, so that \( a_{\text{lt}} = a_{\text{l}}, \quad b_{\text{lt}} = b_{\text{l}} \) for all \( t \).

For the other descriptors, it is quite possible that the optimal coefficients will vary over time. This is particularly true because no great effort has been made to optimize the historical measures of risk \( H_{\text{gt}} \) and \( H_{\text{g}}^{2} \) as predictors of these parameters. The minimum mean square error prediction rule would weigh these descriptors less heavily in times when estimation error variance was larger relative to the variance of the parameter itself, as would occur during periods when the previous historical interval provided unusually little information to estimate the parameters. Thus, \( b_{\text{lt}} \) should vary as a function of the average variance of estimation error in the cross section, and this will vary over time. Second, the risk parameters presumably exhibit a tendency to revert toward the population norm. During a period of unusually rapid reversion, the coefficients \( b_{\text{lt}} \) would
decline in absolute value, since estimates of historical position would be
given less weight. In addition, the optimal prediction rules will be found
to include substantial weights on all of the descriptors, and there is ev-
ery probability that these weights may shift relative to one another over
time.

Lacking information on the timing and magnitude of these changes,
it is desirable to allow for a general form of parameter variation. The
form to be employed here permits the value of the parameter vector to
drift according to a first-order Markov process, with serially uncorre-
lated random increments. Transitory random additions to the vector, that
effect one period only and are serially independent, are added to the un-
derlying vector that is being propagated according to the Markov process.
For economy in the calculations, only the three primary descriptors, \( \mathbf{H}\beta \),
\( \mathbf{H}\sigma^2 \), and \( \sigma \), are given dispersed coefficients. Formally, the model for
\( \beta \) is:

\[
\beta_{nt} = b'x_{nt} + (a_b + e_b)'z_{nt} + \varepsilon_{nt}.
\]

(23)

Here, \( z_{nt} \) is the three-element column vector of primary descriptors.
For clarity, the six descriptors to be used, and the three primary de-
scriptors to be given dispersed parameters, are presented in table 1.
<table>
<thead>
<tr>
<th>$x_{1nt}$</th>
<th>Constant, identically equal to 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{2nt}$</td>
<td>$z_{1nt}$ $\bar{H}_t^\beta - \bar{H}_t^\gamma$</td>
</tr>
<tr>
<td>$x_{3nt}$</td>
<td>$z_{2nt}$ $\bar{H}_t^\omega - \bar{H}_t^\omega$</td>
</tr>
<tr>
<td>$x_{4nt}$</td>
<td>$z_{3nt}$ $\bar{H}_t^\phi - \bar{H}_t^\phi$</td>
</tr>
<tr>
<td>$x_{5nt}$</td>
<td>$\bar{H}_t^2 - \bar{H}_t^2$</td>
</tr>
<tr>
<td>$x_{6nt}$</td>
<td>$\bar{H}_t^\gamma - \bar{H}_t^\gamma$</td>
</tr>
</tbody>
</table>

The term $e^\beta_{nt}$ is a scalar random component in $\beta_{nt}$ that introduces those aspects of $\beta$ that are unrelated with the descriptors and uncorrelated with all other stochastic terms in the model, with $E(e^\beta_{nt}) = 0$, $\text{VAR}(e^\beta_{nt}) = \omega_\beta$, $\text{COV}(e^\beta_{nt}, e^\beta_{ms}) = 0$ for $m \neq n$ or $t \neq s$.

The vector $e^\beta_{bt} = (e^\beta_{\beta,t}, e^\beta_{\omega,t}, e^\beta_{\phi,t})'$ is a vector of random increments to the partial derivatives of $\beta$ with respect to the primary descriptors in period $t$. This vector is assumed to be independent of the descriptors, uncorrelated with all other stochastic terms, with $E(e^\beta_{bt}) = 0$, $\text{VAR}(e^\beta_{bt}) = \Omega_\beta$, $\text{COV}(e^\beta_{bt}, e^\beta_{bs}) = 0$ for $s \neq t$.

Finally, $e^\beta_{bt}$ is the underlying or permanent value of the random increment to the partial derivatives, with respect to the primary descriptors in period $t$, that is propagated according to a first-order Markov process:
\[ E(\hat{b}_t) = 0 ; \text{VAR}(\hat{b}_1) = \frac{\sigma_b^2}{(1-\phi_b^2)} \]

\[ \hat{b}_t = \phi \hat{b}_{t-1} + d_{t-1} \quad t=2,\ldots,T, \]

where \( E(d_{b_s}) = 0, \text{VAR}(d_{b_s}) = \sigma_b^2, \text{COV}(d_{b_r}, d_{b_s}) = 0 \) for \( r \neq s \). The \( d_{b_s}, s=1, T-1 \) are serially independent increments that are uncorrelated with all other stochastic terms and independent of the descriptors. The scalar \( \phi_b, 0 < \phi_b \leq 1 \) introduces a reversion toward zero in the process.

Finally, \( b_\cdot \) is the grand mean value of the partial derivatives of \( \beta \) with respect to the descriptors, over the entire time series of cross sections. For the primary descriptors, \( b_j \) gives the normal value about which the coefficient drifts.

To clarify the function of these different random terms, notice, first, that if all random terms are identically zero, the model asserts that the prediction rule for \( \beta \) is linear and constant, with coefficient vector \( \hat{b} \). The term \( \hat{\beta}_{nt} \) introduces a random and unpredictable component in \( \beta \), that is not correlated across stocks or over time, and that will consequently show up only as a component of variance in individual returns. The term \( \hat{\beta}_{nt} \) is serially uncorrelated, so it will show up as a distinct component of variance in each cross section, multiplying the primary descriptors and thereby inducing a correlation in the residuals that is proportional to the values of the descriptors. The random term \( \hat{b}_t \) introduces a serially correlated element in the partial derivatives that will show up as serial correlation in the contribution of \( r_{nt}X_{njnt} \).
Turning next to the model for $\alpha$, the CAP hypotheses suggest that $\alpha$ may be functionally dependent on the other return parameters, implying that $\alpha$ would be predicted by descriptors having predictive content for those parameters. Thus, in each cross section $t$,

$$\alpha_{nt} = a^t x_{nt} + e_{nt}^t, \quad n=1, \ldots, N(t)$$

where $e_{nt}^t$ introduces that component of $\alpha$ that is uncorrelated with (unpredictable by) the descriptors in that cross section. Will the predictive relationship be changing over time, so that $a^t$ is not constant? Examination of the CAP hypotheses where $a^t = B^t c^t$ suggests that changes in $c^t$ may occur in response to changes in the tax rate, the expected variance of market return, the risk-aversion of the market, or the ex ante excess return in the "zero beta" portfolio $k^t$. Also, changes in $B^t$ may result from variations over time in the predictive content of the descriptors, such as the variation in $b^t$ introduced above.

Consequently, nonstationarity in $a^t = B^t c^t$ is to be expected. It is again possible that this nonstationarity may take the form of random drifting in these parameters over time, on top of which there might be random nonpersistent shocks in each period. Thus, a natural model is again of the form:

$$\alpha_{nt} = a^t x_{nt} + (a^t + e^t a^t) z_{nt} + e^t_{nt}$$

where:

$$E(*a_{-1}) = 0; \quad \text{VAR}(*a_{-1}) = \frac{Q_a}{(1-\phi_a^2)}$$

$$*a_t = \phi a_{t-1} + d_{t-1}, \quad t=2, \ldots, T$$
with \( \mathbb{E}(e_a^r) = 0 \), \( \text{VAR}(e_a^r) = \Omega_a \), \( \text{COV}(e_a^r, e_a^s) = 0 \) for \( r \neq s \)

\[
\mathbb{E}(e_a^s) = 0, \quad \text{VAR}(e_a^s) = \Omega_a, \quad \text{COV}(e_a^s, e_a^r) = 0 \quad \text{for} \quad r \neq s
\]

\[
\mathbb{E}(e_{\alpha_n}^t) = 0, \quad \text{VAR}(e_{\alpha_n}^t) = \omega_{\alpha_n}, \quad \text{COV}(e_{\alpha_n}^t, e_{\alpha_n}^s) = 0 \quad \text{for} \quad m \neq n \quad \text{or} \quad s \neq t.
\]

and where the processes \( e_{\alpha_n}^t \), \( d_{\alpha}^t \), and \( e_{\alpha}^t \) are independent of one another and of the descriptors.

As before, \( \alpha_t^* \) represents the underlying random component of \( \alpha_t \) that is propagated over time by the addition of random shifts, \( d_{\alpha}^t \), so that it obeys a Markov process; \( e_{\alpha}^t \) represents a random increment to the underlying component in period \( t \) that is independent of events in other periods; and \( \phi_{\alpha}^t, 0 < \phi_{\alpha}^t \leq 1 \) introduces a tendency for the random vector \( \alpha_t^* \) to revert to zero.

To complete the model for \( \alpha \), notice that the possibility of misspecification of the appropriate riskless borrowing rate introduces another particular form of random variation. Let the difference between the posited riskless rate, \( i_{Ft}^i \), and the actual rate, \( i_{Zt}^Z \), defined as \( k_t = i_{Zt}^Z - i_{Ft}^F \), be written as:

\[
k_t = \bar{k} + \star_k^t,
\]

where:

\[
\mathbb{E}(\star_k^t) = 0, \quad \text{VAR}(\star_k^t) = q_k^t/(1-\phi_k^t) \]

\[
\star_k^t = \phi_k^t \star_k^{t-1} + d_k^{t-1}, \quad t=2, \ldots, T
\]

with \( \mathbb{E}(d_k^s) = 0 \), \( \text{VAR}(d_k^s) = q_k^s \), \( \text{COV}(d_k^s, d_k^r) = 0 \) for \( s \neq r \), and where
the \( d_k \) process is independent of the descriptors. In this model, the transitory random shocks are omitted, because the fact that \( k_t \) is a spread between two short-term rates suggests that monthly values will be serially dependent to a high degree. The term \( \phi_k, 0 < \phi_k < 1 \) introduces a tendency for \( k_t \) to converge toward zero or, equivalently, for \( k_t \) to converge toward the grand mean value \( \bar{k} \). This again seems plausible since variations in the spread about its normal value would presumably result from temporary conditions in the money markets that might be expected to disappear eventually.

Misspecification of \( i_zt \) leads to a component \( k_t (1-\beta_{nt}) = (k_t + \beta_{nt})(1-\beta_{nt}) \) in \( a_{nt} \). The term, \( k(1-\beta_{nt}) = k(b_{znt}^{\ast} + b_{znt}^{\ast \ast} + b_{znt}^{\ast \ast \ast} + \beta_{nt}) \), is absorbed into the model for \( a_{nt} \) developed previously. Addition of the remaining term to the expression for \( a_{nt} \) yields the complete model:

\[
\alpha_{nt} = a_{nt}^{\ast} + b_{nt}^{\ast \ast \ast} z_{nt} + k_t (1-\beta_{nt}) + \epsilon_{nt}.
\]

The random components in \( \alpha \) and \( \beta \) should be related, according to the CAP hypotheses, just as the systematic components are related. The exact form of the relationship depends on the hypothesis. For instance, under CAP:Z, these random components are linearly related and hence exactly correlated; under the CAP:M, random components in \( \alpha \) are identically zero, and the values for \( \beta \) are unconstrained; under the other CAP hypotheses, the random component for \( \alpha \) is related to random components in return parameters other than beta and hence is less than perfectly correlated with the random component in beta. Thus, the covariance between \( \epsilon_{a_t} \) and \( \epsilon_{b_t} \), between \( d_{a_t} \) and \( d_{b_t} \), and between
\( e_{\alpha nt} \) and \( e_{\beta nt} \) may be nonzero. Let this covariance be denoted by:

\[
(25) \quad \text{COV}(e_{\alpha nt}, e_{\beta nt}) = \omega_{\alpha\beta}, \quad \text{COV}(e_{\alpha t}, e_{\beta t}) = \Omega_{ab}, \quad \text{COV}(d_{\alpha t}, d_{\beta t}) = \Omega_{ab}.
\]

These models of nonconstancy in \( \alpha \) and \( \beta \) may be collected into:

\[\text{AL2): Stochastic Variation in} \quad \alpha \quad \text{and} \quad \beta\]

The behavior of \( \alpha_{nt} \) and \( \beta_{nt} \) is given by equations (23), (24), (25).

Finally, the model for \( v_{nt} \) is taken directly from the market model derived from the multiple factor model:

\[
v_{nt} = g_{zt}^{t} + u_{nt},
\]

where \( E(g_{zt}) = 0, \quad \text{VAR}(g_{zt}) = \pi, \quad \text{VAR}(u_{nt}) = \sigma_{nt}^{2}, \) and \( \text{COV}(g_{zt}, u_{nt}) = 0. \)

Note that a simplification has occurred, with only the three primary descriptors appearing as the loadings on random factors, rather than all five nonconstant descriptors. Combining the models for \( \alpha, \beta, \) and \( v, \) the pooled time-series, cross-sectional model for returns can now be completed.

In the formalism of the model for market and residual returns,

\[
r_{nt} = \alpha_{nt} + \beta_{nt} r_{Mt} + v_{nt}.
\]

Substitution of the expression for \( \alpha_{nt} \) yields:

\[
= a'_{zt} x_{nt} + (a'_{zt} + e_{\alpha t})' z_{nt} + e_{\alpha t} + k_{t} (1-\beta_{nt}) + r_{Mt} + \beta_{nt} r_{Mt} + v_{nt}.
\]

Substitution of the expression for \( \beta_{nt} \) yields, after rearrangement of terms:
\[ \begin{align*} 
&= a' x_{nt} + b' x_{nt} r_{Mt} - (1-b' x_{nt}) k_t + a' z_{nt} + b' z_{nt} (r_{Mt} - k_t) \\
&+ e_{a't} z_{nt} + e_{b'z} z_{nt} (r_{Mt} - k_t) + e_{a't} + e_{b'z} (r_{Mt} - k_t) + v_{nt}. 
\end{align*} \]

Substitution of the multiple-factor model of residual returns yields, after further rearrangement of terms:

\[ \begin{align*} 
&= a' x_{nt} + b' x_{nt} r_{Mt} + a' z_{nt} + b' z_{nt} r_{Mt} + k_t (1-b' x_{nt}) + e_{a't} z_{nt} + e_{b'z} z_{nt} \\
&+ e_{b'z} r_{Mt} + e_{a't} + u_{nt} + e_{b'z} r_{Mt} - k_t \left( \left( b' + e_{b'z} \right) z_{nt} + e_{b'z} \right). 
\end{align*} \]

With the combination of the separate models, it is possible to simplify the structure. Notice that \( e_{a't} \) appears at the same point in the model as \( e_{b'z} \), and that \( a_{nt} \) appears identically to \( u_{nt} \). This occurs because the terms are indistinguishable observationally: the terms \( e_{a't} \) and \( e_{a't} \) are aspects of expectations that cannot be predicted on the basis of the descriptors; the terms \( e_{b'z} \) and \( u_{nt} \) are random components of return with \( ex \ ante \) expected value of zero. The two are indistinguishable for the simple reason that we cannot tell how much of the unpredictable return was anticipated in unobservable expectations (and, hence, belongs in \( \alpha \)) and how much was truly unanticipated (and, hence, belongs in \( v \)).

Moreover, the variance of the unpredictable random component in \( v \) completely swamps any conceivable variance in the warranted expected
returns.\(^\text{2}\)

Also, \(\text{COV}(e_{\beta t}, e_{\alpha t} + v_{nt}) = \text{COV}(e_{\beta t}, e_{\alpha t})\) and \(\text{COV}(e_{\beta t}, e_{\alpha_t} + g_t)\) are negligible, relative to the variance of the random components in \(v\) and may be ignored. Therefore, in the following, \(e_{\alpha_t} + v_{nt}\) will be redefined as \(v_{nt}\), \(e_{\alpha t} + g_t\) will be redefined as \(g_t\), and the covariance of the residual returns with the random process in \(\beta\) will be set to zero.

\(^2\) Experience in previous studies (Rosenberg et al. (1973)) suggests that a typical random factor \(\gamma_j\) may contribute a standard deviation of monthly return on the order of 2 percent when multiplied by a typical value for the descriptor. Thus, 
\[
(\text{VAR}(\gamma_j)E(z_j^2)) \approx (.02)^2.
\]
Similarly, a typical value for the standard deviation of monthly specific return is about .10, implying that 
\[
(\text{VAR}(v_{nt})) \approx (.10)^2.
\]
By contrast, \(e_{\alpha_t}\) and \(e_{\alpha nt}\) introduce fluctuation in equilibrium values of expected return that are warranted by expectations of the other return parameters. The magnitude of the fluctuations in these parameters is therefore determined by the degree of variation in the other return parameters. The standard deviation of expected variation in \(\beta\), unrelated to the descriptors, might conceivably be as great as \(.40 - (\omega_\beta = (.40)^2)\)--and the variations in the coefficients of the descriptors might typically account for a standard deviation of as much as \(.20 - (\text{VAR}(e_{\beta jt})E(z_{jnt}^2)) = (.20)^2\). But even allowing for a major violation of the CAPM, in that the partial derivative of expected return differs from the theoretical value \(E(r_{Mt})\), by a factor of 2, we would still have \(e_{\beta} - E(r_{Mt}) \sim .007\). Thus, when this coefficient is multiplied times the variation in \(\beta\), we find a maximum contribution to the variance of expected return of 
\[
\text{VAR}(e_{\alpha nt}) = (.007)^2(.40)^2 = (.003)^2
\]
and 
\[
\text{VAR}(e_{\alpha jt}) = (.007)^2(.20)^2 = (.001)^2.
\]
Even if variance of the same order of magnitude is also contributed by the other return parameters, so that this variance is multiplied by 4, the variance in \(\alpha\) remains less than 1/100 of the variance in \(v\), and hence is totally negligible.
The next simplification follows from noting that the last term of the expanded model, 
\[ k_t^* \left( (b_t + e_{bt})'z_{nt} + e_{bt}^* \right) \] 

is the product of the random variation in \( k_t^* \) and the random variation in \( e_{bt}^* \). The mean value of this term is zero, and the variance is small, relative to earlier terms that are linear in \( k_t^* \) or in \( b_t^* \), \( e_t^* \), or \( e_t^* \). Therefore, it will be deleted. After these simplifications are incorporated, the model becomes:

\[
r_{nt} = a'x_{nt} + b'x_{nt} r_{Mt} \\
+ a'z_{nt} + b'z_{nt} r_{Mt} + k_t^*(1-b'x_{nt}) \\
+ g_{nt}^* + e_{nt}^* r_{Mt} \\
+ u_{nt} + e_{nt}^* r_{Mt}^*.
\]

The standard deviation of \( k_t^* \), a spread in short-term rates, is unlikely to be greater than 1.2 percent at an annual rate, or .001 per month. This yields \( \text{VAR}(k_t^*) = .000001 \). By contrast, \( E(r_{Mt})^2 \) is on the order of .002, or 2,000 times as great, so that the term \( k_t^*(b_t + e_{bt})'z_{nt} + e_{bt}^* \) is negligible, relative to the random term \( r_{Mt}(b_t + e_{bt})'z_{nt} + e_{nt}^* \). This term is also small relative to \( k_t^*(1-b'x_{nt}) \), since \( (1-b'x_{nt}) \), the explained component of \( \beta \), will have larger variance than the unexplained component, but in this case, the term to be deleted is of the same order of magnitude as the one to be retained. The distinction that suggests that the product of the random terms be discarded is that it contributes a component of variance that is effectively undistinguishable from the other components in the model multiplying \( (z_{Mt} - k) \), whereas the term \( k_t^*(1-b'x_{nt}) \) is the only random coefficient multiplying \( (1-b'x_{nt}) \) and, therefore, may conceivably be identifiable despite the very small variance of \( k_t^* \).
The terms in the first line introduce components of return that are related to the descriptors by constant coefficients; terms in the second line introduce coefficients multiplying the descriptors that drift over time; terms in the third line are serially uncorrelated random coefficients multiplying the descriptors; terms in the fourth line are random terms, uncorrelated across the cross section as well as over time.

Now, $g_t$ and $e_b^t r_{Mt}$ are both random components multiplying $z_{nt}$. For our purposes, there is no importance in the distinction between the two nor in the distinction between $u_{nt}$ and $e_b^t r_{Mt}$. The model may be rewritten as:

\begin{equation}
\begin{aligned}
r_{nt} &= a' x_{nt} + b' x_{nt} r_{Mt} \\
&\quad + e_t^l z_{nt} + e_t^b z_{nt} r_{Mt} + \kappa_t (1-b' x_{nt}) \\
&\quad + g_t' z_{nt} + u_{nt},
\end{aligned}
\end{equation}

where $g_t$ is redefined to include $e_b^t r_{Mt}$, so that $\text{VAR}(g_t) = C + r_{Mt-b}$ and $u_{nt}$ is redefined to include $e_b^t r_{Mt}$, so that $\text{VAR}(u_{nt}) = \sigma_{nt}^2 + r_{Mt-w}^2$.

The variance of the redefined terms $g_{nt}$ and $u_{nt}$ now includes $r_{Mt-b}^2 \omega_b^2$, respectively, so that it varies as a function of the squared market return. The term $r_{Mt}^2$ may be absorbed in the conditional prediction rule of $\sigma_{nt}^2$, in which case, one cause of fluctuations in the market average residual variance, $\sigma_t^2$, will then be the fluctuations in the squared market return. Thus, the final version of the model is given by (27), with
\[
\begin{align*}
\text{E}(\mathbf{a}_1) &= 0; \quad \text{VAR}(\mathbf{a}_1) = \frac{1}{1 - \phi_a^2} q_a \\
\mathbf{a}_t &= \phi_a \mathbf{a}_{t-1} + \mathbf{d}_{a_{t-1}}, \quad t=2, \ldots, T \\
E(\mathbf{d}_{a_s}) &= 0; \quad \text{VAR}(\mathbf{d}_{a_s}) = q_a; \quad \text{COV}(\mathbf{d}_{a_s}, \mathbf{d}_{a_t}) = 0 \quad \text{for} \ s \neq t \\
\text{E}(\mathbf{b}_1) &= 0; \quad \text{VAR}(\mathbf{b}_1) = \frac{1}{1 - \phi_b^2} q_b \\
\mathbf{b}_t &= \phi_b \mathbf{b}_{t-1} + \mathbf{d}_{b_{t-1}}, \quad t=2, \ldots, T \\
E(\mathbf{d}_{b_s}) &= 0; \quad \text{VAR}(\mathbf{d}_{b_s}) = q_b; \quad \text{COV}(\mathbf{d}_{b_s}, \mathbf{d}_{b_t}) = 0 \quad \text{for} \ s \neq t \\
\text{E}(\mathbf{k}_1) &= 0; \quad \text{VAR}(\mathbf{k}_1) = \frac{1}{1 - \phi_k^2} q_k \\
\mathbf{k}_t &= \phi_k \mathbf{k}_{t-1} + \mathbf{d}_{k_{t-1}}, \quad t=2, \ldots, T \\
E(\mathbf{d}_{k_s}) &= 0; \quad \text{VAR}(\mathbf{d}_{k_s}) = q_k; \quad \text{COV}(\mathbf{d}_{k_s}, \mathbf{d}_{k_t}) = 0 \quad \text{for} \ s \neq t \\
\text{E}(\mathbf{g}_{s_t}) &= 0; \quad \text{VAR}(\mathbf{g}_{s_t}) = G + \frac{2}{\varphi_{Mt-b}} \\
\text{COV}(\mathbf{g}_s, \mathbf{g}_t) &= 0 \quad \text{for} \ s \neq t \\
\text{E}(\mathbf{u}_{nt}) &= 0; \quad \text{VAR}(\mathbf{u}_{nt}) = \sigma^2_{nt} = \sigma^2_t \delta_n^t \delta_n^t; \\
\text{COV}(\mathbf{u}_{ms}, \mathbf{u}_{nt}) &= 0 \quad \text{for} \ m \neq n \text{ or } s \neq t
\end{align*}
\]
and with the different stochastic processes (28) through (32) independent of one another and of the descriptors.

For many purposes, it is more convenient to write the model as a multiple-factor model. In matrix form, we have the equivalent model:

\[ z_t = X f_t + \rho_t \]

where:

\[ f_t = a + b_r \frac{r_{Mt}}{\sim_t} + \frac{1 - b_1}{k_t} \begin{pmatrix} 1 \\ b_2 \\ \vdots \\ -b_j \end{pmatrix} \]

\[ + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} a_t \\ b_t \end{pmatrix} + b_t r_{Mt} \]

(33)

and where:

\[ \text{VAR}[\rho_t] = \Sigma_t = \text{DIAG}[\sigma^2_{nt}] \]

\[ \text{COV}[f_t, \rho_t] = 0. \]

Comparing (27) and (33), it is apparent that the market and residual return model, incorporating the complicated stochastic structures for \( \alpha, \beta, \) and \( \nu, \) is equivalent to a multiple-factor model, with a stochastic model for the factors as a function of the market return. Statistical inference may be carried out in whichever context is the more convenient.
III. STATISTICAL INFERENCE

Section III.1 examines hypothesis tests on the system of prediction rules for all parameters and presents estimation methods for the prediction rules for \( \sigma^2 \), \( \phi \), and \( \delta \). Next, under the simplifying assumption that there is no sequential variation in \( a \) and \( b \) \( (Q_a = Q_b = 0) \), three seemingly different but equivalent approaches to estimating the prediction rules for \( \alpha \) and \( \beta \) are developed in Sections III.2 through III.4. The equivalence of these approaches provides a perspective that is used in developing the estimation method for the problem in its full complexity (in III.5) and in discussing the previous literature (in Section V, below).

III.1 The System of Prediction Rules

Because of the irregular distribution of dividends over a calendar year, yield is most well behaved when it is observed on a calendar-year basis. Thus, the calendar year is a natural unit of observation when estimating the prediction rule for yield. Let \( y_{ns} \) denote the yield on stock \( n \) in calendar year \( s \). All other return parameters relate to the behavior of monthly security returns. Therefore, the set of all observed dependent variables on which the estimated prediction rules will be based is \( \{r_{nt}, y_{ns}\} \), where the brackets denote the set of all items. The probability distribution of these variables, as a function of the parameters of the model for returns, conditional on the observed predetermined values of the descriptors, may be written as:
(33) \[ P(\{x_{nt}\}, \{y_{ns}\} | \{x_{nt}\}; a, b, s, h, d; \phi_k, \phi_a, \phi_b, q_k, q_a, q_b, \omega_a, \omega_b, \omega_c, G) \]

or, conditional on the observed market return as well, as:

(34) \[ P(\{x_{nt}\}, \{y_{ns}\} | \{x_{nt}\}, \{r_{Mt}\}; a, b, s, h, d; \phi_k, \phi_a, \phi_b, q_k, q_a, q_b, \omega_a, \omega_b, \omega_c, G) \]

Now, the covariance between the errors in prediction of the annual yields, on the one hand, and the random components in the monthly returns to the securities in the same year, on the other, will not be zero. Nevertheless, it will be convenient to assume away this joint dependence so that (34) may be rewritten as follows:

(A13): Independence of Realized Yields and Total Return

The probability distribution may be decomposed as:

(35) \[ P(\{x_{nt}\} | \{x_{nt}\}, \{r_{Mt}\}; a, b, s, h; \phi_k, \phi_a, \phi_b, q_k, q_a, q_b, \omega_a, \omega_b, \omega_c, G) \]

\[ \times P(\{y_{ns}\} | \{x_{nt}\}, d) \]

This assumption greatly facilitates the testing of hypotheses, since the prediction rule for yield may be estimated independently.

The prediction rule for yield is estimated using the formulation:

(36) \[ (y_{ns} - y_{Ms})/y_{Ms} = d'x_{ns} + \epsilon_{y_{ns}}, s=1931, \ldots, 1966 ; n=1, \ldots, N(s) \]

where \( y_{Ms} \) is the average yield in the market in year \( s \), and where the yield on the individual stock in period \( s \) is divided by the market yield in an effort to achieve homoscedasticity in the residuals. This
regression is run under the simplifying assumption of homoscedastic and
independent residuals to obtain an estimate \( \hat{\sigma} \) and estimated estimation
error variance matrix \( M_q \).

Next, note that the prediction rule for total variance, \( \phi = h'x \),
is redundant—and misspecified as well—in that total variance is deter-
mined as a quadratic function of the other parameters \( \beta \) and \( \sigma^2 \). The use
of \( h \) must be considered a simplistic device to obtain a convenient linear
prediction rule for purposes of hypothesis testing. Since there is no
justification for this approach anyway, it is natural to make the equally
unforgivable simplification—that this prediction rule can be estimated in-
dependently of the prediction rules for \( \alpha \) and \( \beta \).

(A14): Independent Estimation of Total Variance

Estimation error in the prediction rule, \( \phi = h'x \), is independent
of error for \( \alpha \) and \( \beta \).

The estimation method for this prediction rule will be taken up
later.

With the above simplifications, we are left with estimation of the
\( \alpha, \beta, \) and \( \sigma^2 \) processes. From (33), it is seen that the specification of
the \( \alpha \) and \( \beta \) processes enters as a stochastic constraint on the series
of factors \( \epsilon_t \), \( t=1,\ldots,T \), whereas the \( \sigma^2 \) process enters as the specifi-
cation of heteroscedasticity in the residuals about these factors. For
maximum likelihood estimation, we will make use of the assumption:

(A15): Normality

The stochastic terms in the probability distribution of returns
are jointly, normally distributed. (As will be seen below, this assumption is much more closely satisfied than one might expect from the literature on the heavy-tailed distributions of security returns.)

Imagine, first, that the factors $f_{tT}$, $t=1, \ldots, T$ are unconstrained. Then, under the normality assumption, the maximum likelihood estimators for $\hat{s}$, for $f_{1T}, \ldots, f_{T}$, and for $\sigma^2_1, \ldots, \sigma^2_T$ maximize the logarithm of the likelihood:

$$
\text{MAX} \left\{ -\frac{1}{2} \sum_t \left( N(t) \ln(2\pi) + \sum_n \ln(\sigma^2_{t \sim nt}) + \sum_n \frac{(r_{nt} - f'_{t \sim nt})^2}{\sigma^2_{t \sim nt}} \right) \right\}.
$$

The solutions for $f_{1T}, \ldots, f_{T}$, given an arbitrary initial estimator $\hat{s}$ for $s$, are:

$$
\hat{f}_t = \left( \sum_n \frac{x_{nt \sim nt}}{s'_{t \sim nt}} \right)^{-1} \left( \sum_{n \sim nt} \frac{x_{nt \sim nt}}{s'_{t \sim nt}} \right).
$$

These are just the Aitken's Generalized Least Squares Estimators, with heteroscedasticity as predicted by $\hat{s}$. If, on the other hand, a set of estimators $\hat{f}_{1T}, \ldots, \hat{f}_{T}$ are available, then the maximum likelihood estimators for $s$ and $\sigma^2_1, \ldots, \sigma^2_T$ maximize:

$$
\text{MAX} \left\{ -\frac{1}{2} \sum_t \left( N(t) \ln(2\pi) + \sum_n \ln(\sigma^2_{t \sim nt}) + \sum_n \frac{\hat{\rho}^2_{nt}}{\sigma^2_{t \sim nt}} \right) \right\},
$$

where $\hat{\rho}_{nt} = r_{nt} - \hat{f}'_{t \sim nt}$. This is just the condition for maximum likelihood estimation of the linear prediction rule for the residual return.
variance, $\text{VAR}(\hat{\rho}_{nt}) = \sigma^2 s' x_{nt}$. This representation of the problem points up a possible iterative procedure: first, estimate the factors by Generalized Least Squares (38), using an arbitrary initial estimator for $\hat{s}$; then, using the resulting series of estimators for the factors to define the residual returns, reestimate $\hat{s}$ by maximum likelihood (39), and begin another iteration.

For reasons of computational economy, we have substituted a two-pass linear regression for maximum likelihood estimation of $\hat{s}$. This approach behaved well in earlier studies by Rosenberg (1972) and Rosenberg and McKibben (1973). Amemiya (1973) has shown the following:

Suppose that $\rho_{nt}$ is distributed with mean zero and variance $s' x_{nt}$, such that $\rho_{nt}^2$ obeys a gamma distribution (in particular, $\rho_{nt}$ may obey a normal distribution, in which case $\rho_{nt}^2$ follows the $\chi^2_1$ distribution, a special case of the gamma). Then, the following two-pass procedure yields a consistent estimator of $\hat{s}$, with the same asymptotic variance matrix as the maximum likelihood estimator:

**Pass 1:** Obtain a consistent least squares estimator of $\hat{s}$, $\hat{s}$, by:

$$\hat{s} = \left( \sum_{n,t} x_{nt} x_{nt}' \right)^{-1} \sum_{n,t} x_{nt} \rho_{nt}^2.$$

**Pass 2:** Obtain the desired estimator, $s^*$, by the Generalized Least Squares Regression:

$$s^* = \left( \sum_{n,t} \tilde{x}_{nt} \tilde{x}_{nt}' \right)^{-1} \sum_{n,t} \tilde{x}_{nt} \tilde{\rho}_{nt}^2,$$

where $\tilde{x}_{nt} = \frac{x_{nt}}{\hat{s}' x_{nt}}$, $\tilde{\rho}_{nt}^2 = \frac{\rho_{nt}^2}{\hat{s}' x_{nt}}$. 

The estimator $\hat{s}^* - \tilde{s}$ is asymptotically normally distributed with mean $\tilde{s}$ and variance matrix $\text{VAR}(\hat{s}^*) = \mathbb{M}_{\hat{s}^*} = \eta^2 \left( \sum_{n, t} x_{nt} x'_{nt} \right)^{-1}$. Here, $\eta^2 + 1 = \frac{E(\rho^4)}{\text{VAR}(\rho)^2}$ is the kurtosis of the distribution of $\rho$, and $\eta_{2n}^2 = \frac{1}{N(t)} \sum_{n, t} (\hat{\rho}_{nt}^2 - \hat{s}^* x_{nt})^2$ is a consistent estimator of $\eta^2$. This two-pass approach has the advantage that all of the statistics are output by conventional regression packages. Further, the procedure is as efficient as maximum likelihood estimation for nonnormal distributions of a certain class, and directly yields an estimator of the kurtosis of the underlying distribution.

In the present application, two aspects of the specification differ from the pure case with which Amemiya dealt. One is that a multiplicative term $\sigma_t^2$ appears in each period, so that the model is nonlinear. Thus, for our sample, with 432 time periods, there are 432 multiplicative dummy variables to be estimated, as well as the six coefficients in $\tilde{s}$. However, with more than 700 observations for each coefficient to be estimated, and with no collinearity among the multiplicative dummies, we can be fairly confident about the asymptotic approximation.

The second divergence is, in fact, the result of a misspecification in the linear prediction rule. The prediction $\hat{\sigma}_{nt}^2 = \sigma_t^2 x_{nt} x'_{nt}$ can go to zero or even go negative for extreme values of the descriptors. This is clearly a misspecification and should be remedied by selecting an alternative functional form for the prediction rule. We have temporized by truncating the predictions at one-ninth and nine times the average value so that:
This truncation is used in all applications of the predictions, including the weighted least squares in Pass 2 of Amemiya's method and other Generalized Least Squares regressions below. However, the truncation is not reflected in the regression that fits the prediction rule, and this introduces an undesirable discrepancy. Fortunately, the truncation applies to only a small fraction of the observations, so neglect of this problem has little effect.

In estimating the prediction rule for $\hat{\sigma}_{nt}^2$, the following steps were followed:

1. Obtain an initial estimator for $\hat{s}$, with residual returns defined as $\hat{\rho}_{nt} = r_{nt} - r_{Mt}$, from (40).

2. Using the implied prediction rule, truncated as in (42), obtain an initial series of estimators of the factors, $\hat{f}_{1t}, \ldots, \hat{f}_{kt}$, from (38).

3. Using the residual returns defined as $\hat{\rho}_{nt} = r_{nt} - \hat{f}_{t}^t X_{nt}$, obtain Pass 1 estimator for $\hat{s}$, by (40).

4. With residual returns again defined as under step (3), obtain the final Pass 2 estimator of $\hat{s}$, by (41), using the truncated prediction rule yielded by Pass 1.

Asymptotically

Notice that, since these estimators are/equivalent to maximum likelihood estimators of the residual variances in regressions with the unknown
parameters \( f_t \), the estimated variances are downward biased, with bias of the order of \( 6/N(t) \) (since there are 6 unknown parameters to be estimated in each period and \( N(t) \) observations). For our sample, this bias will be about 1/100, so it may be neglected.

Returning to the prediction rule for \( \phi \), it is estimated by a direct application of steps (3) and (4), above, with the total returns \( r_{nt} \) substituted for the residual returns.

We have thus obtained estimators and variance matrices of estimation errors for the prediction rules for \( \phi, \sigma^2, \) and \( \delta \). In the ensuing sections, procedures to obtain estimates for the prediction rules for \( \alpha \) and \( \beta \) will be developed. For the final hypothesis tests, it will be assumed that estimation errors for \( \phi \) and \( \delta \) are independent of one another and of estimation errors for \( \alpha \) and \( \beta \). It will be assumed that all estimators are normally distributed with the estimated means and variance matrices, an assumption justified by asymptotic properties of the estimators in view of the immense sample. The likelihood of any hypothesized values of these prediction rules, \( \hat{\alpha}_0, \hat{\beta}_0, \hat{\gamma}_0, \hat{\delta}_0 \), will be found by computing the likelihood that these values would be drawn from normal distributions with the specified means and variances. Specifically, the log of the likelihood of the hypothesized values will be proportional to:
III.2 Estimating $\alpha$ and $\beta$ through "Market-Conditional" Regressions

To simplify the exposition in this and the next two sections, the serial correlation in $\alpha$, $\beta$, and $k$ introduced by the sequentially varying components $x_{\alpha}$, $x_{\beta}$, $x_{k}$ will be deleted, and the serially independent residual random factors will be retained in a general form, with

$$\text{VAR}(g_{t}) = G_{t}^{-1}G_{t} \quad \text{and} \quad g_{t} \quad \text{independent of the predetermined descriptors,}$$

the market return, and the residual returns $\rho_{t}$.

The model for returns becomes:

$$\Sigma_{t} = X_{t}a + (r_{M_{t}}X_{t})b + X_{t}g_{t} + \rho_{t}$$

(44)

$$= X_{t}a + (r_{M_{t}}X_{t})b + v_{t}, \quad \text{where} \quad v_{t} = X_{t}g_{t} + \rho_{t}.$$  

Therefore,

$$E(v_{t}) = 0, \quad \text{VAR}(v_{t}) = V_{t} = X_{t}G_{t}X_{t}^{\prime} + \Sigma_{t}, \quad \text{COV}(v_{t}, X_{t}) = \text{COV}(v_{t}, r_{M_{t}}X_{t}) = 0.$$  

This model expresses the returns as a linear function of the unknown coefficients $a$ and $b$, with observable regressor matrices $X_{t}$ and $r_{M_{t}}X_{t}$, and with "disturbances" $v_{t}$ that are uncorrelated with the regressors.

Thus, conditional on the observed values of the descriptors and the market
returns, we have an ordinary linear regression model with heteroscedastic disturbances.

Conditional on the predetermined descriptors and the observed market return, the minimum variance linear unbiased estimators of \( \hat{a} \) and \( \hat{b} \) are given by Generalized Least Squares. The model may be rewritten as:

\[
\begin{align*}
\tilde{r}_{t} &= \tilde{X}_{t} (\tilde{I}: \tilde{r}_{Mt}) \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} + \tilde{\nu}_{t}, \\
&= t = 1, \ldots, T.
\end{align*}
\]

Since the disturbances are serially uncorrelated, by (A9) and (A10), the Generalized Least Squares Estimator is found to be:

\[
\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \left( \sum_{t} \begin{pmatrix} 1 & \tilde{r}_{Mt} \\ \tilde{r}_{Mt} & \tilde{r}_{Mt} \end{pmatrix} \right)^{-1} \left( \begin{pmatrix} 1 \\ \tilde{r}_{Mt} \end{pmatrix} \tilde{X}_{t}^{-1} \tilde{X}_{t} \end{pmatrix}^{-1} \left( \sum_{t} \begin{pmatrix} 1 & \tilde{r}_{Mt} \\ \tilde{r}_{Mt} & \tilde{r}_{Mt} \end{pmatrix} \right) \right)
\]

\[(45)\]

where \( \otimes \) denotes the Kronecker product of two matrices. Also, the variance matrix of estimation errors is:

\[
\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \tilde{M}_{a} & \tilde{M}_{ab} \\ \tilde{M}_{ba} & \tilde{M}_{b} \end{pmatrix} = \left( \sum_{t} \begin{pmatrix} 1 & \tilde{r}_{Mt} \\ \tilde{r}_{Mt} & \tilde{r}_{Mt} \end{pmatrix} \right)^{-1} \left( \sum_{t} \begin{pmatrix} 1 \otimes \tilde{X}_{t}^{-1} \tilde{X}_{t} \end{pmatrix} \right)
\]

\[(46)\]

These formulas provide the desired estimators for \( \hat{a} \) and \( \hat{b} \), provided that the variances \( \tilde{G}_{t} \) and \( \tilde{\Sigma}_{t} \) are known. These may be called "market-conditional" regressions, for want of a better term, because the market
return is multiplied times the descriptors to obtain regressions that predict the expected security return conditional upon market return.

III.3 Equivalence to an Alternative Two-Step Procedure, the "Time-Series Test"

In any period \( t \), let the contributions of the descriptors to security return, ignoring the distinction between the contributions through \( \alpha \) and \( \beta \), be written as in (33):

\[
x_t = x_t f_t + \rho_t.
\]

As the first step, for each \( t \), let \( \hat{f}_t \) be estimated by Aitken's Generalized Least Squares (GLS) applied to the cross section of data in period \( t \):

\[
\begin{align*}
\hat{f}_t &= \left( \begin{array}{c} \hat{f}_{1t} \\ \vdots \\ \hat{f}_{Jt} \end{array} \right) = \left( X_t' \Sigma_t^{-1} X_t \right)^{-1} X_t' \Sigma_t^{-1} \tilde{e}_t, \\
\end{align*}
\]

where:

\[
\text{VAR}(\hat{f}_t - f_t) = M_{f_t} = (X_t' \Sigma_t^{-1} X_t)^{-1}.
\]

As the second step, let the time series of estimates of the \( \hat{f}_t \) vector be analyzed to estimate its dependence upon the market return, using the regression model:

\[
\begin{align*}
\hat{f}_{1t} &= a_1 + b_1 r_{Mt} + \epsilon_{f1t} \\
& \vdots \\
\hat{f}_{Jt} &= a_J + b_J r_{Mt} + \epsilon_{fJt}, \\
\end{align*}
\]

\((47a)\) for \( t = 1, \ldots, T.\)
This is a case that Zellner (1962) has called "Seemingly Unrelated Equations." The variables $\hat{f}_1$ through $\hat{f}_J$ do not appear in the same equations, but they are nevertheless jointly dependent because the residual terms $e_{1t}, \ldots, e_{Jt}$ are cross-correlated. In fact,

$$e_{ft} = \hat{f}_{ft} - (a + br_{mt}) = \hat{f}_{ft} + f_{ft} - (a + br_{mt})$$

$$= (\hat{f}_{ft} - f_{ft}) + g_{ft}.$$

The first term is the estimation error in the cross-sectional regression, with mean equal to zero and variance $(X_t'\Sigma_t^{-1}X_t)^{-1}$. The second term has mean zero and variance $g_t$. These two terms are uncorrelated, because the first is a linear function of $\beta_t$, and $\hat{\alpha}_t$ and $g_t$ are uncorrelated from (22). Therefore,

$$E(e_{ft}) = 0; \text{VAR}(e_{ft}) = g_t + (X_t'\Sigma_t^{-1}X_t)^{-1} \equiv \Lambda_t.$$

Thus, $\Lambda_t$ is the variance matrix of the disturbances in the system of equations in period $t$. Unless $\Lambda_t$ is diagonal, which requires that the residual factors be uncorrelated and the estimation errors in the cross-sectional factor regression be uncorrelated as well, the set of equations are jointly dependent.

The formulas for efficient Generalized Least Squares estimation in this "Seemingly Unrelated Equations" model are well known. (See, e.g., Johnston (1972), pp. 238-240.) They simplify greatly in this special case because the explanatory variables are identical in the $J$ equations. To show the relationship between this approach and the preceding, the estimation equations are best expressed as:
\[
\begin{align*}
\left( \begin{array}{c}
\hat{\alpha} \\
\hat{\beta}
\end{array} \right) &= \left( \begin{array}{cc}
1 & r_{Mt} \\
2 & r_{Mt}
\end{array} \right) \otimes \Lambda_t^{-1} \left( \begin{array}{c}
1 \\
r_{Mt}
\end{array} \right) \otimes \Lambda_t^{-1} \frac{\hat{r}}{t}.
\end{align*}
\]

This equation may be simplified, thanks to the matrix identity:

\[
\Lambda_t^{-1} = \left( C_t + (x_t^E_{-t} X_{-t})^{-1} \right)^{-1} = x_t^\prime (x_t C_t x_t^E_{-t} + X_{-t})^{-1} x_t = x_t^\prime v_{-t}^{-1} x_t,
\]

which may be verified directly by matrix manipulations. Using this identity, some further manipulations yield:

\[
\Lambda_t^{-1} \frac{\hat{r}}{t} = x_t^\prime v_{-t}^{-1} x_t \left( x_t C_t x_t^E_{-t} + X_{-t} \right)^{-1} x_t = x_t^\prime v_{-t}^{-1} r_t.
\]

From these two identities, we see that the estimating equation (48) is identical to the minimum-variance linear unbiased estimator in (45). This confirms that the two-step procedure, when correctly applied, is identical to the "market-conditional" regression defined in the preceding section. However, notice that in order for the two-step procedure to provide a minimum-variance linear unbiased estimator, with the correct sampling theory, the following requirements must be met:

1. The cross-sectional estimators of the factors, \( \hat{f}_t \), must be the GLS estimators (47) that are responsive to heteroscedasticity in \( \rho_t \).

2. The \( J \) time-series regressions for the different descriptors must be estimated jointly, as in (48). There are two cases, unlikely to be encountered in reality, in which (48) decomposes into separate estimating equations: either \( \Lambda_t \) may be diagonal for all \( t \), or \( \Lambda_t = \Lambda \) may be any constant matrix (see the derivation of (61), below). In the latter case, the equations may be
estimated separately, but the estimation errors in the different regressions are correlated.

(3) The estimated factors $\hat{\beta}_{jt}$ must be weighted in proportion to $A^{-1}_t$, and therefore must be given different weights in different time periods, unless $A^{-1}_t$ is identical for all $t$.

As will be seen in Part V, several authors have employed variations of this two-step procedure in previous tests of CAP hypotheses, but have not satisfied these requirements.

III.4 Equivalence to a Test on the Moments of the Factors

The two approaches considered thus far employed regressions on the market return, thus estimating the conditional distribution directly. As an alternative, we may study the joint distribution of security returns itself to estimate the mean returns to the factors and the mean specific returns, and the variances of these components. From this estimated distribution, we can then compute the implied distribution for any hypothesized market return and the covariances between the risk components and that return. The difference between this approach and the two previous ones is that here the joint distribution of returns is examined first, with no need for a market portfolio to be specified. Then, at the final stage, the consistency of the joint distribution with the CAP hypotheses applied to a hypothetical market portfolio is tested.

In the present paper, we undertake a limited analysis of this approach that is sufficient to show the analogy with the previous tests. To facilitate this demonstration, we assume that the distribution of returns in the multiple-factor model developed in Part II is completely stationary: that is, $x_{t}, w_{mt}, \bar{f}_{t}, \overline{F}_{t}, \overline{D}_{t}$, and $\Sigma_{t}$ are constant over time. The time subscripts will therefore be deleted.
Suppose that one wanted to "speculate on the factor returns"—that is, to construct an investment portfolio that would obtain the return applying to some factor \( j \). One might define a vector of investment proportions, \( w_j \), such that \( \sum_n w_{jn} x_{jn} = 1 \). One might further desire that the return to the speculation on the \( j \)th factor be unaffected by the returns on the other factors, which would result if \( \sum_n w_{jn} x_{in} = 0 \) for \( i \neq j \).

One might finally desire that the return to the speculation be minimally influenced by the specific returns, which would result if \( \text{VAR}(\sum_n w_{jn} \epsilon_{jn}) = w_{j}^T \Sigma w_j \) were minimal. When these conditions are combined, the requirements for the \( J \) "investment portfolios" may be expressed by the matrix equations:

\[
\begin{align*}
\text{minimize} & \quad \sum_{j \neq \hat{j}} w_{j}^T \Sigma w_j, \quad j=1,\ldots,J \\
\text{subject to} & \quad X_\sim' \left( w_1 \ldots w_j \right) = I_\sim.
\end{align*}
\]

(51)

But this problem is identical to minimum variance linear unbiased estimation of \( f \) in the regression equation (7). Therefore, the optimal investment portfolios are just the weights of the Aitken's GLS estimators for the factors:

\[
W_p^T \equiv (w_1 \ldots w_j)' = \left( X_\sim^T \Sigma^{-1}_\sim X_\sim \right)^{-1} X_\sim^T \Sigma^{-1}_\sim.
\]

(52)

The total investment value in any "portfolio" is \( \sum_n w_{jn} = \sum_n w_{jn} x_{jn} \), since \( x_1 = 1 \). Since \( W_p^T X = I \), it follows that the sum of investments in the first portfolio is one, and in all other portfolios, the sum of investments is zero.
The returns achieved on these \( J \) portfolios will be given by the vector:

\[ \bar{r}_P = \bar{w}_P \bar{r} = \bar{w}^T (X \bar{f} + \bar{\rho}) = \bar{f} + \bar{w}^T \bar{\rho}. \]

The mean return will therefore be:

\[ \bar{r}_P = \bar{f} + \bar{w}^T \bar{\rho}, \tag{53} \]

and the variance of return will be:

\[ \bar{v}_P = \bar{f} + \bar{w}^T \Sigma \bar{P} = \bar{f} + (X' \bar{\Sigma}^{-1} \bar{X})^{-1} \bar{X}' \bar{\Sigma}^{-1} \Sigma \bar{P} \]

where \( \Sigma \bar{P} \) will be recognized as the variance of estimation error for the GLS estimator of \( \bar{f} \) given in (47). Finally, the covariance of returns on these portfolios with the market portfolio will be:

\[ c_{P} = \text{COV}(r_P, r_M) = \bar{w}_P \Sigma \bar{X} \bar{X}' \Sigma \bar{w}_M + \bar{w}_P \Sigma \bar{w}_M = \frac{\bar{f} \bar{x}_M}{\Sigma \bar{X}} + (X' \Sigma^{-1} \bar{X})^{-1} X' \Sigma^{-1} \Sigma \bar{w}_M \]

\[ = \frac{\bar{f} \bar{x}_M}{\Sigma \bar{X}} + \Sigma^{-1} \Sigma \bar{w}_M = \frac{\bar{f} \bar{X}_M}{\Sigma \bar{X}} = \bar{v} \Sigma \bar{w}_M, \tag{55} \]

where \( \bar{x}_M \) is the vector of market weights on the factors.

Now, the first portfolio constitutes an investment opportunity, just as any other security, so that the CAPM condition becomes:

\[ \bar{r}_{P1} = k + c_{P1}(\bar{r}_M - k)/\bar{v}_M. \]

For the other portfolios, each with net investment value of zero, the slightly modified CAPM condition is:
\[ \bar{r}_{Pj} = c_{Pj}(e_{-k})/\bar{V}_M, \quad j=2,\ldots,J. \]

When these \( J \) conditions are examined, they are seen to be equivalent to the CAP:Z conditions stated in Section 1.5, provided that:

\[ a = \bar{r}_P - (\bar{r}_M/V_M)c_P; \quad b = c_P/V_M. \]

Expressions (56) will be recognized as the formulas for the intercepts and slope coefficients in regressions of the portfolio returns onto the market return, so that this approach is conceptually equivalent to the previous two. It remains to verify that optimal estimators of the joint distribution of the factors, when substituted into equations (53) through (56), will result in an empirical test that is equivalent to the two preceding tests. This problem will now be considered.

Because of the assumed constancy in the joint distribution, the minimum variance linear unbiased estimator of \( \bar{r}_P \) may be shown to be the average of the GLS estimators of the factors:

\[ \hat{\bar{r}}_P = \frac{1}{T} \sum_t \hat{r}_{P,t} = \frac{1}{T} \sum_t \hat{M}_t \hat{X}_t \hat{\Sigma}^{-1} \hat{X}_t = \frac{1}{T} \sum_t \hat{f}_t. \]

To obtain an estimator of the variance of portfolio returns, we observe that:

\[ E\left( \sum_t \hat{f}_t \hat{f}_t' \right) = E\left( \sum_t \hat{M}_t \hat{X}_t \hat{\Sigma}^{-1} (\hat{X}_t \hat{X}_t') \hat{\Sigma}^{-1} \hat{X}_t \right) \]

\[ = E\left( \sum_t \hat{M}_t (\hat{X}_t \hat{\Sigma}^{-1} \hat{X}_t + c_t) (\hat{X}_t' \hat{X}_t' + d_t) \hat{\Sigma}^{-1} \hat{X}_t \right) \]

\[ = E\left( \sum_t \hat{X}_t \hat{\Sigma}^{-1} (\hat{X}_t + c_t) (\hat{X}_t' + d_t) \hat{\Sigma}^{-1} \right) \]
\[
\sum_t \left( \overline{R}_{P,P} + \overline{R} + W_{P,P} \overline{R}_{P,P} \right) = T(\overline{R}_{P,P} + \overline{R}_{P,P} + \overline{V}_P).
\]

Similarly,

\[
E(\overline{R}_{P,P}) = \frac{1}{T^2} E \left( \sum_t \sum_s \overline{R}_{t,s} \overline{R}_{s,t} \right)
\]

\[
= \frac{1}{T^2} \left( T^2 \overline{R}_{P,P} + T \overline{V}_P \right) = \overline{R}_{P,P} + \frac{1}{T} \overline{V}_P.
\]

Therefore, an unbiased estimator of the variance of returns is given by:

\[
\hat{V}_P = \frac{1}{T-1} \left( \sum_t \overline{R}_{t,t} - \overline{R}_{P,P} \right).
\]

This estimator also has a strong optimality property, in that it is the Minimum Norm Quadratic Unbiased Estimator (MINQUE) (see Rao (1972) and Swamy (1971)). From (55), an unbiased estimator of the covariances with the market return is then given by:

\[
\hat{c}_p = \hat{V}_P x_p.
\]

Finally, the familiar estimators for the mean and variance of market returns are:

\[
\hat{\mu} = \frac{1}{T} \sum_t R_{Mt} ; \quad \hat{\sigma}^2 = \frac{1}{T-1} \left( \sum_t R_{Mt}^2 - \hat{\mu}^2 \right).
\]

We now demonstrate that the estimators \( \hat{\alpha} \) and \( \hat{\beta} \) from (56), obtained by substitution of (57) through (60), are identical to the estimators \( \hat{\alpha} \) and \( \hat{\beta} \) obtained in the previous tests. The first step is to
derive the simplifications in the earlier estimators (48) that result when the distribution is stationary:

\[
\begin{align*}
\left( \hat{\beta} \right) &= \left( \sum_t \left( \begin{array}{c}
1 & \tau_{Mt} \\
\tau_{Mt} & \tau_{Mt}^2
\end{array} \right) \right)^{-1} \left( \sum_t \left( \begin{array}{c}
1 \\
\tau_{Mt}
\end{array} \right) \right) \times \left( C+M_f \right)^{-1} F_t \\
\left( \hat{\beta} \right) &= \left\{ \left( \sum_t \left( \begin{array}{c}
1 & \tau_{Mt} \\
\tau_{Mt} & \tau_{Mt}^2
\end{array} \right) \right)^{-1} \right\} \times \left( C+M_f \right)^{-1} \sum_t \hat{F}_t \\
\left( \hat{\beta} \right) &= \left\{ \left( \sum_t \left( \begin{array}{c}
1 & \tau_{Mt} \\
\tau_{Mt} & \tau_{Mt}^2
\end{array} \right) \right)^{-1} \right\} \times \left( \sum_t \hat{F}_t \tau_{Mt} \right)
\end{align*}
\]

(61) \[= \frac{1}{T \sum_t \tau_{Mt}^2 - \left( \sum_t \tau_{Mt} \right)^2} \left( \sum_t \tau_{Mt}^2 \sum_t \hat{F}_t - \sum_t \tau_{Mt} \sum_t \hat{F}_t \tau_{Mt} \right) \]

The crucial simplification occurs in the second step above, where constancy of \( A_t = C + M_f \) allows the formula for the inverse of the Kronecker product of two matrices to be applied. The formulas (57) and (60) may be substituted into this expression to obtain:
\[
\left( \widehat{\alpha} \right)^R = \frac{1}{T(T-1)\hat{V}_M} \left( (T(T-1)\hat{V}_M + T_2 \hat{\triangle}_M \hat{\triangle}_P - T_\hat{\triangle}_M \hat{\triangle}_P \hat{r}_M) \right. \\
\left. \hat{r}_M \right) \\
\left( \hat{r}_M \right)^R \left( \hat{r}_M \right)
\]

\[
(62)
\]

To prove equality between this expression and (56), the term in square brackets must be shown to equal \( \hat{\triangle}_P \). Evaluating the latter from (58) and (59),

\[
\hat{\triangle}_P = \hat{V}_P x_M = \frac{1}{T-1} \left( \sum_{t=1}^{T} \hat{\triangle}_P \hat{r}_M \right) x_M
\]

\[
= \frac{1}{T-1} \sum_{t=1}^{T} \hat{\triangle}_P \hat{r}_M x_M
\]

where the last expression is obtained by adding and subtracting terms in \( r_M = \hat{r}_M x_M \). The first term is equal to the term in square brackets in (62).

Thus, the proof will be complete if the second term is equal to zero. This will hold exactly iff

\[
(I - \Sigma^{-1} \hat{X} \Sigma^{-1} \hat{X}^{-1} \hat{X}' \hat{X}) w_M = 0,
\]
iff

$$W_M = E^{-1}X(X'X^{-1}X)^{-1}X'W_M.$$ 

Thus, the equality will hold if the vector of proportional investments is preserved by the idempotent matrix $$E^{-1}X(X'X^{-1}X)^{-1}X',$$ which, in turn, will hold iff the market portfolio is some linear combination of the factor portfolios—that is, for some vector $$\lambda,$$ $$W_M = W_P \lambda.$$ In particular, if the market portfolio is included as one of the factors, this equality holds and the methods are identical.

When the equality fails to hold, the two earlier methods differ by inclusion in $$\hat{c}_p$$ of the term:

$$\frac{1}{T-1} \left( \sum_{t} (\hat{f}_t - \bar{r}_t) r_t' \right) (I - E^{-1}X(X'X^{-1}X)^{-1}X') W_M.$$ 

The postmultiplication by the idempotent matrix removes the contribution of the factors to $$\hat{r}_t,'$$ as well as the component of $$\hat{p}_t,'$$ in the space spanned by $$X',$$ and these are the only components of $$\hat{r}_t,'$$ that enter into $$\hat{c}_t.$$ As a result, $$r_t'(I - E^{-1}X(X'X^{-1}X)^{-1}X')$$ is uncorrelated with

$$\hat{r}_t = (X'X^{-1}X)^{-1}X'X^{-1}X't,$$ since

$$E[(X'X^{-1}X)^{-1}X'X^{-1}X't r_t'(I - E^{-1}X(X'X^{-1}X)^{-1}X')]$$

$$= (X'X^{-1}X)^{-1}X'X^{-1}(XFX'+E)(I - E^{-1}X(X'X^{-1}X)^{-1}X')$$

$$= FX' - FX' + (X'X^{-1}X)^{-1}X' - (X'X)^{-1}X' = 0.$$ 

Therefore, the earlier methods differ by the addition of a term with zero mean and variance of order $$1/(T-1)$$ to $$\hat{c}_p,$$ so that they are asymptotically
equivalent. However, it may be shown that the method developed in this section is slightly more efficient, in that the variance of estimation error for \( \hat{\sigma}_p \) is somewhat smaller. The increase in efficiency is achieved by exploiting the information in (55) on the structure of the covariance between \( \hat{f}_c \) and \( r_{Mt} \).

### III.5 Estimation of the Prediction Rules for \( \alpha \) and \( \beta \) in the Generalized Context

We now return to the general model stated in equations (28) through (33). For simplicity, the parameters of the residual return variance matrices, \( \Sigma_t \), are presumed to be known up to a scale parameter and equal to the estimators obtained in step (4) of the procedure outlined in section III.1. (There is no difficulty, other than computational, in jointly estimating these parameters with those of the \( \alpha \) and \( \beta \) processes by the methods developed below.) Define a universal scale parameter \( \sigma^2 \) for all variance components, including the residual returns. This parameter appears multiplicatively in every variance component.

There is an important distinction to be drawn among the parameters of the model: The variance components \( (\sigma_a^2, \sigma_b^2, \sigma_k^2, \xi^2) \)---up to a scale parameter---and the convergence rates \( (\phi_a^2, \phi_b^2, \phi_k^2) \) are parameters of the stochastic specification that enter nonlinearly in the model; let \( \theta \) denote this set of parameters. For any stochastic specification \( \theta \), the evaluation of the likelihood function is equivalent to the computation of a generalized least squares (GLS) regression, in which the regression parameters are \( a \) and \( b \), the scale parameter for disturbance variance is \( \sigma^2 \), and the likelihood is determined by the goodness of fit of the regression.
Under assumption (A15) that all stochastic terms are normally distributed, the maximum likelihood approach to estimation is set out in Rosenberg (1973b, sections IV and V). Maximum likelihood proceeds by a search of the space \( \mathcal{X}(\theta) \) of permissible values of the stochastic specification parameters; for each set of values of these parameters, \( \theta \), the likelihood function is evaluated by the GLS formulas that yield the exact maximum likelihood estimators for the remaining parameters \( a, b, \) and \( \sigma^2 \). When the maximum is found, the stochastic specification at the maximal point is the maximum likelihood estimator, \( \hat{\theta}_{ML} \), and the GLS regression at that specification yields the MLEs for \( a, b, \) and \( \sigma^2 \).

To implement the approach, the following are required: initial estimators for the stochastic specification, \( \theta^0 \), to provide a starting point for the search; the formulas to carry out the GLS regression for each specification and evaluate the likelihood; and a search procedure, which is facilitated by computation of the partial derivatives of the likelihood function with respect to the specification parameters.

Initial estimators for the stochastic specification may be obtained as follows. In step (2) of the procedure to estimate the prediction rule for \( \sigma^2_{nt} \), estimates \( \hat{f}_t \) of the factor vectors in all periods are obtained. From (47), the estimation error variances of these estimators are not constant. Nevertheless, one can apply the same approach as in equations (57) and (58) for this more general case to obtain the initial estimator:

\[
\text{VAR}\left(\hat{f}_t - \bar{f}\right) = \frac{1}{T-1} \sum_{t=1}^{T} \left( \hat{f}_t - \bar{f} \right) \left( \hat{f}_t - \bar{f} \right)'
\]
Initial estimators of the other parameters specifying the stochastic process, \( \phi_k, \phi_a, \phi_b, \Sigma_a, \Sigma_b \), and the initial allocation of variance between \( \Sigma \) and \( \Sigma_b \) are provided by heuristic arguments of the sort in footnotes 2 and 3.

For the search of the specification space, we rely on an optimization procedure of Fletcher and Powell (1965). Designation of step sizes and of the criterion for convergence is facilitated by the known sampling properties of a likelihood function. Variance components, which must be positive semi-definite (and symmetric if a matrix), are expressed as the product of a triangular matrix with itself, a device that reduces the number of parameters to a minimum and assures positive semi-definiteness and symmetry. Our experience with this approach in stochastic parameter regression problems has been universally satisfactory, and a local maximum— as opposed to a global maximum—has never been encountered.

It remains to present the formulas for the evaluation of the likelihood functions (the implicit GLS regressions) and for the partial derivatives of the likelihood functions, where appropriate. As an intermediate step, the model without sequential variation is estimated first. The specification parameters, \( \theta \), are then \( \Sigma \) and \( \Sigma_b \). From equations (27) through (32), the model becomes:

\[
\begin{align*}
\tilde{z}_t &= X_a + (r_m X) b + r_m Z b + \rho_t \\
&= X_a + (r_m X) b + \nu_t \\
&= \tilde{z}_m + \nu_t,
\end{align*}
\]
where:

\[ O_t \equiv (X_t : r_{Mt-t} X_t) ; \quad m \equiv (a' : b')' ; \]

\[ E[Y_t] = 0 ; \quad E(y_t y_t') = Z_t (\sigma^2 G_t Z_t') + \sigma^2 \Sigma_t \equiv \sigma^2 \Sigma_t \]

\[ \text{COV}(y_t, X_t) = \text{COV}(y_t, r_{Mt-t} X_t) = 0 \]

\[ \Sigma_t = \text{DIAG}(\delta^2_{nt}) ; \quad G_t = G + r_{Mt-b}^2 \Omega. \]

This is a "randomly dispersed parameter" regression problem with nonconstant parameter dispersion matrix \( G_t \) of a special form and with a presumably known pattern of residual heteroscedasticity. Following equations (7), (10), (5*), (8*), and (9*) in Rosenberg (1973a), the logarithm of the likelihood is easily derived to be:

\[ L(\theta) = \text{constant} - \frac{1}{2} (\delta + N \ln(N-k)s^2) \]

where \( \delta \) and \( s^2 \) are given as functions of \( \theta \) by the equations,

\[ \delta = \sum_{t} \delta_{t} = \sum_{t} \left( \ln|G_t| + \ln|B_t| \right) \]

\[ s^2 = \frac{1}{N-k} (u-h'H^{-1}h) \]

with \( B_t = G_t^{-1} + Z_t \Sigma_t^{-1} Z_t' \)

\[ k = 2J \]

\[ N = \sum_{t} N(t) \]
- \[ u = \sum_t u_t = \sum_t \left( r_t^{1} \Sigma_t^{-1} \Sigma_t - r_t^{1} \Sigma_t^{-1} Z_t B_t^{-1} Z_t^{1} \Sigma_t^{-1} \right) \]

- \[ h = \sum_t h_t = \sum_t \left( q_t^{1} \Sigma_t^{-1} \Sigma_t - q_t^{1} \Sigma_t^{-1} Z_t B_t^{-1} Z_t^{1} \Sigma_t^{-1} \right) \]

- \[ H = \sum_t H_t = \sum_t \left( q_t^{1} \Sigma_t^{-1} O_t - q_t^{1} \Sigma_t^{-1} Z_t B_t^{-1} Z_t^{1} \Sigma_t^{-1} O_t \right) \]

Maximization of this function is equivalent to minimization of:

\[ \phi(\Sigma) = \delta + N \ln((N-k)s^2). \]

The partial derivatives of this function, with respect to the specification parameters, are given by the matrix equations:

\[ \frac{\partial \phi}{\partial \Sigma} = \frac{\partial \delta}{\partial \Sigma} + \frac{N}{(N-k)s^2} \frac{\partial ((N-k)s^2)}{\partial \Sigma} \]

where

\[ \frac{\partial \delta}{\partial \Sigma} = \sum_{t=1}^{T} \frac{\partial \delta}{\partial \Sigma_t} = \sum_{t=1}^{T} \left( G_t^{-1} - G_t^{-1} B_t G_t^{-1} \right) \]

\[ \frac{\partial ((N-k)s^2)}{\partial \Sigma} = \sum_{t=1}^{T} \frac{\partial ((N-k)s^2)}{\partial \Sigma_t} \]

\[ = - \sum_{t=1}^{T} G_t^{-1} B_t^{-1} \Sigma_t^{-1} (r_t^{1} - O_t)^{-1} (r_t^{1} - O_t)^{-1} \Sigma_t^{-1} G_t^{-1} \]
and

\[ \frac{\partial \ell}{\partial \Omega_b} = \frac{\partial \delta}{\partial \Omega_b} + \frac{N}{(N-k)s^2} \frac{\partial ((N-k)s^2)}{\partial \Omega_b} \]

where:

\[ \frac{\partial \delta}{\partial \Omega_b} = \sum_{t=1}^{T} \frac{1}{M_t} \left( \frac{\partial \delta}{\partial \gamma_t} \right) \]

\[ \frac{\partial ((N-k)s^2)}{\partial \Omega_b} = \sum_{t=1}^{T} \frac{1}{M_t} \left( \frac{\partial ((N-k)s^2)}{\partial \gamma_t} \right) \]

These partial derivatives may be employed in the minimization procedure of Fletcher and Powell. At the maximum likelihood estimates of \( \hat{\gamma} \) and \( \hat{\Omega}_b \), the explicit GLS equations yielding the maximum likelihood estimators of \( \hat{a}, \hat{b} \), and the scale parameter \( s^2 \) are:

\[ \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \hat{H}^{-1} \hat{h} \]

\[ s^2 = \frac{N-k}{N}s^2 \]

\[ \begin{pmatrix} \hat{M}_{aa} & \hat{M}_{ab} \\ \hat{M}_{ba} & \hat{M}_{bb} \end{pmatrix} = \sigma_2 H^{-1} \]

where \( h, \hat{h} \), and \( s^2 \) are computed at \( \hat{\gamma} \) and \( \hat{\Omega}_b \). Equations (69) confirm that the GLS regression is carried out implicitly in evaluating the likelihood function by (63).
Asymptotic confidence regions for the estimated stochastic specification are provided by the Hessian of \( \mathcal{L}(\theta) \), with \( \theta - \hat{\Theta}_{ML} \) being asymptotically normally distributed with mean zero and variance matrix:

\[
\Sigma_{\theta} = \left( \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta \partial \theta} \right)^{-1}.
\]

In the preceding derivation, the GLS regression and the evaluation of the likelihood function involve the four terms, \( \Sigma, \nabla, \delta, \) and \( s^2 \), each of which is a summation over the \( T \) periods. The decomposition into a summation results because the stochastic terms are serially uncorrelated, so that the variance matrix for the GLS regression is block diagonal, with one block per period. In the more general case including sequential parameter variation, the stochastic terms are serially correlated. A computationally feasible solution to the GLS regression problem is obtained through a transformation of the model that eliminates this serial correlation. The development of the appropriate transformation is provided in greater detail in Rosenberg (1973c, Section II.3) and Rosenberg and Ohlson (1975).

To facilitate the derivation of the solution for the present case, note that in the multiple factor model (33), with known residual return heteroscedasticity \( \Sigma_t \), the information concerning the factors \( \hat{f}_t \) in period \( t \) is captured by the sufficient statistics \( \hat{f}_t \) and \( \Sigma_{\hat{f}_t} \):

\[
\hat{f}_t = \mathcal{N} (\hat{f}_t, \sigma^2_{\hat{f}_t}).
\]
The model with sequential parameter variation may therefore be written in the notation of a regression model as:

\[
\begin{pmatrix}
\hat{r}_t \\
\end{pmatrix} = \begin{pmatrix}
a + br_{-1} \t M_t \\
\end{pmatrix} + \begin{pmatrix}
0 \\
0 \\
\end{pmatrix} \cdot \begin{pmatrix}
a_t \\
k_t + b_t \t \mathbf{r}_{-1} \\
\end{pmatrix} + \begin{pmatrix}
\text{etc.} \\
\text{etc.} \\
\end{pmatrix} \cdot \begin{pmatrix}
\text{estimates} \\
\end{pmatrix}
\]

\[
(72)
= X_{1t} m + X_{2t} \eta_t + u_t,
\]

where:

\[
X_{1t} = (I : I_r \t \mathbf{M}_t) ; \quad X_{2t} = \begin{pmatrix}
1-b_1 : 0 : 0 \\
-b_2 \\
-b_3 : I : r_t \mathbf{M}_t \\
-b_4 \\
: 0 : 0 \\
-b_J \\
\end{pmatrix}
\]

\[
m = \begin{pmatrix}
a \\
a_t \\
\end{pmatrix} \quad \eta_t = \begin{pmatrix}
k_t \\
a_t \\
b_t \\
\end{pmatrix}
\]

\[
u_t = g_t + \hat{r}_t - r_t,
\]

with \(\mathbb{E}(u_t) = 0;\) \(\text{VAR}(u_t) = \sigma^2 \left( C + \mathbf{r}_t \mathbf{M}_t \mathbf{r}_t + \mathbf{M}_t \mathbf{r}_t \right) = \sigma^2 \mathbf{R}_t,\) where \(\eta_t\) is
specified by the stationary first-order Markov process

\[(73) \quad \eta_t = \phi \eta_{t-1} + d \eta_{t-1},\]

with \( \text{VAR}(d \eta_t) = \sigma^2 \begin{pmatrix} q_k & 0 & 0 \\ 0 & q_a & 0 \\ 0 & 0 & q_b \end{pmatrix} = \sigma^2 q; \quad \phi = \begin{pmatrix} \phi_k & 0 & 0 \\ 0 & \phi_a & 0 \\ 0 & 0 & \phi_b \end{pmatrix} \)

and where the \( y, \eta, \) and \( X \) processes are independent of one another.

In this model, the serially independent random terms \( g \) and \( e_b \) are treated as nuisance parameters and merged with the estimation error \( \hat{f}_t - \tilde{f}_t \), but the sequentially varying parameters \( \eta \) are retained explicitly. This is done because the aforementioned transformation to obtain serial independence is most easily understood as a recursive estimation procedure for \( \eta_t, t=1, T \). In particular, the transformation will involve the mean and variance of the distribution of \( \eta_t \), conditional on \( m, \sigma_\eta^2, X \), and the observations of \( \tilde{f}_t \) prior to date \( s \), where \( s \) will take the values of \( t-1 \) or \( t \). Let \( f_{s}^s \) denote the set of all observations of \( \tilde{f}_t \) through to date \( s \). Then define:

\[(74) \quad \mu^*_t|_s = \text{E}[\eta_t| m, X, \tilde{f}_s] \]

\[\sigma^2_{\eta}^*_t|_s = \text{VAR}[\eta_t| m, \sigma^2, X, \tilde{f}_s].\]

The initial values of these moments, prior to observation of any estimated factors, are taken from the stationary distribution for \( \eta \) implied by \( \phi \) and \( \sigma^2 q \).
\[ u^*_1|0 = 0 \]

\[
\begin{pmatrix}
\frac{q_k}{1-\phi_k^2} & 0 & 0 \\
0 & \frac{q_a}{1-\phi_a^2} & 0 \\
0 & 0 & \frac{q_b}{1-\phi_b^2}
\end{pmatrix}
\]

(75)

\[ \sigma^2 m^*_{-1}|0 = \sigma^2 \]

Next, a recursive equation for these moments may be derived by induction.

Suppose that \( u^*_{t-1} \) is of the form:

\[ u^*_{t-1} = \xi_t|t-1 + \eta_t|t-1 \]

That is, it is linear in \( \eta_t \). Notice that \( u^*_{1|0} \) satisfies this assumption trivially. Then, to derive \( u^*_{t|t} \) and \( m^*_{t|t} \), we note that \( u^*_{t-1} \) and \( m^*_{t-1} \) specify a prior distribution for \( \eta_t \), and that \( u^*_{t|t} \) and \( m^*_{t|t} \) specify the posterior distribution for \( \eta_t \), after incorporation of the information provided by the linear relation

\[ \hat{\eta}_t = X_t \eta_m + X_t \eta_{t-t} + u_t \]

But conditional on \( m \), this has the form of a linear regression for \( \eta_t \), with disturbances that are independent of all prior stochastic terms and hence of the prior distribution. Therefore, from the familiar formulas for the incorporation of the information in a linear regression into a prior distribution:
\[ u^*_t | t = u^*_t | t-1 + M^*_t | t-1 \cdot X_{2t}^* \cdot e_t^* \]

(76)

\[ M^*_t | t = M^*_t | t-1 - M^*_t | t-1 \cdot X_{2t}^* \cdot e_t^* \]

(77) where \[ F^*_t \equiv X_{2t}^* \cdot M^*_t | t-1 \cdot X_{1t}^* + R_t = \frac{1}{\sigma^2} \text{VAR}(\hat{f}_t | m, \sigma, X, \hat{\gamma}_t^{-1}) \]

(78) and \[ e^*_t \equiv \hat{f}_t - X_{1t} \cdot \hat{m} - X_{2t} \cdot \hat{\gamma}_t^{-1} = \hat{\gamma}_t - E[\hat{f}_t | m, X, \hat{\gamma}_t^{-1}] \].

Note that the linear form for \( u^*_t \) is preserved, since,

\[ u^*_t | t = \hat{f}_t | t + \hat{e}_t | t \]

(79) where \[ \hat{f}_t | t = \hat{f}_t | t-1 + M^*_t | t-1 \cdot X_{1t}^* \cdot e_t^* | t-1 \]

\[ \hat{e}_t | t = \hat{e}_t | t-1 - M^*_t | t-1 \cdot X_{2t}^* \cdot e_t^* | t-1 \].

Thus, the step from the prior distribution for \( \eta_t \) to the posterior distribution for \( \eta_t \) has been completed.

Next, the posterior distribution for \( \eta_t \) is extrapolated to a prior distribution for \( \eta_{t+1} \). Since \( \eta_{t+1} \) is the affine function of \( \eta_t \) given in (73), with a random increment \( d \eta_t \) that is independent of previous stochastic terms, the extrapolated moments are immediately found to be:

\[ u^{*}_{t+1} | t = \hat{f}_{t+1} | t + \hat{e}_{t+1} | t \]

(80)

where \[ \hat{f}_{t+1} | t = \Phi \cdot \hat{f}_t | t \]

\[ \hat{e}_{t+1} | t = \hat{e}_t | t \]

\[ M^{*}_{t+1} | t = \Phi M^*_t | t \cdot \Phi' + Q. \]
This completes the recursion from the prior distribution for $\eta_t^*$ to the prior distribution for $\eta_{t+1}^*$. Since linearity in $\zeta$ is preserved, the proof by induction is complete.

Next, note that the random variables $e_t^*, t=1,\ldots,T$ are each a linear transformation of the form $e_t^* = \hat{f}_t^* + \hat{A}_t^* \hat{z}_{t-1}^*$. Therefore, the transformation from the set of variables $\hat{z}_1^*, \ldots, \hat{z}_T^*$ to the set of variables $e_1^*, \ldots, e_T^*$ is a nonsingular linear transformation with a determinant (or Jacobean) of 1. Consequently, the likelihood of the $\{\hat{z}_t\}$ is the same as the likelihood of the $\{e_t^*\}$. Conditional on $\zeta$, the $e_t^*$ are serially independent by construction, for each is the difference between $\hat{f}_t$ and $E(\hat{f}_t | \hat{f}_1, \ldots, \hat{f}_{t-1}, \zeta, \eta)$. Hence, the logarithmic likelihood function, conditional on $\zeta$, $\sigma$, is:

$$
   l(\zeta, \sigma) = \text{constant} - \frac{1}{2} \left( \sum_{t=1}^{T} \ln |\sigma_t^2| + \sum_{t=1}^{T} e_t^* (\sigma_t^2)^{-1} e_t^* \right).
$$

Since $e_t^* = \hat{f}_t - X_{2t-1} w_t + X_{2t} z_t | \zeta_{t-1} - (X_{1t} + X_{2t} \zeta_t) \mu$, the likelihood function may be seen to take the form of a generalized least squares regression on $\zeta$, with the linear relation in each period $t$ being:

$$
   \hat{f}_t^* \equiv \hat{f}_t - \hat{X}_{2t-1} \hat{u}_t^* | \zeta_{t-1} = (X_{1t} + X_{2t} \zeta_t | \zeta_{t-1}) \mu + \hat{u}_t^* \equiv T \mu + \eta_t^*,
$$

such that the $u_t^*$ are serially uncorrelated with zero mean and variance $\sigma^2_{u_t^*}$. This is the desired form for the generalized least squares regression. It is now a simple matter to verify that the likelihood function is given by the previous formula (65), with the revised terms:
\[ \delta = \frac{1}{T} \sum_{t=1}^{T} \delta_t = \frac{1}{T} \sum_{t=1}^{T} \ln |F_t^*| \]

\[ s^2 = \frac{1}{N-k} (\mathbf{v}'\mathbf{h}^{-1}\mathbf{v}) \]

\[ k = 2J \]

\[ N = TJ \]

(83)

\[ \mathbf{v} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{v}_t = \frac{1}{T} \sum_{t=1}^{T} \hat{F}_t'^* \hat{F}_t^* \mathbf{v}_t \]

\[ \mathbf{h} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}_t = \frac{1}{T} \sum_{t=1}^{T} \hat{F}_t'^* \hat{F}_t^* \mathbf{h}_t \]

\[ \mathbf{H} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{H}_t = \frac{1}{T} \sum_{t=1}^{T} \hat{T}_t'^* \hat{T}_t^* \mathbf{H}_t \]

The computation proceeds recursively, one period at a time. Each recursion consists of extrapolation of the prior distribution for \( \eta_t \) in the next period (the initial extrapolation by (75), subsequent extrapolations by (76) and (77)); construction of the transformed random variables (by (77) and (80)); summation of the GLS statistics (by (83)); and computation of the posterior distribution for \( \eta_t \) (by (76) and (79)). One sequence of recursions is analogous to the simple summation for the previous randomly dispersed parameter case, and requires about three times as many arithmetic operations. However, computation of the partial derivatives of the likelihood function with respect to the specification parameters is more cumbersome in this case, so that a search algorithm
that does not employ derivatives is preferable. We have employed Powell's algorithm (1964), with satisfactory results. Once the maximum likelihood solution is obtained, the MLEs for \(a, b,\) and \(\sigma^2\) are again given by (69).

IV. THE EMPIRICAL RESULTS

The empirical study duplicates the format of earlier studies as closely as possible, to facilitate comparison. The CRSP tape provides \(i_{nt}\) for all stocks listed on the New York Stock Exchange for all months from February 1926 to December 1966. The rate of return on the market index \(i_{Mt}\) is taken as the rate of return to the Fisher (1966) "Link-Relative" index, with reinvestment of dividends. (This is a misspecification, in that Capital Asset Pricing Theory suggests that a value-weighted index, rather than an equal-weighted index such as this, should be used. For a discussion of this problem, see Miller and Scholes (1972, pp. 63-66).) The risk-free rate, \(i_{fr}\), is the ninety-day treasury bill rate quoted at the beginning of the month, expressed as a monthly rate of interest.

To begin, the descriptors specified in Section I.4, as transformed in Table I, are computed for the five-year period from 1926 to 1930. The descriptors are construed as available if fifty monthly returns (out of a possible sixty) and dividend yields over the latest calendar year are present. Those firms, \(n,\) for which all descriptors are available, make up the sample for the 1931 calendar year. Whatever monthly returns, \(i_{nt}\), are available in year 1931 for these firms are included in the data sample, with descriptor values as computed from the five previous calendar years. Next, descriptors are computed for the five-year period ending in 1931,
the sample of firms with available descriptors is redetermined, and all monthly returns in 1932 for these firms are added to the sample. The procedure continues to 1966, with the descriptor values for 1966 being computed over the years 1961-1965. The available cross section grows from 419 firms in 1931 to 874 firms in 1966. In all, there are 315,227 monthly returns.

IV.1 The Prediction Rule for Yield

The prediction rule for yield, as defined in (36), was estimated as:

\[
\frac{y_{ns} - y_{Ms}}{y_{Ms}} = -0.0054 + 0.1225 \beta_{ns} - 0.0743 H_{ns}^2
\]

\[
+ 0.7126 \delta_{ns} - 0.0771 \beta_{ns}^2 + 0.00008 H_{ns}
\]

\[
\text{Sample size} = 25,711; R^2 = 0.4521; \text{Residual mean square} = 0.3311.
\]

The data sample includes all firms \( n \) and calendar years \( s \), such that (i) historical descriptors for firm \( n \) are available for year \( s \), and (ii) dividend data are available for the complete calendar year \( s \). Consequently, there is a slight ex-post selection bias, resulting from the exclusion of data for the calendar year in which a stock was delisted. This problem was thought to be negligible for the current study but should be resolved in later work. The constant term would be equal to zero if the average values of all descriptors were zero; the nonzero constant arises because the descriptors were standardized to have zero average
values over the sample of monthly returns, a slightly different sample
than the present one.

Notice that \( H_0 \) plays the dominant role in prediction, but that
the estimated coefficient of \( .7126 \) is far smaller than the unit value im-
plied by naive expectations of type (E1). The other descriptors provide
statistically significant, although relatively unimportant, adjustments,
so that the rational prediction rule cannot be of form (E2).

IV.2 The Specification of Variance

The first step was to estimate the conditional prediction rules
for heteroscedasticity. First, prediction rules for \( \text{VAR}(r_{nt} - r_{Mt}) \), for
both alternative "rate-of-return" and "logarithmic-return" models speci-
fied in Section I.2, were fitted by equation (40). The results are given
in the first two columns of Table 2. Using the resulting truncated pre-
diction rules of form (42), the moments of the transformed returns
\( (r_{ns} - r_{Ms}) / \text{VAR}(r_{ns} - r_{Ms})^{1/2} \) were computed. The results are given in
Table 3.

Table 3 shows that a large part of the skewness and kurtosis of
the rates of return is removed by the transformation to logarithmic re-
turns, and further skewness and the majority of remaining kurtosis is
removed by the conditional prediction rule for heteroscedasticity. The
transformed logarithmic return relatives, after division by their pre-
dicted standard deviations, exhibit kurtosis of 5.52 and skewness of
.2079 (in comparison to values of 3.0 for kurtosis and 0.0 for skewness
in the normal distribution). This result is highly encouraging, in
### TABLE 2

PREDICTIONS FOR HETROSCEDASTICITY
TAKEN FROM FIRST-AND SECOND-PASS REGRESSIONS

Sample Period 1931-1966; N = 315,227

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>( \text{Return} )</th>
<th>( \text{Logarithmic Return} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \text{VAR}(r_{nt} - r_{Mt}) )</td>
<td>( \text{VAR}(r_{nt} - r_{Mt}) )</td>
</tr>
<tr>
<td></td>
<td>1st Pass</td>
<td>2nd Pass</td>
</tr>
<tr>
<td>Constant</td>
<td>1.0000 (0.0105)</td>
<td>1.0000 (0.0049)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( H\beta )</td>
<td>-2.6835 (0.1419)</td>
<td>-2.6864 (0.0653)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( H\sigma^2 )</td>
<td>2.5849 (0.1303)</td>
<td>1.2562 (0.0575)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( H\delta )</td>
<td>-.1836 (0.0121)</td>
<td>-.1333 (0.0058)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( H\beta^2 )</td>
<td>1.3770 (0.0771)</td>
<td>1.1882 (0.0351)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( H\sigma )</td>
<td>-2.9386 (0.2643)</td>
<td>-.7642 (0.1172)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( R^2 )</td>
<td>.0381</td>
<td>.1085</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>Mean</td>
<td>Variance</td>
</tr>
<tr>
<td>----------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td><strong>Rate-of-Return Model</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Market return, $i_{Mt}$</td>
<td>0.015116</td>
<td>0.007747</td>
</tr>
<tr>
<td>Risk-free rate, $i_{FT}$</td>
<td>0.001210</td>
<td>0.000001</td>
</tr>
<tr>
<td>Excess market return, $r_{Mt}$</td>
<td>0.013907</td>
<td>0.007763</td>
</tr>
<tr>
<td><strong>Individual Returns</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_{nt}$</td>
<td>0.013028</td>
<td>0.016496</td>
</tr>
<tr>
<td>$\hat{\theta}<em>{nt} = r</em>{nt} - \beta_{nt} - \beta_{nt} r_{Mt}$</td>
<td>-0.000135</td>
<td>0.010098</td>
</tr>
<tr>
<td>$\hat{\eta}_{nt}/\hat{\sigma}(\cdot)$</td>
<td>-0.000335</td>
<td>0.005726</td>
</tr>
<tr>
<td><strong>Logarithmic Return Model</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Market return, $\ln(1+i_{Mt})$</td>
<td>0.011507</td>
<td>0.006823</td>
</tr>
<tr>
<td>Risk-free rate, $\ln(1+i_{FT})$</td>
<td>0.001208</td>
<td>0.000001</td>
</tr>
<tr>
<td>Excess market return, $r_{Mt}$</td>
<td>0.010299</td>
<td>0.006834</td>
</tr>
<tr>
<td><strong>Individual Returns</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{r}_{nt}$</td>
<td>0.006132</td>
<td>0.013147</td>
</tr>
<tr>
<td>$\hat{r}_{nt}/\hat{\sigma}(\cdot)$</td>
<td>0.006794</td>
<td>0.006659</td>
</tr>
<tr>
<td>$\hat{\theta}<em>{nt} = \hat{r}</em>{nt} - \alpha_{nt} - \beta_{nt} r_{Mt}$</td>
<td>-0.001826</td>
<td>0.007703</td>
</tr>
<tr>
<td>$\hat{\eta}_{nt}/\hat{\sigma}(\cdot)$</td>
<td>-0.000765</td>
<td>0.004795</td>
</tr>
<tr>
<td>$\hat{\rho}_{nt}/\hat{\sigma}(\cdot)$</td>
<td>0.000002</td>
<td>0.007315</td>
</tr>
</tbody>
</table>
| NOTE.—For footnotes, see next page.
FOOTNOTES TO TABLE 3

\(^a\) Skewness \([x] = \frac{E[(x-\overline{x})^3]}{E[(x-\overline{x})^2]^{3/2}};\)
Kurtosis \([x] = \frac{E[(x-\overline{x})^4]}{E[(x-\overline{x})^2]^{2}}.\)

For the normal distribution, skewness = 0.0, and kurtosis = 3.0.

\(^b\) From regression \#3, Table 7.

\(^c\) The notation \(\hat{\sigma}(\cdot)\) denotes division by the square root of the predicted variance for the numerator. The predicted variance is taken from a truncated prediction rule of form (42) and is then standardized so that the geometric average of all predicted variances is unity. Under this standardization, the reduction in variance for the transformed population of returns, relative to the initial population, is another indication of the success in explaining variance fluctuations (see, e.g., Rosenberg (1972)). In this case, the prediction rule for variance is taken from the regression in column (1) of Table 2.

\(^d\) The prediction rule for variance is from column (4) of Table 2.

\(^e\) From regression \#3, Table 8.

\(^f\) The prediction rule for variance is from column (6) of Table 2.

\(^g\) From regressions of form (38), in step 2 of Section III.1.

\(^h\) The prediction rule for variance is from column (6) of Table 2.
comparison with skewness of .4025 and kurtosis of 19.46 for the untransformed logarithmic return relatives, and skewness of 5.633 and kurtosis of 147.45 for the untransformed rates of return.

However, the achieved skewness and kurtosis still differ significantly from normality. The remaining difference may stem from three causes: data error (see Rosenberg and Houglet (1974)); imperfections in the conditional prediction rule for heteroscedasticity, which could potentially be removed by an improved specification of that rule; and deviation from lognormality in the underlying generating process for returns. In later research, we intend to push the analysis of the first two sources further, with the hope that most remaining kurtosis will be explained. For the present study, the results of methods based upon maximum likelihood for the normal distribution must be viewed with caution, since the kurtosis of the residual logarithmic returns, after the achieved correction for heteroscedasticity, is 6.83.

The effect of this nonnormality on the hypothesis tests may be shown to be negligible in a linear model, and the same argument appears to hold for the tests on the variance specification, based on the asymptotic distribution of the likelihood ratio. This argument may be sketched as follows: All hypothesis tests to be used take the form of "t," "F," or likelihood ratio tests, where the exact or approximated sampling distribution of the test statistic is based upon the behavior of the ratio of two mean squares, the numerator being the explained mean square and the denominator being the residual mean square. Skewness and kurtosis affect these mean squares by increasing their sampling variance, altering their
distributions from \( \chi^2 \), and removing their independence. For example, the variance of a mean square increases as \((\text{kurtosis} - 1)\), so that the erroneous assumption of a normal kurtosis of 3, when the actual kurtosis were 6.83, would lead to an understatement of the sampling variance of the mean square by a factor of 2/5.83.

However, because of the very large number of degrees of freedom in the denominator—over 314,700 in all cases—its sampling variance is negligible relative to that of the numerator, so that misstatement of the distribution of the denominator has no noticeable effect on the sampling distribution of the test statistic. This is fortunate, since the kurtosis of the residual returns that enter the denominator is clearly nonnormal, and if the denominator were not effectively an exact estimate, its increased variance and possible dependence with the numerator would have to be taken into account.

Thus, the distribution of the test statistic is determined by the random variables that enter the numerator of the explained sum of squares. These are linear or nonlinear functions of the individual residual returns, but all share the property of being weighted averages of all 315,000 residual returns, with weights that rarely differ by an order of magnitude. Because of this, and thanks to the near independence of residual returns, we can rely on the limiting behavior of sums of independent random variables to insure near normality in the distributions of the terms entering the numerator. As a result, the approximation to the sampling distribution of the test statistics, based upon the assumption of normality, is reliable. This argument would break down if the variances
of the individual residual returns were not finite, but there is no indica-
tion of infinite variance in the population of transformed residual
returns.

Returning to the main theme, the results of Table 3 confirm that
the distributions of market returns and of individual security returns
are much closer to lognormality than to normality. The logarithmic-return
model is therefore chosen for subsequent analysis, although results for
the rate-of-return model will be reported for comparative purposes where
the computations are not overly expensive. However, we stress that much
remains to be done before the logarithmic-return model, with linear condi-
tional expectations as assumed in (A2), can be considered to have been
confirmed. For example, approximate lognormality in the residual returns
is consistent with the model of form \( (1+i_{nt}) = (1+i_{Mt})^{\beta_{nt}}\exp(\alpha_{nt} + \nu_{nt}) \)
that we have employed, but is also consistent with the model \( (1+i_{nt}) =
(1+\alpha_{nt} + \beta_{nt} i_{Mt})\exp(\nu_{nt}) \). We have chosen the former because of the attend-
ant advantage of linearity in the logarithms, and because of an implaus-
able property of the latter—namely, that the conditional expectation of
the compound security return depends on the pattern of market returns over
time, and is even a function of beta over a sample period where the market
return equals the risk-free return. However, we know of no confirmation
of the preferability of the former specification, in terms of goodness of
fit to the data.

The procedure to estimate the conditional prediction rules for \( \phi \)
and \( \sigma^2 \) was next completed for the logarithmic-return model. (The pro-
visional estimates of \( a \) and \( \beta \), on which this procedure depends, are
reported below.) First- and second-pass results are given in Table 2, and the moments of the transformed returns resulting from the second-pass prediction rule are given in Table 3. (The first-pass prediction rule achieved results that were somewhat further from normality.) The kurtosis of residual returns is higher than that for total return, because of the removal from the latter of the market component, that exhibits low kurtosis when transformed by division by the cross-sectional standard deviation of individual security returns, and hence reduces the kurtosis of total return.

It is instructive to compare the results for the two passes in Table 2. The first pass achieves a relatively high $R^2$ and highly significant estimated coefficients. The sampling distribution for these is misstated, however, because of the overlooked heteroscedasticity in the disturbances. The second-pass regression, with the GLS correction, exhibits lower $R^2$. This is not surprising, because of the nature of the disturbance variance. In fact, examination of (41) will show that the optimal GLS transformation (based upon the true relationship) will result in an expected $R^2$ of zero. This is not indicative of a poor fit, because the GLS transformation eliminates the constant term, and an $R^2$ of zero corresponds to perfect explanation of the fluctuation in variance. The improvement in the second pass is confirmed by lower standard errors than the (probably downward biased) standard errors in the first pass.

The first-pass estimators are inefficient but unbiased, in a correctly specified model, and the large sample size (approximately 700 observations per estimated parameter) encourages one to expect that the
estimators will approach the population values. The substantial changes in the estimated coefficients from the first to second passes are therefore disturbing. These changes indicate that the first set of estimated coefficients was rather far from the true values. The fact that the change from the first to the second pass, which is really no more than a change in the weights accorded to the data, brings such a large shift in the estimated coefficients suggests to us that the model is probably misspecified in other ways than those discussed in Section III.1. If we believed that the model had been well-specified, we would have carried out another pass in the hope that the coefficients would stabilize from the second to third passes, and that $R^2$ in the third passes would be near zero. However, suspecting that the model was seriously misspecified, so that a later in-depth analysis was desirable, we chose to defer the rather arduous computations to a later project and to employ the results of the second passes for present purposes.

The magnitude of the improvement in model specification resulting from the conditional prediction rule for heteroscedasticity can be measured in a number of ways: the reduction in skewness and kurtosis already mentioned; the statistical significance of the estimated coefficients in the fitted relation; the improvement in the resulting estimators for the regression coefficients $\alpha$ and $\beta$; or the overall improvement in goodness of fit to the data. The latter can be captured by the improvement of the sample likelihood. Consider the following hypothesis test:
Maintained hypothesis, $H_0$: $a, b$ unconstrained; $\sigma^2_{nt} = \sigma^2$

Alternative hypothesis, $H_1$: $a, b$ unconstrained; $\sigma^2_{nt} = \sigma^2_t (s'x_{nt})$. 

The alternative hypothesis corresponds to conditionally predictable heteroscedasticity and includes 432 cross-sectional average variances, $\sigma^2_t$, and 6 regression coefficients, $s$. The former hypothesis includes only the global average, $\sigma^2$. Thus, the hypotheses are nested, with the maintained hypothesis corresponding to a special case of the alternative hypothesis, where 437 exact constraints are imposed on the parameters. An asymptotic test for the rejection of the maintained hypothesis in favor of the alternative is obtained from the logarithm of the likelihood ratio: specifically, let $\mathcal{L}_1$ and $\mathcal{L}_0$ be the sample likelihoods obtained from maximum likelihood estimators for the parameters of the two hypotheses, and let $df$ denote the number of linearly independent parameters that are constrained under the maintained hypothesis but free under the alternative hypothesis. Then, let $\Delta \mathcal{L} = \ln(\mathcal{L}_1) - \ln(\mathcal{L}_0) = \ln(\mathcal{L}_1 / \mathcal{L}_0)$ be the logarithm of the likelihood ratio. An asymptotic approximation to the distribution of this statistic, under the maintained hypothesis, is that $2\Delta \mathcal{L}$ obeys the $\chi^2$ distribution with $df$ degrees of freedom. Likelihood ratio tests of this form will be used repeatedly in the balance of the paper. In all cases, the critical value will be at the 99.5 percent level.

In the present case, the maximum likelihood estimator for the maintained hypothesis has been computed, so that $\mathcal{L}_0$ is available, but the
estimated, truncated, conditional prediction rule of form (42) is not the maximum likelihood estimator, so that the computed likelihood is a lower bound for $\mathcal{L}_1$. Therefore, the likelihood ratio test is biased against rejection of the maintained hypothesis. Nevertheless, the computed sample likelihood yields $2\Delta \lambda = 125,000$ against a critical point of 520! Thus, there is overwhelming evidence for rejection of the homoscedastic hypothesis in favor of conditionally predictable heteroscedasticity.

In view of the large sample size, it is possible to achieve great statistical significance in estimating a relationship, even though the relationship is not of important magnitude. The next question must therefore be whether the estimated relationship accounts for an important fluctuation in variance across observations. One indication of the importance is provided by the magnitude of the estimated coefficients in the prediction rule. From column 6 of Table 2, the fitted prediction rule is seen to include a quadratic function of $\ln \sigma$, which resembles an approximation to a Bayesian adjustment procedure, plus a substantial negative adjustment for historical yield and a positive adjustment for historical beta. The standard deviation of conditional predictions for $\sigma_{nt}$ is about .03; in other words, 67 percent of the predicted values for the standard deviation of the monthly logarithmic return lie between .04 and .12, with a population median of .09; conversely, 33 percent of the predictions differ from the population median by enough to fall outside this interval.

The next major results concern the possible rejection of the one-factor model in favor of a multiple-factor model with known factor loadings. The multiple-factor models (64) that were estimated include
residual factor variance of the form:

\[
C_t = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0 & G_\beta & G_\beta,\sigma^2 & 0 & 0 & 0 \\
0 & G_\beta,\sigma^2 & G_\beta,\delta & 0 & 0 & 0 \\
0 & G_\beta,\delta & G_\beta,\delta^2 & G_\delta & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\end{pmatrix} + \begin{pmatrix}
\omega_0 & 0 & 0 & 0 & 0 & 0 \\
0 & \omega_\beta & 0 & 0 & 0 & 0 \\
0 & 0 & \omega_\delta & 0 & 0 & 0 \\
0 & 0 & 0 & \omega_\delta & 0 & 0 \\
0 & 0 & 0 & 0 & \omega_\delta & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\end{pmatrix}
\]

The (3 × 3) matrix \( C_t \) introduces a component of variance that is the same in all periods, regardless of the market return. The matrix \( \Omega_\delta \) introduces the component of variance that increases as the square of the market return, and corresponds to random dispersion in the contributions of the descriptors to the beta coefficients.

The impact of the random factors on security returns is graphed in Figure 1. Figure 1(a) shows the standard deviation of the partial relationship between security returns and historical yield in a month when the excess return on the market is zero, so that only the matrix \( G \) makes a contribution to variance. The slope of the relationship is given by the value of the random factor \( g_\delta \), with variance \( G_\delta \), or standard deviation \( G_\delta^{1/2} \). The importance of the factor can be expressed by measuring the standard deviation of return at one standard deviation of \( \hat{H} \delta \) from the population average, equal to

\[
G_\delta^{1/2} \sqrt{E\left( \frac{(\mu_\delta - \hat{\mu}_\delta)^2}{\sigma_\delta} \right)}^{1/2}
\]

This is the standard deviation of the contribution of the factor to the return on a security
The solid line is the expected relationship, with slope $a_\delta + b_\delta r_{Mt}$.

The dashed lines show possible relationships, as determined by the value of the residual factor, with slopes $a_\delta + b_\delta r_{Mt} + g_\delta t$.

The braces $\{\}$ indicate the width of the $\pm$ two-standard-deviation band of resultant returns for a security or portfolio with a given $H_\delta$. The brackets $[\!]$ show the values reported in Table 5.
(or portfolio) for which historical yield is one standard deviation away from the average for all stocks. Figure 1(b) shows the same relationship for the case where the excess market return is equal to its root mean square over the sample history, or .083. Since the market return is positive, and there is a negative partial relationship between beta and the historical yield, the expected value of the slope of the relationship is now downward sloping. The standard deviation of the slope is now equal to \( (C_0 + \frac{\sigma^2}{N^2 M \omega_0})^{1/2} \). In other words, in association with the greater variance in the market, as indicated by the greater market return, the variance of the residual factors is also increased. Table 4 gives the estimated values for five different specifications of the multiple-factor model, ranging from the specification in (85) through a series of restrictions to the diagonal model. For each specification, the number of exact constraints, relative to the more general specification in (85), and the log likelihood ratio are also given. Table 5 reports the likelihood ratio tests for each pair of nested models, and also for tests of these models against the homoscedastic specification. Table 6 reports the MLEs and approximate standard errors for the parameters in (85).

First, to summarize the results of the likelihood ratio tests, the diagonal model is conclusively rejected against any of the alternative specifications of the multiple-factor model. Each of the restricted specifications for the factor variance may be rejected in favor of the most general specification (85) that was tried. In terms of the contribution to goodness of fit, the diagonal entries in \( \Omega_b \) are the most
TABLE 4

ESTIMATED COMPONENTS OF VARIANCE FOR THE MULTIPLE-FACTOR MODEL OF RESIDUAL RETURNS

<table>
<thead>
<tr>
<th>Model No.</th>
<th>Additional Restrictions on the Variance Specification (85)</th>
<th>df</th>
<th>2Δα</th>
<th>100 x Standard Error of Residual Factors, When ( r_M = 0 )</th>
<th>Correlations Between Residual Factors, When ( r_M = 0 )</th>
<th>100 x Standard Error of Residual Factors, When ( \sigma_H = 0.083 = (E(r^2))^{1/2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>As in (85)</td>
<td>--</td>
<td>--</td>
<td>0 x 0.709 x 0.745 x 0.527</td>
<td>(-.281 \quad .288 \quad .270)</td>
<td>(.373 \quad 1.066 \quad 1.416 \quad .717)</td>
</tr>
<tr>
<td>2</td>
<td>G diagonal</td>
<td>3</td>
<td>51.7</td>
<td>0 x 0.647 x 0.684 x 0.503</td>
<td>(0 \quad 0 \quad 0)</td>
<td>(.371 \quad 1.102 \quad 1.438 \quad .722)</td>
</tr>
<tr>
<td>3</td>
<td>(\Omega_0 = 0)</td>
<td>4</td>
<td>2150</td>
<td>0 x 0.996 x 1.322 x 0.686</td>
<td>(-.475 \quad .328 \quad -.167)</td>
<td>(0 \quad 0 \quad 0 \quad 0)</td>
</tr>
<tr>
<td>4</td>
<td>(G = 0)</td>
<td>6</td>
<td>942</td>
<td>0 x 0 x 0 x 0</td>
<td>(0 \quad 0 \quad 0)</td>
<td>(.358 \quad 2.265 \quad 2.35 \quad 1.175)</td>
</tr>
<tr>
<td>5</td>
<td>(G = 0; \Omega_{-b} = 0)</td>
<td>10</td>
<td>19,230</td>
<td>0 x 0 x 0 x 0</td>
<td>(0 \quad 0 \quad 0)</td>
<td>(0 \quad 0 \quad 0 \quad 0)</td>
</tr>
</tbody>
</table>

\(^a\)The prediction rule for heteroscedasticity in residual returns is taken from column 6 of Table 2 for all cases. Thus, specification (5) is the diagonal model with predictable heteroscedasticity.

\(^b\)The magnitude of the residual variance is expressed in terms of the standard deviation of residual return (in percent) per month. Where the residual factor multiplies a nonconstant descriptor, the return is taken as that value for a descriptor differing by one standard deviation from the population average. For instance, under specification (1), in a month when the market return is zero, the residual return associated with the "Hβ factor" leads to a return with standard deviation of 0.709 for a firm with an Hβ value of 1.4 (one standard deviation from the mean), or a standard deviation of \(2 \times 0.709 = 1.418\) for a firm with an Hβ value of 1.8.

\(^c\)I.e., \(\rho_{β,δ} = G_{β,δ} / \sqrt{(G β_δ)^2}\).
<table>
<thead>
<tr>
<th>Alternative Specification</th>
<th>Maintained Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2Δ\ell = 51.7</td>
</tr>
<tr>
<td></td>
<td>\alpha = [12.8]</td>
</tr>
<tr>
<td></td>
<td>df = 3</td>
</tr>
<tr>
<td>2</td>
<td>2Δ\ell = 2,098</td>
</tr>
<tr>
<td></td>
<td>\alpha = [7.9]</td>
</tr>
<tr>
<td></td>
<td>df = 1</td>
</tr>
<tr>
<td>3</td>
<td>2Δ\ell = n.a.</td>
</tr>
<tr>
<td></td>
<td>\alpha = [18.6]</td>
</tr>
<tr>
<td></td>
<td>df = 6</td>
</tr>
<tr>
<td>4</td>
<td>2Δ\ell = 18,288</td>
</tr>
<tr>
<td></td>
<td>\alpha = [14.9]</td>
</tr>
<tr>
<td></td>
<td>df = 4</td>
</tr>
<tr>
<td>5</td>
<td>2Δ\ell = 125,000</td>
</tr>
<tr>
<td></td>
<td>\alpha = [520]</td>
</tr>
<tr>
<td></td>
<td>df = 437</td>
</tr>
</tbody>
</table>

aFor each pair of hypotheses, \( H_1 \) and \( H_0 \), such that \( H_0 \) is a constrained version of \( H_1 \), the log likelihood ratio, 2Δ\ell, the critical value for the 99.5 percent level of confidence, \( \alpha \), and the number of parameters, df, that are constrained under \( H_0 \) but not under \( H_1 \), are given.

bSpecifications 1 through 5 are as stated in Table 4. Specification 6 is the diagonal model with homoscedastic residual returns.

cSpecifications 3 and 4 are not nested.
TABLE 6

VARIANCE COMPONENTS IN THE MULTIPLE-FACTOR MODEL OF RESIDUAL RETURNS:
ML ESTIMATES,\footnote{a} STANDARD ERRORS,\footnote{b} AND ESTIMATION ERROR CORRELATIONS\footnote{b}

<table>
<thead>
<tr>
<th></th>
<th>$G_{\beta}$</th>
<th>$G_{\sigma^2}$</th>
<th>$G_{\delta}$</th>
<th>$G_{\beta,\sigma^2}$</th>
<th>$G_{\beta,\delta}$</th>
<th>$G_{\sigma^2,\delta}$</th>
<th>$\omega_0$</th>
<th>$\omega_{\beta}$</th>
<th>$\omega_{\sigma^2}$</th>
<th>$\omega_{\delta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML Estimates $\times 10^4$</td>
<td>3.117</td>
<td>1.410</td>
<td>.553</td>
<td>-.588</td>
<td>.379</td>
<td>.239</td>
<td>20.021</td>
<td>573.949</td>
<td>542.211</td>
<td>70.445</td>
</tr>
<tr>
<td>Standard Errors of Estimates $\times 10^4$</td>
<td>.391</td>
<td>.106</td>
<td>.058</td>
<td>.111</td>
<td>.064</td>
<td>.062</td>
<td>.142</td>
<td>133.890</td>
<td>53.271</td>
<td>10.611</td>
</tr>
<tr>
<td>Correlation of Estimation Errors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G_{\beta}$</td>
<td>1.</td>
<td>.073</td>
<td>.548</td>
<td>.104</td>
<td>.261</td>
<td>.531</td>
<td>-.619</td>
<td>-.825</td>
<td>-.031</td>
<td>-.197</td>
</tr>
<tr>
<td>$G_{\sigma^2}$</td>
<td>1.</td>
<td>.140</td>
<td>-.186</td>
<td>.014</td>
<td>.026</td>
<td>-.117</td>
<td>-.471</td>
<td>-.487</td>
<td>.236</td>
<td></td>
</tr>
<tr>
<td>$G_{\delta}$</td>
<td>1.</td>
<td>.450</td>
<td>.450</td>
<td>.705</td>
<td>-.521</td>
<td>-.763</td>
<td>.221</td>
<td>-.232</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G_{\beta,\sigma^2}$</td>
<td>1.</td>
<td>.147</td>
<td>.415</td>
<td>-.227</td>
<td>-.339</td>
<td>.203</td>
<td>.018</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G_{\beta,\delta}$</td>
<td>1.</td>
<td>.208</td>
<td>-.142</td>
<td>-.330</td>
<td>-.093</td>
<td>-.032</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G_{\sigma^2,\delta}$</td>
<td>1.</td>
<td>-.488</td>
<td>-.809</td>
<td>.093</td>
<td>.028</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\omega_0$</td>
<td>1.</td>
<td>.761</td>
<td>-.168</td>
<td>.166</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\omega_{\beta}$</td>
<td>1.</td>
<td>-.316</td>
<td>.227</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\omega_{\sigma^2}$</td>
<td>1.</td>
<td>-.254</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\omega_{\delta}$</td>
<td>1.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\footnote{a}{The results reported in Row 1 of Table 5 are a transformation of these values.}

\footnote{b}{From the Hessian of the likelihood function, as in (70).}
important, the diagonal entries in $\Sigma$ the next most important, and the off-diagonal entries in $\tilde{\Sigma}$ the least important. In fact, the maintained hypothesis that the off-diagonal entries in $\tilde{\Sigma}$ are zero, while the diagonal entries in $\tilde{\Sigma}$ and $\Omega_b$ are nonzero, can be rejected with a value of the test statistic that is just four times the critical value at the 99.5 percent level of confidence (51.7 against 12.8), while all the other restrictions on the parameters can be rejected with test statistics that are at least fifty times the critical value.

Next, to summarize the importance of the residual factors, note in Table 4 that the standard deviation of return associated with each factor is about .7 percent per month at minimum, for a security or portfolio having a descriptor value that is one standard deviation from the average. This minimal value occurs when the market return is zero. The standard deviation of monthly return to each factor rises to about 1.0 percent per month in months where the market return is equal to its root mean square and increases further with larger absolute market returns.

The asymptotic approximations to the standard errors of the estimated variances in Table 6 suggest that we can have some confidence in the order of magnitude of these variances. The specification could be improved by inclusion of residual factors for the other descriptors $H^2$ and $HG$. It was our expectation, however, that, because of high collinearity between these variables and the previous descriptors, the more general model would be little distinguished from the one we have employed, either in terms of improvement to the sample likelihood or in terms of
the implied estimators and sampling distribution for \( a \) and \( b \). As will be seen in the next section, these estimators are relatively unaffected by the variance specification.

IV.2 Prediction of \( a \) and \( b \)

To illustrate the effects of the specification of the components of variance in the model, prediction rules for \( a \) and \( b \) will be reported for a series of models with progressively more sophisticated treatment of variances. Tables 7 and 8 report the most naive possible model, the diagonal model with homoskedastic variance (Assumptions A6 and A7)—Table 7 for the rate-of-return model, Table 8 for the logarithmic-return model. Tables 9 and 10 report the GLS estimates for the diagonal model with predictable heteroskedasticity (Assumptions A6 and A5), for rate-of-return and logarithmic-return models, respectively. In each case, three regressions are reported: #1 is the estimated prediction rule with the constant and \( H\beta \) as the only descriptors; #2 is the prediction rule with \( H\delta \) added; and #3 includes all six descriptors. The unconstrained regressions are reported at the top of each panel, and the constrained regressions that maximize the sample likelihood, subject to the CAP:2 (specified in (6a)) are reported in the bottom panel. The unconstrained regressions were conducted by an ordinary regression program, in the form of "market-conditional" regressions (44). The constrained regressions were carried out by maximization of the nonlinear functions of \( k \),

\[
\ell(k) = \frac{N}{2} (\ln N - 1) - \ln \sqrt{2\pi} - \frac{N}{2} \ln \left( \frac{\prod_{t=1}^{T} e^{y_{t} - \mu}}{\prod_{t=1}^{T} e^{y_{t} - \mu}} \right),
\]

(83)
<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Regression #1</th>
<th>Regression #2</th>
<th>Regression #3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{a} \times 10^2 )</td>
<td>( \hat{a} \times 10^2 )</td>
<td>( \hat{a} \times 10^2 )</td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>.0072 (0.0182)</td>
<td>.0072 (0.0182)</td>
<td>.0072 (0.0181)</td>
</tr>
<tr>
<td></td>
<td>1.0003 (0.0024)</td>
<td>1.0003 (0.0024)</td>
<td>1.0003 (0.0023)</td>
</tr>
<tr>
<td>( H\beta )</td>
<td>-.4812 (0.0458)</td>
<td>-.4396 (0.0483)</td>
<td>-.5816 (0.1994)</td>
</tr>
<tr>
<td></td>
<td>.8030 (0.0065)</td>
<td>.6422 (0.0071)</td>
<td>.7098 (0.0302)</td>
</tr>
<tr>
<td>( H\sigma^2 )</td>
<td>-.3425 (0.1646)</td>
<td>-.0449 (0.0275)</td>
<td>-.0075 (0.0291)</td>
</tr>
<tr>
<td></td>
<td>.0178 (0.0259)</td>
<td>-.1427 (0.0026)</td>
<td>-.11236 (0.0028)</td>
</tr>
<tr>
<td>( H\delta )</td>
<td>-.0106 (0.1037)</td>
<td>-.0106 (0.1037)</td>
<td>-.0106 (0.1037)</td>
</tr>
<tr>
<td></td>
<td>-.0106 (0.0160)</td>
<td>-.0106 (0.0160)</td>
<td>-.0106 (0.0160)</td>
</tr>
<tr>
<td>( H\sigma^2 )</td>
<td>1.0893 (0.3506)</td>
<td>1.0893 (0.3506)</td>
<td>1.0893 (0.3506)</td>
</tr>
<tr>
<td>( \hat{R}^2 )</td>
<td>.3761</td>
<td>.3876</td>
<td>.3895</td>
</tr>
</tbody>
</table>

**Table 7**

*Market Conditional Regressions, Rate-of-Return Model, with No Heteroscedasticity Correction*

Sample Period = 1931-1966; N = 315,227
TABLE 8
MARKET-CONDITIONAL, LOGARITHMIC RETURN REGRESSIONS,
WITH NO HETEROSEDASTICITY CORRECTION
Sample Period 1931-1966; N = 315,227

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Regression #1</th>
<th>Regression #2</th>
<th>Regression #3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(a \times 10^2) &amp; (b)</td>
<td>(a \times 10^2) &amp; (b)</td>
<td>(a \times 10^2) &amp; (b)</td>
</tr>
<tr>
<td>Constant</td>
<td>(-.3875) (.0158) &amp; .9801 (.0022)</td>
<td>(-.3875) (.0158) &amp; .9801 (.0022)</td>
<td>(-.3875) (.0158) &amp; .9801 (.0022)</td>
</tr>
<tr>
<td>(Hβ)</td>
<td>(-.8190) (.0938) &amp; .7562 (.0059)</td>
<td>(-.6162) (.0421) &amp; .6422 (.0065)</td>
<td>(-.5962) (.1737) &amp; .7821 (.0235)</td>
</tr>
<tr>
<td>(Ho^2)</td>
<td>(-.3853) (.1433) &amp; -.0722 (.0235)</td>
<td>(-.071) (.0903) &amp; -.1126 (.0146)</td>
<td></td>
</tr>
<tr>
<td>(Hδ)</td>
<td>.2707 (.0239) &amp; -.1047 (.0025)</td>
<td>.2403 (.0253) &amp; -.0875 (.0027)</td>
<td></td>
</tr>
<tr>
<td>(Hβ^2)</td>
<td>(.0071) (.0903) &amp; -.1126 (.0146)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Hσ)</td>
<td>.6175 (.3053) &amp; .3215 (.0486)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(R^2)</td>
<td>.4100</td>
<td>.4134</td>
<td>.4142</td>
</tr>
</tbody>
</table>

| Constant   | \(.9798\) (.0022) | \(.9806\) (.0022) | \(.9804\) (.0022) |
| \(Hβ\)     | \(.7563\) (.0059) | \(.6427\) (.0065) | \(.7806\) (.0276) |
| \(Ho^2\)   | \(-.0708\) (.0235) |
| \(Hδ\)     | \(-.1042\) (.0025) | \(-.0870\) (.0027) |
| \(Hβ^2\)   | \(-.1115\) (.0146) |
| \(Hσ\)     | \(.3192\) (.0490) |
| \(k\)      | \(.0099\) (.0005) | \(.0109\) (.0006) | \(.0107\) (.0005) |
TABLE 9
MARKET-CONDITIONAL REGRESSIONS, RATE-OF-RETURN MODEL, WITH HETEROSCEDASTICITY CORRECTION FROM COLUMN 1, TABLE 2
Sample Period 1931-1966; N = 315,227

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Regression #1</th>
<th>Regression #2</th>
<th>Regression #3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{a} \times 10^2$</td>
<td>$\hat{b}$</td>
<td>$\hat{a} \times 10^2$</td>
</tr>
<tr>
<td>Constant</td>
<td>.0920</td>
<td>.9489</td>
<td>.0386</td>
</tr>
<tr>
<td></td>
<td>(.0107)</td>
<td>(.0023)</td>
<td>(.0114)</td>
</tr>
<tr>
<td>$H^2$</td>
<td>-.5358</td>
<td>.6880</td>
<td>-.4529</td>
</tr>
<tr>
<td></td>
<td>(.0339)</td>
<td>(.0075)</td>
<td>(.0342)</td>
</tr>
<tr>
<td>$H^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H^6$</td>
<td></td>
<td></td>
<td>.2142</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.0190)</td>
</tr>
<tr>
<td>$H^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H^6$</td>
<td></td>
<td></td>
<td>.4867</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.2766)</td>
</tr>
<tr>
<td>$r^2$</td>
<td>.3463</td>
<td></td>
<td>.3485</td>
</tr>
<tr>
<td>Constant</td>
<td>.9474</td>
<td>.9779</td>
<td>.9943</td>
</tr>
<tr>
<td>$H^2$</td>
<td>.6864</td>
<td>.6241</td>
<td>.6080</td>
</tr>
<tr>
<td>$H^2$</td>
<td></td>
<td></td>
<td>-.2128</td>
</tr>
<tr>
<td>$H^6$</td>
<td></td>
<td>-.0971</td>
<td>-.0843</td>
</tr>
<tr>
<td>$H^2$</td>
<td></td>
<td></td>
<td>-.0359</td>
</tr>
<tr>
<td>$H^6$</td>
<td></td>
<td></td>
<td>.6208</td>
</tr>
<tr>
<td>$k$</td>
<td>.0085</td>
<td>.0093</td>
<td>.0090</td>
</tr>
<tr>
<td>Descriptor</td>
<td>Regression #1</td>
<td>Regression #2</td>
<td>Regression #3</td>
</tr>
<tr>
<td>------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
</tr>
<tr>
<td></td>
<td>$\hat{a} \times 10^2$</td>
<td>$\hat{b}$</td>
<td>$\hat{a} \times 10^2$</td>
</tr>
<tr>
<td>Constant</td>
<td>-.1180</td>
<td>.9519</td>
<td>-.1710</td>
</tr>
<tr>
<td></td>
<td>(.0109)</td>
<td>(.0024)</td>
<td>(.0113)</td>
</tr>
<tr>
<td>$H_B$</td>
<td>-.6945</td>
<td>.6987</td>
<td>-.5774</td>
</tr>
<tr>
<td></td>
<td>(.0319)</td>
<td>(.0072)</td>
<td>(.0323)</td>
</tr>
<tr>
<td>$H_O^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_D$</td>
<td></td>
<td></td>
<td>.3059</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.0196)</td>
</tr>
<tr>
<td>$H_B^2$</td>
<td></td>
<td>.0587</td>
<td>.0587</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0072)</td>
<td>(.0074)</td>
</tr>
<tr>
<td>$H_O$</td>
<td></td>
<td>.4399</td>
<td>.5757</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.3455</td>
<td>.3480</td>
<td>.3489</td>
</tr>
</tbody>
</table>

Sample Period 1931-1966; N = 315,227
where \( N = \sum_{t=1}^{T} N_t \) and \( e_{t,t} = x_{t,t} - 1^k - (1-k)x_{t,t} \).

For any given \( k \), the MLE of \( b \) is given by:

\[
(87) \quad b = \left( \sum_{t=1}^{T} x_{t,t}^* v_{t,t}^{-1} x_{t,t}^* \right)^{-1} \left( \sum_{t=1}^{T} x_{t,t}^* v_{t,t}^{-1} (x_{t,t} - 1) \right),
\]

where: \( x_{t,t}^* = x_{Mt,t} - k x_{t,t} \).

Then, \( a \) is given in terms of \( b \) by (6a), and the estimation error variance matrix, \( M_b \), is:

\[
(88) \quad M_b = s^2 \left( \sum_{t=1}^{T} x_{t,t}^* v_{t,t}^{-1} x_{t,t}^* \right)^{-1},
\]

where: \( s^2 = \frac{1}{N} \sum_{t=1}^{T} e_{t,t}^* v_{t,t}^{-1} e_{t,t} \).

The standard error for \( k \) is found by the asymptotic approximation that, in general, for a parameter vector \( \hat{p} \) of dimension \( df \), confidence regions for \( \hat{p} \) may be obtained for the relation that:

\[
(89) \quad 2[\hat{\lambda}_{\text{ML}} - \lambda(p)] \sim \chi^2_{df}.
\]

In this case, with \( df = 1 \), the asymptotic approximation to the sampling distribution for \( k \) is easily obtained by computing \( 2\hat{\lambda}(k) \) for a grid of values of \( k \). Figure 2 shows an illustrative example for regression #3 of Table 10.
FIGURE 2

PLOT OF \(2\Delta \lambda(k)\) FOR REGRESSION #3, TABLE 10, SHOWING THE IMPLIED DISTRIBUTION FOR \(k - \hat{\lambda}_{ML}\), WHICH EFFECTIVELY COINCIDES WITH \(\mathcal{N}(0, 0.00465)\).

The confidence regions that are shown are determined by \(F(x)\), the cumulative distribution for the \(\chi^2_1\) distribution, where \(F(1) = .683\) and \(F(7.87) = .995\).
The estimate of \( k \) is primarily determined by the ratios \( \hat{\alpha}_\beta / \hat{\beta} \) and \( \hat{\alpha}_\delta / \hat{\delta} \), since these are the most precisely estimated ratios. \( k \) would be substantially smaller if \( b_\beta \) were 1 as in naive prediction of \( \beta \). The reported estimates for \( k \) nearly equal the mean excess market return. For the rate-of-return model, \( \hat{k} \) is estimated at .0093 per month in regression #3 of table 9, against mean excess market return of .0139, leaving the difference of .0045 per month as the estimated compensation for systematic risk. For the logarithmic-return model, the estimate \( \hat{k} = .0105 \) in Regression #3 of Table 10 almost exactly offsets the mean excess market return of .0103, so that, with an estimated standard error of estimate of .0005, the null hypothesis of zero compensation for systematic risk cannot be rejected; but judgment on this point must be deferred until the results of more sophisticated models are reported below.

We have carried on further analysis for the logarithmic-return model only, because the cost of continued analysis of the rate-of-return model did not seem justifiable in view of the inferiority of the normal approximation to the rates of return.

The first two columns of the top panel of Table 11 report the maximum likelihood estimators (69) for \( a \) and \( b \), in the multiple-factor model as estimated in Table 6. The estimates are close to those in Table 10, Regression #3, where the variance correction for residual factors was omitted; and the signs and relative magnitudes of the estimates are similar to those in Table 8, where all corrections for heteroscedasticity were omitted. This robustness against misspecification of the residual
TABLE II
LOGARITHMIC-RETURN MODEL: ESTIMATED SYSTEMS OF PREDICTION RULES FOR THE MULTIPLE-FACTOR MODEL WITH PREDICTABLE HETEROSCEDASTICITY

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>$\hat{\alpha} \times 10^2$</th>
<th>$\hat{\beta}$</th>
<th>$\hat{\gamma}$</th>
<th>$\hat{\delta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unconstrained Model$^a$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-.2160</td>
<td>.9895</td>
<td>.9914</td>
<td>-.0054</td>
</tr>
<tr>
<td></td>
<td>(.0145)</td>
<td>(.0040)</td>
<td>(.0041)</td>
<td>(.0036)</td>
</tr>
<tr>
<td>$H_\beta$</td>
<td>-.4194</td>
<td>.7045</td>
<td>-.0358</td>
<td>.1225</td>
</tr>
<tr>
<td></td>
<td>(.1572)</td>
<td>(.0345)</td>
<td>(.0354)</td>
<td>(.0396)</td>
</tr>
<tr>
<td>$H_\sigma^2$</td>
<td>-.4648</td>
<td>-.1558</td>
<td>.8441</td>
<td>-.0743</td>
</tr>
<tr>
<td></td>
<td>(.1442)</td>
<td>(.0330)</td>
<td>(.0358)</td>
<td>(.0327)</td>
</tr>
<tr>
<td>$H_\delta$</td>
<td>.2142</td>
<td>-.0761</td>
<td>-.1034</td>
<td>.7126</td>
</tr>
<tr>
<td></td>
<td>(.0479)</td>
<td>(.0096)</td>
<td>(.0036)</td>
<td>(.0058)</td>
</tr>
<tr>
<td>$H_\beta^2$</td>
<td>-.0460</td>
<td>-.0576</td>
<td>.1472</td>
<td>-.0771</td>
</tr>
<tr>
<td></td>
<td>(.0652)</td>
<td>(.0144)</td>
<td>(.0225)</td>
<td>(.0206)</td>
</tr>
<tr>
<td>$H_\sigma$</td>
<td>.4446</td>
<td>.5362</td>
<td>-.2384</td>
<td>.0001</td>
</tr>
<tr>
<td></td>
<td>(.2291)</td>
<td>(.0496)</td>
<td>(.0627)</td>
<td>(.0695)</td>
</tr>
</tbody>
</table>

Model$^b$ Constrained to MLEs of $c_\beta$, $c_\phi$, and $c_\delta$

| | $H_\beta$ | .7005 | -.0351 | .1210 |
| | $H_\alpha^2$ | -.2698 | -.1559 | .8509 | -.0797 |
| | $H_\delta$ | .2257 | -.0762 | -.1034 | .7126 |
| | $H_\beta^2$ | -.0499 | -.0575 | .1476 | -.0768 |
| | $H_\sigma$ | -.1026 | .5403 | -.2518 | .0122 |

MLEs of $c_\beta$, $c_\phi$, and $c_\delta$ and Estimation Error Variances

| | Estimate | Std. Error | Error Correlations |
| | $c_\beta$ | -.00364 | (.00159) | 1 | .457 | -.214 |
| | $c_\phi$ | -.275 | (.086) | 1 | .405 |
| | $c_\delta$ | .563 | (.190) | 1 |

$^a$MLEs for $a$ and $b$ from (69), for $h$ from Table 2, Column 4, and for $d$ from (84).

$^b$MLEs from maximization of (43), and estimation error moments from the Hessian of (43).
variance is encouraging, for it suggests that any remaining misspecifications may not importantly affect the results.

The estimated standard errors of \( a \) and \( b \) in Table 11 are quite similar to those in Table 10, except for \( a_0 \) and \( b_0 \), where the standard errors increase by a factor of 2.5. The true standard errors for Table 10 were larger because a less efficient estimator was in use, but the computed values were biased because of the misspecified residual variance on which their computation was based. Typically, the bias is downward, and this was clearly the case for the contribution of \( H_0 \), since the apparent standard error increases although the new estimator is actually more efficient. For all descriptors except \( H_0 \), the increase in efficiency approximately offsets the bias in the previous standard errors.

The upper panel of Table 11 is completed by restating the prediction rules for \( \phi \) and \( \delta \) that were reported previously. This system of prediction rules is then subjected to the hypothesis-testing procedure stated in (43), for which the estimators and their error variance matrices (not reported here for want of space) are the ingredients. The procedure in (43) applies to the set of all estimated coefficients in the prediction rules except the constant terms. (The constant terms simply set the sample mean values, because all descriptors are taken as differences from their means.) Accordingly, the hypothesis test consists in subjecting the twenty coefficients for the five nonconstant descriptors to the five constraints implied by the CAP hypotheses. The lower panel of Table 11 reports the maximum likelihood estimators for \( c_B \), \( c_0 \), and \( c_1 \), and for
\(a, b, d, h\), obtained by the imposition of (43). Since three parameters 
\((c_\beta, c_\delta, c_\phi)\) are free in the five constraints, only two degrees of freedom are actually sacrificed by the imposition of (43). In Table 12, it is seen that the constraint (43) cannot quite be rejected at the 99.5 percent level \((2\Delta \lambda = 9.015\ \text{against a critical point of 10.6 for} \ \chi^2_2)\).

Tests of more restrictive CAP hypotheses, implemented by further constraints in (43), are also reported in Table 12. Both \(c_\delta = 0\) and \(c_\phi = 0\) can be rejected at the 99.5 percent level, but \(c_\phi = -.5\), as suggested in the CAP:L cannot quite be rejected against the alternative hypothesis that \(c_\phi\) (as well as \(c_\delta\) and \(c_\beta\)) is unconstrained. The sixth hypothesis, \(c_\delta = c_\phi = 0\), coincides with the previously reported hypothesis CAP:Z. The MLE for \(c_\beta\) under this hypothesis is \(-.00923\), implying \(\hat{\kappa} = .00923\), quite similar to the earlier estimates. However, the hypothesis can be rejected against all alternative hypotheses that \(c_\phi\) and \(c_\delta\) are unequal to zero, and under the alternative hypotheses, \(\hat{\kappa}\) falls sharply.

As the error correlations in the lower panel of Table 11 indicate, there is nonnegligible multicollinearity among the prediction rules, and both the negative compensation attribute to total risk, \(c_\phi < 0\), and the positive compensation for yield, \(c_\delta > 0\), result in increased compensation attributed to systematic risk—that is, to an increase in \(c_\beta\). It follows that any test of the CAP:Z lacks robustness against misspecification of the compensations to yield and total risk, and that, consequently, a test of this hypothesis in isolation is ill-advised. Instead, it must be tested in the context of all plausible hypotheses.


**TABLE 12**

Tests of the CAP Hypotheses, \( \alpha = c_\beta \beta + c_\phi \phi + c_\delta \delta \),

In the logarithmic multiple-factor model with predictable heteroscedasticity

<table>
<thead>
<tr>
<th>Hypothesis: ( H_1 )</th>
<th>( \alpha ) is not a linear function of ( \beta, \phi, \delta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_2 )</td>
<td>( c_\beta, c_\delta, c_\phi ) unconstrained</td>
</tr>
<tr>
<td>( H_3 )</td>
<td>( c_\beta, c_\delta ) unconstrained; ( c_\phi = -0.5 )</td>
</tr>
<tr>
<td>( H_4 )</td>
<td>( c_\beta, c_\delta ) unconstrained; ( c_\phi = 0 )</td>
</tr>
<tr>
<td>( H_5 )</td>
<td>( c_\beta, c_\phi ) unconstrained; ( c_\delta = 0 )</td>
</tr>
<tr>
<td>( H_6 )</td>
<td>( c_\beta ) unconstrained; ( c_\phi = c_\delta = 0 ); i.e., ( \alpha = 0 )</td>
</tr>
<tr>
<td>( H_7 )</td>
<td>( c_\beta = c_\phi = c_\delta = 0 ); i.e., ( \alpha = 0 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alternative Hypothesis</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ( 2 \Delta \lambda )</td>
<td>9.015</td>
<td>15.807</td>
<td>19.279</td>
<td>18.028</td>
<td>41.510</td>
<td>98.555</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>[10.6]</td>
<td>[12.8]</td>
<td>[12.8]</td>
<td>[12.8]</td>
<td>[14.9]</td>
<td>[16.8]</td>
</tr>
<tr>
<td>( df )</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2 ( 2 \Delta \lambda )</td>
<td>6.792</td>
<td>10.264</td>
<td>9.013</td>
<td>32.495</td>
<td>89.540</td>
<td></td>
</tr>
<tr>
<td>( \alpha )</td>
<td>[7.9]</td>
<td>[7.9]</td>
<td>[7.9]</td>
<td>[10.6]</td>
<td>[12.8]</td>
<td></td>
</tr>
<tr>
<td>( df )</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3 ( 2 \Delta \lambda )</td>
<td>( \alpha )</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>( df )</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4 ( 2 \Delta \lambda )</td>
<td>( 22.231 )</td>
<td>79.276</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha )</td>
<td>NA</td>
<td>[7.9]</td>
<td>[10.6]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( df )</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 ( 2 \Delta \lambda )</td>
<td>( 23.482 )</td>
<td>80.527</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha )</td>
<td>[7.9]</td>
<td>[10.6]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( df )</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 ( 2 \Delta \lambda )</td>
<td>( 57.045 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha )</td>
<td>[7.9]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( df )</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Estimated Coefficients Under Maintained Hypothesis**

| \( c_\beta \) | \( -0.00364 \) | \( -0.00172 \) | \( -0.00592 \) | \( -0.00464 \) | \( -0.00923 \) | 0. |
| \( c_\phi \) | \( -0.275 \) | \( -0.5 \) | 0. | \( -0.380 \) | 0. | 0. |
| \( c_\delta \) | \( 0.563 \) | \( 0.368 \) | \( 0.808 \) | 0. | 0. | 0. |

\(^a\)See Table 5 for explanation. All hypotheses are implemented by maximization of (43), subject to the appropriate constraints.
IV.3 A Test of Misspecification

The estimated error variances of $$\hat{\epsilon}_{t}$$ and $$\hat{\beta}$$ are determined by the estimated variances of $$\hat{\epsilon}_{t} = a - bx_{t}, t=1,\ldots,T$$, which are assumed to be of the form $$\sigma^{2}(G + r^{2} \Omega_{b} + M_{t})$$. These are the components of variance in security returns that lie in the subspaces spanned by $$x_{t}, t=1,\ldots,T$$. In the approach developed thus far, the model was fitted by maximum likelihood in the space of security returns. $$G$$ and $$\Omega_{b}$$ were estimated, subject to certain constraints on their structure (only 10 parameters out of a possible 30 are free, with the others assumed to be zero), so as to explain best the excess of the components of variance in the subspaces spanned by $$x_{t}$$, over that variance attributable to the components of specific risk in these subspaces. The contribution of specific risk, in turn, was estimated so as to explain the variance components in the entire space of security returns. The largest part of this information comes from the subspaces orthogonal (with respect to the norm $$\Sigma_{t}$$) to $$x_{t}$$, for these subspaces are of dimension $$N - J$$, whereas the subspace spanned by $$x_{t}$$ is of dimension $$J$$ only.

This approach is the efficient one when the model (64) is correctly specified. However, if (64) is misspecified, it is possible for the variance components in the spaces spanned by the $$x_{t}$$ to be misestimated. The free components of $$G$$ and $$\Omega_{b}$$ will correctly capture the variances in the corresponding dimensions of the subspaces, but in the remaining dimensions, the scale of variance will be set, not by the components of variance in these dimensions, but rather by the typical
components of variance in the much larger orthogonal subspaces. Put in
another way, the relative magnitude of specific risk in the spaces
spanned by the $X_t$, relative to the variance of specific risk in the or-
thogonal subspaces, is fixed by the model of heteroscedasticity of spe-
cific risk, and the estimated scale of specific risk is overwhelmingly
determined by the orthogonal subspaces. Consequently, if the model in-
appropriately specifies the relative magnitudes, the variance in the
subspaces spanned by the $X_t$ will be misestimated.

As a robust check of the importance of this problem, it is de-
sirable to reestimate directly the scale of variance in the subspaces
spanned by the $X_t$. This is accomplished by using the model (71) and
(72), with no sequential variation—that is, with $\hat{\eta}_t \equiv 0$ in (72).
This model is exactly equivalent to (64) in its implications for the
components of variance in the subspaces spanned by the $X_t$ or, in other
words, for the variances of the $f_t$. As a check against misspecification,
the relative magnitudes of the estimates $\hat{G}, \hat{\Omega}_b$, and $\hat{M}_t$ obtained
from Section IV.2 may be retained, and an optimal scale adjustment may
be estimated by maximum likelihood, such that

$$\text{VAR}(\hat{f}_t - \hat{a} - \hat{b}_t \hat{M}_t) = \sigma_a^2 \left( \hat{G} + r^2 \hat{\Omega}_b + \hat{M}_t \right).$$

This is easily accomplished as a trivial case of

the procedure for estimation in the sequentially varying parameter model.
We found an MLE of $\hat{\sigma}_a^2 = 1.52$. Since the expected value in the absence
of misspecification is 1.0, and there are 2,580 degrees of freedom, mis-
specification is clearly proved. This is somewhat disappointing, al-
though not surprising in view of our doubts about the prediction rule
for heteroscedasticity. Until the misspecification is corrected, a more
accurate sampling theory for the estimators is obtained by using these
adjusted estimates of the variance of the $\hat{f}_t$. These imply an increase
of 1.52 in variance, or 1.25 in standard error, for the estimates of $a$
and $b$.

IV.4 Sequential Parameter Variation

The remaining step is to introduce sequential variation in the
parameters. The specification (73) was chosen to allow nonstationarity
in the intercept $k$, in the contribution of descriptors to expected re-
turn $a$, and to the covariance with the market $b$. There are ten spec-
ified parameters to be estimated for sequential variation:

$$\phi_k, q_k, \phi_a, q_a, \beta, q_{a, \sigma^2}, q_a, \delta; \phi_b, q_b, \beta, q_{b, \sigma^2}, q_b, \delta.$$ 

Also, the specification parameters for residual variance $\sigma^2$, $\Omega_b$, and $s$
should ideally be jointly estimated by maximum likelihood. However,
knowing, from a priori reasoning, that the contribution of the sequen-
tially varying components to total variance is very small relative to
residual variance, we concluded that the estimates of residual variance
would change little after the estimation of sequential variation. Ac-
cordingly, the relative magnitudes of $\hat{G}$, $\hat{\Omega}_b$, and $\hat{M}_t$, $t=1, T$
were
accepted as fixed, and the overall scale parameter $\sigma^2_a$ was again es-
timated jointly with the specification for sequential variation. The
estimated adjustment, $\sigma^2_a$, did remain at 1.52, as we expected.
This simplification leaves ten parameters over which the likelihood function must be optimized, a relatively large number for an optimization algorithm when analytical partial derivatives are not available. Nevertheless, convergence was achieved without great difficulty.

The improvement in sample likelihood at the MLEs for sequential variation, against the maintained hypothesis of no sequential variation, was $2\Delta \ell = 14.51$. With df = 10, this allows rejection of stationarity at the 80 percent level of confidence only, according to the asymptotic approximation that $2\Delta \ell \sim \chi^2_{10}$. The MLEs are reported in Table 13. It is impossible to identify nonstationarity in the contribution to $\alpha$, with the likelihood function almost flat over the range that might be considered reasonable, a priori. The optimal estimate of $\phi_a = .17$, corresponding to an extremely rapid convergence rate, suggests first-order serial correlation in the factor returns, a violation of Assumption (A10), but the effect is far from statistically significant. For $k$ and for $\beta$, there is more information in the sample. In fact, if the improvement in sample likelihood, which is entirely due to these terms, is attributed to the six parameters related to $k$ and $\beta$, stationarity may be rejected at the 97.5 percent level of confidence. No significant nonstationarity in $b_\delta$ is found, but nonstationarity in $k$, $b_\beta$, and $\sigma^2$ is substantial enough to warrant attention. Applying the optimal "smoothing" formulas (Rosenberg and Ohlsen, 1975) for the sequentially varying parameters $k^*$, $b_{*\beta}$, and $\sigma^{*2}$, we obtain the time series of confidence regions in Figure 3. There is apparent nonstationarity in the
TABLE 13

MAXIMUM LIKELIHOOD ESTIMATES OF SEQUENTIAL VARIATION

<table>
<thead>
<tr>
<th>Convergence Rate</th>
<th>Incremental Standard Deviation</th>
<th>Stationary Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\phi}_k = 0.878$</td>
<td>$\sqrt{\hat{\sigma}_k} = 0.0022$</td>
<td>$\sqrt{\hat{\sigma}_k/(1-\hat{\phi}_k^2)} = 0.000000$</td>
</tr>
<tr>
<td>$\sqrt{\hat{q}_{a,\beta}} = 0.000000$</td>
<td>$\sqrt{\hat{q}_{a,\beta}/(1-\hat{\phi}_a^2)} = 0.000000$</td>
<td></td>
</tr>
<tr>
<td>$\hat{\phi}_a = 0.179$</td>
<td>$\sqrt{\hat{q}_{a,\sigma^2}} = 0.000001$</td>
<td>$\sqrt{\hat{q}_{a,\sigma^2}/(1-\hat{\phi}_a^2)} = 0.000001$</td>
</tr>
<tr>
<td>$\sqrt{\hat{q}_{a,\delta}} = 0.000000$</td>
<td>$\sqrt{\hat{q}_{a,\delta}/(1-\hat{\phi}_a^2)} = 0.000000$</td>
<td></td>
</tr>
<tr>
<td>$\sqrt{\hat{q}_{b,\beta}} = 0.0304$</td>
<td>$\sqrt{\hat{q}_{b,\beta}/(1-\hat{\phi}_b^2)} = 0.1019$</td>
<td></td>
</tr>
<tr>
<td>$\hat{\phi}_b = 0.954$</td>
<td>$\sqrt{\hat{q}_{b,\sigma^2}} = 0.0228$</td>
<td>$\sqrt{\hat{q}_{b,\sigma^2}/(1-\hat{\phi}_b^2)} = 0.0750$</td>
</tr>
<tr>
<td>$\sqrt{\hat{q}_{b,\delta}} = 0.0116$</td>
<td>$\sqrt{\hat{q}_{b,\delta}/(1-\hat{\phi}_b^2)} = 0.0381$</td>
<td></td>
</tr>
</tbody>
</table>
FIGURE 3

Plots of the two standard deviation bands (68.3% confidence regions) for the sequentially varying components of \( b_0, b_0^2, \) and \( k. \)

The MLEs lie at the midpoints of the confidence region.
prediction rule for $\beta$, which can hopefully be removed by improved specification of the prediction rule. The nonstationarity in $k$ suggests a search for an improved measure of the risk-free return.

Since it is impossible to identify nonstationarity in $\alpha$, the appealing possibility of a test of adherence to the CAP hypotheses in the sequential variation of $\alpha$, $k$, and $\beta$ is ruled out. Thus, tests of CAP hypotheses are restricted to tests on the mean parameters $\underline{\alpha}$ and $\underline{\beta}$, as before. However, these tests are improved by the more accurate specification of residual variance provided by the model of sequential variation. Tables 14 and 15 report the hypotheses tests in the context of sequential parameter variation, in identical format to Tables 11 and 12. The estimates for $\underline{\beta}$ and $\underline{\alpha}$ are now taken from (83) and (69). The results are the most reliable, in our view, among those we have obtained, since they reflect the most accurate specification of residual variance. There is little change in the estimated coefficients. There is a reduction in the significance of the test statistics, partly due to the increased estimate of residual variance, with $\hat{\sigma}^2_a = 1.52$, and partly due to residual correlation stemming from sequential parameter variation.

The linear constraint implied by the CAP hypotheses cannot be rejected at the 90 percent level of confidence, nor can the additional constraint $c_\phi = -.5$. The alternative constraint $c_\phi = 0$ can be rejected at the 99 percent confidence level. The constraint $c_\delta = 0$ can be rejected at the 97.5 percent confidence level. The twin constraints $c_\phi = c_\delta = 0$, equivalent to the CAP:Z are strongly rejected at the 99.5 percent level, as is the further constraint, equivalent to the CAP:M, $c_\phi = c_\delta = 0$. 
**TABLE 14**

**ESTIMATED SYSTEMS OF PREDICTION RULES FOR THE LOGARITHMIC MULTIPLE-FACTOR MODEL WITH PREDICTABLE HETEROSEDASTICITY AND SEQUENTIAL PARAMETER VARIATION**

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>$\hat{a} \times 10^2$</th>
<th>$\hat{b}$</th>
<th>$\hat{h}$</th>
<th>$\hat{d}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unconstrained Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>$.2201$</td>
<td>$.9892$</td>
<td>$.9914$</td>
<td>$-.0054$</td>
</tr>
<tr>
<td></td>
<td>$(.0181)$</td>
<td>$(.0049)$</td>
<td>$(.0041)$</td>
<td>$(.0036)$</td>
</tr>
<tr>
<td>$H_B$</td>
<td>$-.3877$</td>
<td>$.5912$</td>
<td>$-.0358$</td>
<td>$.1225$</td>
</tr>
<tr>
<td></td>
<td>$(.2175)$</td>
<td>$(.0588)$</td>
<td>$(.0354)$</td>
<td>$(.0396)$</td>
</tr>
<tr>
<td>$H_\sigma^2$</td>
<td>$-.4724$</td>
<td>$-.1419$</td>
<td>$.8441$</td>
<td>$-.0743$</td>
</tr>
<tr>
<td></td>
<td>$(.1819)$</td>
<td>$(.0503)$</td>
<td>$(.0358)$</td>
<td>$(.0327)$</td>
</tr>
<tr>
<td>$H_\delta$</td>
<td>$.2275$</td>
<td>$-.0742$</td>
<td>$-.1034$</td>
<td>$.7126$</td>
</tr>
<tr>
<td></td>
<td>$(.0599)$</td>
<td>$(.0120)$</td>
<td>$(.0036)$</td>
<td>$(.0058)$</td>
</tr>
<tr>
<td>$H_B^2$</td>
<td>$-.0564$</td>
<td>$-.0510$</td>
<td>$.1472$</td>
<td>$-.0771$</td>
</tr>
<tr>
<td></td>
<td>$(.0826)$</td>
<td>$(.0182)$</td>
<td>$(.0225)$</td>
<td>$(.0206)$</td>
</tr>
<tr>
<td>$H_\sigma$</td>
<td>$.4123$</td>
<td>$.5247$</td>
<td>$-.2384$</td>
<td>$.0001$</td>
</tr>
<tr>
<td></td>
<td>$(.2958)$</td>
<td>$(.0624)$</td>
<td>$(.0627)$</td>
<td>$(.0695)$</td>
</tr>
</tbody>
</table>

| **Model Constrained to MLEs of $c_\beta$, $c_\phi$, and $c_\delta$** |           |           |           |           |
| $H_B$       | $.2079$              | $.6901$   | $-.0355$  | $.1217$   |
| $H_\sigma^2$ | $-.2951$            | $-.1446$  | $.8484$   | $-.0775$  |
| $H_\delta$  | $.2289$              | $-.0767$  | $-.1034$  | $.7126$   |
| $H_B^2$     | $.0565$              | $-.0506$  | $-.1475$  | $.0770$   |
| $H_\sigma$  | $-.0935$             | $.5294$   | $-.2469$  | $.0072$   |

**MLEs of $c_\beta$, $c_\phi$, and $c_\delta$ and Estimation Error Variances**

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Error Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_\beta$</td>
<td>$-.00361$</td>
<td>$(.00243)$</td>
<td>$1$</td>
</tr>
<tr>
<td>$c_\phi$</td>
<td>$-.295$</td>
<td>$(.109)$</td>
<td>$.391$ $-.189$</td>
</tr>
<tr>
<td>$c_\delta$</td>
<td>$.565$</td>
<td>$(.232)$</td>
<td>$1$ $1$</td>
</tr>
</tbody>
</table>

*See Table 11 for explanation.*
# Table 15

Tests of the CAP Hypotheses, $a = c_\beta^2 + c_\phi h + c_\delta d$

For the Logarithmic Multiple-Factor Model with Sequential Variation

<table>
<thead>
<tr>
<th>Hypothesis: $H_1$ a is not a linear function of $b$, $h$, $d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_2$ $c_\beta$, $c_\phi$, $c_\delta$ unconstrained</td>
</tr>
<tr>
<td>$H_3$ $c_\beta$, $c_\delta$ unconstrained; $c_\phi = -0.5$</td>
</tr>
<tr>
<td>$H_4$ $c_\beta$, $c_\delta$ unconstrained; $c_\phi = 0$</td>
</tr>
<tr>
<td>$H_5$ $c_\beta$, $c_\phi$ unconstrained; $c_\delta = 0$</td>
</tr>
<tr>
<td>$H_6$ $c_\beta$ unconstrained; $c_\phi = c_\delta = 0$ i.e., $a = 0$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alternative Hypothesis</th>
<th>Maintained Hypothesis$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>$2\Delta l$ = 4.994</td>
</tr>
<tr>
<td></td>
<td>$\alpha = [10.6]$</td>
</tr>
<tr>
<td></td>
<td>df = 2</td>
</tr>
<tr>
<td>2</td>
<td>$2\Delta l$ = 3.570</td>
</tr>
<tr>
<td></td>
<td>$\alpha = [7.9]$</td>
</tr>
<tr>
<td></td>
<td>df = 1</td>
</tr>
<tr>
<td>3</td>
<td>$2\Delta l$ =</td>
</tr>
<tr>
<td></td>
<td>$\alpha =$ NA</td>
</tr>
<tr>
<td></td>
<td>df =</td>
</tr>
<tr>
<td>4</td>
<td>$2\Delta l$ =</td>
</tr>
<tr>
<td></td>
<td>$\alpha =$ NA</td>
</tr>
<tr>
<td></td>
<td>df =</td>
</tr>
<tr>
<td>5</td>
<td>$2\Delta l$ =</td>
</tr>
<tr>
<td></td>
<td>$\alpha = [7.9]$</td>
</tr>
<tr>
<td></td>
<td>df =</td>
</tr>
<tr>
<td>6</td>
<td>$2\Delta l$ =</td>
</tr>
<tr>
<td></td>
<td>$\alpha = [7.9]$</td>
</tr>
<tr>
<td></td>
<td>df =</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimated Coefficients Under Maintained Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_\beta$</td>
</tr>
<tr>
<td>$c_\phi$</td>
</tr>
<tr>
<td>$c_\delta$</td>
</tr>
</tbody>
</table>

$^a$See Table 5 for explanation. All hypotheses are implemented by maximization of (43), subject to the appropriate constraints.
V. PREVIOUS TESTS OF CAPITAL ASSET PRICING HYPOTHESIS

Miller and Scholes (1972) provide a thoughtful, critical review of the methodology in the earliest tests of CAP hypotheses. This discussion will consider some major studies published since that review, with the goals of reconciling other empirical results with those presented above and of placing other testing procedures within the methodological context that has been developed. Section V.1 takes up the doubtful merits of grouping stocks into portfolios as a preliminary of hypothesis testing. Section V.2 examines the applications of the two-step, "time-series" test in studies by Black, Jensen, and Scholes (1972) and by Black and Scholes (1974). Section V.3 considers the direct analysis of estimated factors by Fama and MacBeth (1973).

V.1 The Use of Groups of Stocks, or "Portfolios"

Virtually every previous study has employed the computation of returns on groups of stocks, or "portfolios," as an intermediate step in statistical analysis. First, the securities are grouped according to some criterion. Then, average values of the descriptors for each group are computed and used as regressors, and the average value of security returns are computed and used as the dependent variables. Since the average return for a group of securities is equal to the return on the portfolio with equal investment weights in those securities, this approach may be interpreted as the analysis of the behavior of "portfolios." This is a valuable property from a heuristic point of view, since the properties of portfolios selected on the basis of predetermined descriptors are of
interest to investors, but it is not relevant to the statistical efficiency of the method. The "portfolios" have identity only insofar as they are weighted combinations of individual securities chosen by certain criteria.

Since the present study has employed the returns on individual stocks as the unit of analysis, the first order of business must be to contrast these two approaches. To do this, it is necessary to postulate some model for the joint distribution of individual security returns, and then to compare the method of analysis that operates at the level of the individual security return with one that operates at the level of grouped returns. It will be seen that a grouping procedure can be represented as a special case of the analysis of individual security returns, with descriptors that take a particular form.

The first grouping procedure that appeared in the CAP literature was "contemporaneous grouping," in which the observations for period $t$ are grouped by percentiles of the descriptors to be used for period $t$. Thus, the grouping criterion is contemporaneous with the descriptors to be used in the regression. Once the securities are grouped, the regression may proceed in two ways. The average values of the descriptors in each group may be used as regressors, with the estimated slope coefficient retaining the same interpretation as in the regression on the raw descriptors. Alternatively, dummy variables for each group may be defined as descriptors, in which case the fitted coefficient of each dummy is the estimated height of a step function, and the number of estimated coefficients equals the number of groups.
If the return parameter is a linear function of the descriptors, either approach to regression by grouping will always reduce efficiency relative to linear regression on the raw descriptors. The only possible improvement in efficiency occurs when the true relationship is nonlinear, in which case a linear regression may be an even worse misspecification than a grouped regression. To illustrate this point, in Figure 4, a nonlinear relationship between the descriptor $H\bar{\beta}$ and $\beta$ is posited. This is the form of the relationship that would hold if $H\bar{\beta}^2$ were correlated with $(\beta-1)^2$, since then the measurement error variance in $H\bar{\beta}$ would also be correlated with $(H\bar{\beta}-1)^2$. A linear prediction rule for $\beta$, would then yield a poor fit, as indicated in Figure 4. A grouped regression, indicated by the dashed step function, would yield a better fit by virtue of its ability to flatten out the extremes. The predicted values for $\beta$, at the points of decile mean values for $H\bar{\beta}$, are constrained to lie on a straight line in this approach. Finally, an unconstrained step function, using dummy variables for the ten deciles, achieves a still better fit, although at the cost of using up ten degrees of freedom instead of two.

The situation drawn in the figure happens to be favorable to the use of groups. However, regression methods based on grouping always sacrifice statistical efficiency relative to a regression using the correct functional form. As will be seen in Table 16, below, prediction rules for $\beta$ obtained from various grouping procedures provide poorer explanation and hence reduced statistical efficiency, relative to a truncated linear prediction rule. The traditional reason for grouping was to achieve computational economy without a great loss of statistical efficiency. In this age of excellent computer hardware, this can no longer be considered a strong justification.
The second grouping technique that has pervaded the literature is "prior grouping," in which the securities are grouped by partiles of the values of descriptors taken from a "prior interval" that terminated before the "historical interval" in which descriptors for period $t$ are estimated. This has been correctly viewed as an instrumental variable method, but it has not been widely understood that the grouping aspect of the method is incidental, and that a potentially more efficient instrumental variable estimator can be constructed without grouping.

**FIGURE 4**

- The curved line gives $E(\beta|\bar{R})$.
- The straight line is the estimated linear regression.
- The dashed line gives the prediction for $\beta$ yielded by a regression onto decile average values for $\bar{R}$. The diamonds give the predictions for $\beta$ at the points of means of $\bar{R}$ that must lie on the straight line with slope equal to the estimated regression coefficient.
- The heavy horizontal lines give the fitted prediction rule for an unconstrained step function, where the average prediction error in each decile is zero.
To demonstrate this fact, suppose, for example, that $\hat{\beta}_{nt}$ and $\hat{\sigma}_{nt}^2$ are descriptors for stock $n$, computed by the formulas for $\hat{\beta}_{nt}$ and $\hat{\sigma}_{nt}^2$, but computed over the five-year interval from six to ten years before year $t$. Thus, $\hat{\beta}_{nt} = \hat{\beta}_{n,t-5}$ and $\hat{\sigma}_{nt}^2 = \hat{\sigma}_{n,t-5}^2$. Because of the serial independence of security returns, it is reasonable to assume that estimation errors in two nonoverlapping intervals are independent. Therefore, $\hat{\beta}$ and $\hat{\sigma}^2$ may serve as instruments for the regression of $\hat{r}_{nt}$ on $\hat{\beta}$ and $\hat{\sigma}^2$ if it is desired to estimate the dependence of $\hat{r}_{nt}$ on the true five-year historical averages of $\beta$ and $\sigma$.

(In terms of the typology of expectation formation in Section I.4, regression on $\hat{\beta}$ and $\hat{\sigma}^2$ would be appropriate under "Type 1" naive expectations, while the instrumental-variable regression would be appropriate under "Type 2" naive expectations.) The instrumental-variable approach is well known (see, e.g., Johnston (1972), pp. 278-281). For the heteroscedastic case, the instrumental-variable estimator is:

\[
\hat{\beta}_t = (X\hat{X}^{-1}X)^{-1}X\hat{X}^{-1}X_t^{-1}X_t, 
\]

where $X_{t-5}$ is the matrix of instruments. An equivalent formulation that is somewhat less widely known, is:

\[
\hat{\beta}_t = (X\hat{X}^{-1}X_t)^{-1}X\hat{X}^{-1}X_tX_t^{-1}X_t, 
\]

where the superscript "i" indicates a fitted value taken from a regression on the instruments:
\[
X^i_t = X^i_{t-5}(X^i_{t-5}X^{-1}_{t-5})^{-1}X^i_{t-5}X^{-1}_{t-5}X^i_t
\]

(92)

\[
X^i_t = X^i_{t-5}(X^i_{t-5}X^{-1}_{t-5})^{-1}X^i_{t-5}X^{-1}_{t-5}X^i_t.
\]

Thus, the instrumental-variable regression is equivalent to regression of the fitted values for \( X^i_t \) onto the fitted values for \( X^i_t \). It may further be shown that the statistical efficiency of the instrumental regression is proportional to the percentage of variance of \( X^i_t \) that is explained by the prior instruments.

Let us now replace the prior instruments by a set of groupings of the stocks by particles of these instruments, resulting in \( K \) groups in all. Let \( ID_1, \ldots, ID_K \) be dummy variables for these \( K \) groups, each equal to 1 if the stock is in that group and equal to 0 otherwise. Then we may carry out an instrumental-variable regression, using the fitted values from regressions of form (92), but with the \( N(t) \times K \) matrix of instrumental dummies--\( D_{t-5} \), say--replacing the matrix \( X^i_{t-5} \) of instrumental descriptors. The fitted values will now take the form:

\[
X^i_t = D_{t-5}X^a_t
\]

(93)

\[
X^i_t = D_{t-5}X^a_t.
\]

where \( X^a_t \) is a \( K \times J \) matrix of fitted group average values for the descriptors, and where \( X^a_t \) is a \( K \times 1 \) vector of fitted group average returns (or portfolio returns). Will this alternative instrumental procedure
be more efficient statistically? Only if the grouping procedure predicts a higher percentage of the variance of the historical descriptors than the regression on the prior descriptors. Again, this will be possible only if the relationship between the historical descriptors and the prior descriptors is so conveniently nonlinear that a fitted step function explains a larger percentage of variance than a fitted regression line.

When the formulas for the fitted values (93) are substituted into (91), the instrumental estimator becomes:

\[ \hat{\beta}_t = (X_t' \bar{X}_t) \bar{X}_t \bar{X}_t' \bar{X}_t = \mathbf{D}_t \mathbf{V}^{-1} \mathbf{D}_t - 5. \]

Thus, the estimation equation based on instrumental dummies takes the form of a regression of group average values of return upon group average values of the descriptors. However, the regression is only optimal when the weighting matrix \( \mathbf{Z} \) is used, and this matrix has no necessary interpretation in terms of the portfolio returns. In fact, the typical element \( \mathbf{E}_{ij} = \sum_{m \in \text{group } i} \sum_{n \in \text{group } j} \mathbf{V}_{mn} \) can only be computed by analysis of the covariances among individual security returns. Only under a strong simplifying assumption, such as the assumption that the covariance between pairs of stocks is constant across all pairs drawn from any two groups, so that the covariance depends only on the groups in which the securities lie, will the matrix \( \mathbf{Z} \) be interpreted as the inverse of the covariance matrix of portfolio residual returns. When such assumptions are satisfied, we have finally reached an estimation method that is defined in terms of portfolio returns:
\[ \hat{\mathbf{r}}_t = \left( \mathbf{x}_t' \mathbf{y}_t^{-1} \mathbf{x}_t \right)^{-1} \mathbf{x}_t' \mathbf{y}_t^{-1} \mathbf{r}_t, \]

where \( \mathbf{y}_t \) is the variance matrix for the residual portfolio returns.

The probable sacrifices in efficiency in a portfolio estimator based on instrumental groupings, relative to an instrumental-variable analysis of individual security returns, have now been traced:

1. There is a loss of efficiency unless the prior groupings predict as much variance in the historical descriptors as do the prior descriptors themselves.

2. There is a further loss of efficiency unless the correct weighting matrix \( \mathbf{z} \) is used.

V.2 Applications of "Time Series Tests"

In "The Capital Asset Pricing Model: Some Empirical Tests," Black, Jensen, and Scholes (BJS) are concerned exclusively with CAP hypotheses referring to \( \alpha \) and \( \beta \). The only descriptors are dummy variables for the deciles of historical beta, so that grouping is used to estimate prediction rules in the form of unconstrained step functions. For each month, BJS compute average returns for each of the ten portfolios. This is equivalent to the cross-sectional factor regression (47), with the factors being the returns to the portfolios. We note that heteroscedasticity is ignored, thus violating requirement 1 of Section III.3. Then, the second time-series step (48) is carried out with the ten regressions for the different factors being estimated separately (thus violating requirement 2) and with all periods equally weighted (thus violating requirement 3). The
output of this step is a pair, \( \hat{a}_j \) and \( \hat{b}_j \), for each factor \( j \) (or, in more transparent terms, an \( \hat{\alpha} \) and a \( \hat{\beta} \) for each decile portfolio).

Since there is no constant descriptor, the CAP:Z condition takes the form that \( a_j = \overset{\sim}{k}_Z (1-b_j), j=1,\ldots,10. \) (The fact that the CAP:Z conditions take this form can also be understood in terms of the discussion on "speculation on the factors" in Section III.4, for in the absence of a constant descriptor, the unbiased "factor portfolios" each have investment weights summing to one.) A test of this hypothesized constraint, as we have already discussed, involves minimization of:

\[
\left( \begin{array}{c} a - \hat{a} \\ b - \hat{b} \\
\end{array} \right) \left( \begin{array}{cc} M_a & M_{ab} \\ M_{ab} & M_b \\
\end{array} \right) \left( \begin{array}{c} a - \hat{a} \\ b - \hat{b} \\
\end{array} \right) \text{ subject to (6a).}
\]

BJS first make the simplistic assumption that \( M_a \approx \mathbb{I} \), and that \( M_{ab} = \tilde{M}_b = 0 \), in which case the optimal value of \( \overset{\sim}{k}_Z \) is obtained by regression of \( \hat{a}_j \) on \( (1-\hat{b}_j) \), or equivalently from the intercept of the regression of the time-series average portfolio return \( \bar{r}_{Pj} \) on \( \hat{b}_j \). BJS obtain this estimator \( \hat{k}_Z = .00338 \). They approximate its sampling distribution by use of the estimator:

\[
\hat{M}_a = \frac{\sum_{j=1}^{10} (\hat{a}_j - \hat{k}_Z (1-\hat{b}_j))^2}{9} \mathbb{I}.
\]

This indeed yields an unbiased estimator of \( \text{VAR}(\hat{k}_Z) \), provided that \( \tilde{M}_a \approx \mathbb{I}, \tilde{M}_{ab} = \tilde{M}_b = 0. \) However, because of the cross correlation in the portfolio residual returns, some of which is shown by BJS to arise from a random component multiplying beta (in other words, a random serially
independent factor with loading equal to beta), the requirement $M_a \propto I$ is strongly violated. The approach to this problem, as developed above, would be to estimate $\hat{M}_a$, $\hat{M}_{ab}$, and $\hat{M}_b$ directly and to deduce therefrom the confidence region for $\hat{k}_Z$. BJS employ an alternative approach.

They first present an elaborate derivation resulting in an apparently different estimator for $\bar{k}_Z$. Actually, the difference is very slight: the first estimator was obtained by regression of $\bar{r}_{Pj}$ onto $\hat{b}_j$; the second estimator may be shown to be equivalent to the same least squares regression, constrained to pass through the point $(x = 1, y = \bar{r}_M)$.¹

¹On pages 93–95 of their study, BJS carry out the regression, $\bar{r}_{Pj} = \gamma_0 + \gamma_1 \hat{b}_j$, where $\bar{r}_{Pj}$ is the average return to the jth factor (decile) over the history. From the familiar formulas for least squares regression:

$$\gamma_0 = \bar{r}_P - \frac{\sum_j (\hat{b}_j - \bar{b})(\bar{r}_{Pj} - \bar{r}_P)}{\sum_j (\hat{b}_j - \bar{b})^2},$$

(98)

where $\bar{r}_P = \frac{1}{10} \sum_j \bar{r}_{Pj}$ is the average return across all stocks in the sample and, therefore, similar to the market average return, and where $\bar{b}$ is the average estimated beta across all deciles and, therefore, approximately one.

Then, on pages 107–109, an alternative "(approximately) efficient" estimator for the above intercept is derived to be:

$$\bar{R}_Z^* = \frac{1}{T} \sum_t \left( \frac{\sum_j (1 - \hat{b}_j)(r_{Pjt} - \hat{b}_j r_{Mt})}{\sum_j (1 - \hat{b}_j)^2} \right).$$

Inverting the order of summation, and simplifying, we find that:

$$\bar{R}_Z^* = \bar{r}_M - \frac{\sum_j (\hat{b}_j - 1)(\bar{r}_{Pj} - \bar{r}_M)}{\sum_j (\hat{b}_j - 1)^2}.$$

(99)

Comparing (98) with (99), it is seen that the former is the intercept of
The first regression, being unconstrained, passes through the point of means of the variables \( x = \bar{b} \equiv \text{average portfolio beta}, \ y = \bar{r}_p \equiv \text{average portfolio return} \). Thus, the two estimators are identical if the universe of stocks in the portfolios is identical to the universe of stocks in the market, since, then, \( \bar{r}_p = \bar{r}_M, \ \bar{b} = 1 \). Otherwise there will be a difference. However, with this insight into the relationships between the estimators, it is doubtful that the second, constrained estimator is superior. For tests of the CAP hypotheses, the constraint should be tested rather than imposed, for it is one of the implications of the hypotheses. In the BJS case, fortunately, the difference between the constrained and unconstrained estimates is so small that this test is apparently satisfied.

BJS compute the constrained estimator separately for each monthly cross section in the sample. The global average of these estimators, \( \sum_{t} \hat{k}_{zt} / t \), which is taken as the estimator of \( \hat{k}_Z \), possesses the properties alluded to in the previous paragraph. Thus, the computation of a time series of estimators achieves nothing in terms of increased efficiency for regression that is unconstrained and therefore passes through the point \( (\bar{r}_p, \bar{b}) \), and the latter is the intercept of a least-squares fitted line that is constrained to pass through the point \( (\bar{r}_M, 1) \).

Therefore, if \( \bar{r}_p = \bar{r}_M \) and \( \bar{b} = 1 \), the two procedures are equivalent. Otherwise, the difference between the two can be computed. From the data given in BJS, Table 2, we have the approximate values \( \bar{r}_p - \bar{r}_M = .0002, \ \bar{b} - 1 = .007 \). Substitution of these values yields, to the fourth decimal place, \( \gamma_0 - \bar{r}_Z = \bar{r}_p - \bar{r}_M = .0002 \). In fact, the reported values, \( \gamma_0 = .00359 \) and \( \bar{r}_Z = .00338 \) differ by roughly this amount.

the global estimator. However, computation of the time series of estimators is an extremely valuable innovation for another reason: it provides an unbiased estimator of the sampling variance of the estimated average, which is \( \text{VAR}(\hat{k}_Z) \). Formally, if for each \( t \), \( \text{E}(\hat{k}_{Zt}) = \bar{k}_Z \), \( \text{VAR}(\hat{k}_{Zt}) = \sigma_t^2 \), and \( \text{COV}(\hat{k}_{Zt}, \hat{k}_{Zs}) = 0 \) for \( s \neq t \), then \( \text{E}\left( \frac{1}{T} \sum_t \hat{k}_{Zt} \right) = \bar{k}_Z \), \( \text{VAR}\left( \frac{1}{T} \sum_t \hat{k}_{Zt} \right) = \frac{1}{T^2} \sum_t \sigma_t^2 \) and \( \frac{1}{(T-1)T} \sum_t (\hat{k}_{Zt} - \bar{k}_Z)^2 \) is an unbiased estimator of \( \text{VAR}(\hat{k}_Z) \).

This approach to estimating the variance is robust against misspecification of the cross-sectional distribution of security returns and also against any misspecification in the joint distribution over time except serial dependence in \( \hat{k}_{Zt} \). Misspecification may lead to inefficient estimators \( \hat{k}_{Zt} \), and to inefficient weighted combination of those estimators in the global estimator, but, in the absence of serial dependence, this procedure unbiasedly estimates the variance of whatever global estimator is used. The importance of this device is highlighted by the fact that BJS find the unbiased estimator for the variance to be more than ten times as great as the value computed in the earlier regression. BJS treat the distribution of \( \hat{k}_{Zt} \) in much the same fashion as the vector \( \hat{\pi}_t \) is examined in the present study, where the sampling distribution of any function of the estimated factors (such as the BJS \( \hat{k}_Z \)) is deduced from the distribution for the factors themselves.

This test of misspecification of the cross-sectional distribution is closely analogous to the test of misspecification applied in Section IV.3.
One indication of the improvements in specification in the multiple-factor model of Section IV.2, relative to that of BJS, is that the upward adjustment in variance is found to be 1.52, as opposed to the BJS value greater than 10.

BJS do not consider the use of logarithmic return relatives, nor do they consider the possibility of compensation to yield or specific risk. Both of these have been found to have important influences on the results in the present study. BJS do not carry out the final step of testing the CAP:2 against the alternative model, where $a$ and $b$ are unconstrained. Their paper concludes with an appendix concerning the use of grouping to deal with the problem of measurement error in $\hat{b}$. Grouping does indeed reduce the magnitude of the variance matrix $\tilde{\Sigma}_b$. The procedure developed in the current study in no way requires that $\tilde{\Sigma}_b = 0$, so that this function of grouping is irrelevant. What the power of the present approach does depend upon is the percentage of variance in $\hat{\beta}$ that is predicted by the descriptors, and grouping should be employed only when it leads to an increase in this percentage.

A recently published study by Black and Scholes (BS), "The Effects of Dividend Yield and Dividend Policy on Common Stock Prices and Returns" (1974) examines CAP hypotheses concerning $\alpha$, $\beta$, and $\delta$. The approach is again a two-step procedure, with several innovative features.

BS construct descriptors for $\beta$ and $\delta$, $b^f_\beta$ and $b^f_\delta$, that correspond to predicted values from a rational prediction rule fitted over the immediately preceding years. The details, in the terminology of the
present study, are as follows: For each year \( s \), securities are partitioned lexicographically, first by quintiles of \( \hat{H}S_s \), then within these groups by quintiles of \( \hat{H}S_s \). For each year \( s \), this procedure results in twenty-five predetermined groups, \( i = 1, \ldots, 25 \), for which dummy variables \( D_is \) may be defined. These may be used to implement prediction rules for \( \beta \) and for \( \delta \) in the form of fitted step functions, with the predictions for any year \( t \) being given by:

\[
\hat{\beta}_{nt} = \sum_{i=1}^{25} \hat{\beta}_{it} D_{nit} \\
(100) \quad n = 1, \ldots, N(t) \\
\hat{\delta}_{nt} = \sum_{i=1}^{25} \hat{\delta}_{it} D_{nit}
\]

In the BJS study, \( \hat{\beta}_{it} \) for each \( i \) was identical for all \( t \). The values were fitted over the entire sample history as is appropriate for expectations of Type 4, corresponding to a rational prediction rule fitted over the entire sample. Alternatively, the prediction rule can be fitted only over previous history. BJS chose this approach. Using five years of data previous to year \( t \) (years \( s = t-1, t-2, t-3, t-4, t-5 \)), the average values of \( \beta \) and \( \delta \) for those stocks that were classified into group \( i \) by the predetermined groupings for those years are computed. These averages, \( \hat{H}S_{it}^f \) and \( \hat{H}S_{it}^f \), are equal to the estimated coefficients in the rational prediction rules (100) fitted over those five years. They specify the fitted step functions which would have achieved minimum mean square error in predicting \( \beta \) and \( \delta \) in that period.
For year $t$, the descriptors for $\beta$ and $\delta$ are the predictions obtained from these rules:

$$
\hat{H}_t^\beta D_{it \text{ nit}} \quad \text{and} \quad \hat{H}_t^\delta D_{it \text{ nit}} \quad i = 1, \ldots, 25.
$$

If these descriptors were asserted to be equal to investors' expectations, the approach would be appropriate for Type 3 expectations, corresponding to a rational prediction rule fitted to prior data. However, BS employ these descriptors as the inputs for a complete two-step analysis in which the predictions are further adjusted to an optimal fit over the entire sample history. The fitted prediction rule is of the form:

$$
\hat{\beta}_{nt} = b_0 + b_1 \left( \sum_{i=1}^{25} \hat{H}_t^\beta D_{it \text{ nit}} \right) + b_2 \left( \sum_{i=1}^{25} \hat{H}_t^\delta D_{it \text{ nit}} \right) \quad (101)
$$

$$
\hat{\delta}_{nt} = d_0 + d_1 \left( \sum_{i=1}^{25} \hat{H}_t^\delta D_{it \text{ nit}} \right) + d_2 \left( \sum_{i=1}^{25} \hat{H}_t^\delta D_{it \text{ nit}} \right)
$$

In fitting this rule, BS, like BJS and the present authors, subscribe to expectations of Type 4. The device of a moving prediction rule fitted to the recent past is presumably used to track possible variations in the coefficients of the prediction rules.

This is clearly an interesting possibility. Whether or not the properties of the fitted prediction rule are improved is an empirical matter. On the one hand, if the parameters of the optimal prediction rule are drifting over time, some of that drift can, in principle, be captured in the BS device. On the other hand, the use of fitted prediction
rules over short periods as descriptors introduces measurement error in
the descriptors that can be expected to worsen predictive performance.
In our replication of the BS methodology, we found that the \( \beta \) prediction
rule of form (101) achieved an \( R^2 \) of .3499, whereas the prediction rule
using \( HB \) and \( H_0 \) achieved an \( R^2 \) of .3572. Thus, the BS device sub-
stantially worsens the explanatory power of the prediction rule. Some
other device, such as improved definitions of the descriptors, should be
employed to achieve stationarity.

Following the remarks in Section V.1, it is apparent that the use
of grouping to fit the prediction rules is incidental. Any other func-
tional form could have been used. The lexicographic grouping does achieve
substantial between-group variation in the descriptors, but some predic-
tive potential is sacrificed by discarding within-group variation in \( HB \)
and \( H_0 \).

To emphasize this point, we applied several grouping procedures to
the data sample utilized by BS. Using the lexicographic ordering proposed
by BS, three types of descriptors were constructed: contemporaneously
grouped descriptors, where the descriptor is equal to the average value
in the period where grouping occurred; prior grouped descriptors, where
the descriptor is the average value for the past five years, for groups
constructed using data lagged six to ten years; and the BS descriptors
just described. For the first two approaches, inclusion in the sample
required ten previous years of data for the firm; this resulted in a
sample of 240,527. For the BS approach, inclusion in the sample followed
the BS criterion, resulting in 280,799 data points. As seen in Table 16, use of the raw descriptors H8 and H0 resulted in higher explanatory power than the grouped descriptors; in fact, the loss of explained variance from the prior grouping or from the BS procedure is about 20 percent. Another disadvantage of either of these two instrumental grouping procedures is the necessary sacrifice of five years of data, since the sample must begin in 1936 rather than 1931. Finally, it is noteworthy that the instrumental groupings substantially fail in their goal of eliminating prediction error; in all cases, the fitted prediction rule includes a highly significant negative weight on historical yield.

A further sacrifice in statistical efficiency is entailed by simplifications where the securities are arbitrarily assigned equal weights, and where the structure of residual variances is implicitly assumed to be of the form where the optimal cross-sectional regression treats each group or "portfolio" as a single stock, as in (99).

Within this simplified context, BS consider the problem of efficiently conducting a series of cross-sectional regressions to estimate \( f_t, t = 1, \ldots, T \), followed by the second step, time-series regressions of form (47a). In the present study, the optimal cross-sectional regression has been shown to be of form (47), or (99) if portfolios are to be treated as stocks. This was done in two ways: first, in III.3, by demonstrating that this form of cross-sectional regression, when combined with efficient conduct of the time-series regression, is equivalent to the efficient single-step "market-conditional" regression developed in III.2;
TABLE 16

COMPARISON OF ALTERNATIVE GROUPING PROCEDURES IN PREDICTION RULES
FOR $\beta$, FITTED BY MARKET-CONDITIONAL, RATE-OF-RETURN
REGRESSIONS FOR 1936-1966, OF THE FORM

$$r_{nt} = \hat{c}_0 + \sum \hat{c}_j (x_{jnt} r_{Mt})$$

WHERE $x_1 = 1, x_2$ IS A DESCRIPTOR OF BETA AND $x_3$ IS A DESCRIPTOR OF YIELD

<table>
<thead>
<tr>
<th>Sample A: $N = 240,527$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CONST</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample B: $N = 280,799$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CONST</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
second, in III.4, by showing that this cross-sectional regression defines the minimum variance speculations on the factors of return that are unbiased with respect to other factors of return. BS pioneered the latter approach in deriving their cross-sectional regression, and, indeed, our development in Equations (51) through (53) is little more than a systematic restatement of their innovative derivation. All that has been added is the statement that the Aitken's Generalized Least Squares, as expressed in (99) for portfolios, is the solution to the problem of "minimum variance unbiased speculation" on the factors, as defined in BS (1974, conditions (8) through (12)). BS appear to have been unaware that an explicit solution to their problem existed.\(^1\)

\(^1\)In Section III of their paper, BS use an approach motivated as follows:

(a) We construct a portfolio whose expected return is the quantity we want to estimate.
(b) To avoid bias, we select stocks for the portfolio at each point in time using only information that was available at that time.
(c) To get an efficient estimator, we select a portfolio with the smallest possible variance of return, subject to conditions (a) and (b).

Now, a "portfolio" return is just a weighted sum of individual security returns. Upon substitution of the term "weighted average of security returns" for the word "portfolio," the language of BS makes it clear that they intend to derive an unbiased estimator that is (a) linear in security returns (a weighted sum of security returns), (b) based on predetermined descriptors, and (c) minimum variance. Since the Aitken's Generalized Least Squares Estimator is the minimum variance linear unbiased estimator, one suspects that BS are attempting to rederive this.

To illustrate that Aitken's GLS does provide the solution to the requirements of BS, it is sufficient to restate those conditions slightly. First, to convert the problem to our formalism, note that the three characterizations of each group provided by BS are the descriptors \(x_{lit} \neq 1\),
It is a praiseworthy step to introduce GLS into the cross-sectional regressions, so that the appropriate response (99) to the variance matrix of residual returns $\gamma^a_t$ is made. However, it remains to obtain an appropriate estimator for this variance matrix. BS chose an unfortunate value for this matrix, with the result that the heteroscedasticity correction is vacuous: the BS estimators $\hat{\gamma}^a_t$ are identical to those that would have been obtained by ordinary least squares.\footnote{The authors choose the form:}

$$v^a_t = \sigma_{ee} I + \sigma_{mm} \begin{pmatrix} H^f_{lt} \\ H^f_{25,t} \end{pmatrix} \begin{pmatrix} H^f_{lt} \\ H^f_{25,t} \end{pmatrix}'$$
remains in violation of Requirement 1 of Section III.3 for efficiency in the two-step procedure.

Having discussed the BS approach to the cross-section estimation problem, we turn next to their execution of the second (time-series) step. Here BS ignore the "seemingly unrelated equations" aspect entirely, thus portfolio residual variance is assumed to be the same for all portfolios, equal to $\sigma_{\varepsilon \varepsilon}$. In addition, the variance of the market, $\sigma_{mm}$, produces a covariance of rank 1 among all portfolios. This is equivalent to a random market factor multiplying the descriptor $H_{0t}^f$ for each portfolio. Recall that, in deriving the estimators for $\hat{f}_{0t}$ and $\hat{f}_{\beta t}$, BS have required that $\varepsilon_{0t}$ and $\varepsilon_{\beta t}$ are orthogonal to $H_{0t}$, in order to avoid bias from the contribution of $H_{0t}$. As a result, one would expect that the variance of the coefficient of $H_{0t}$, whatever it was, would not influence these estimators in any way. In fact, one may verify, by a series of matrix operations applied to the Aitken's Generalized Least Squares Estimator with $V^a$ of the above form, that the estimators for $f_0$, $f_{\beta}$, and $f_{\beta}$ are unaffected by the term $\sigma_{mm} H_{0t} H_{0t}'$ and will be the same if this term is omitted entirely. (This result is an instance of a general theorem, Rao (1965), that the optimal estimator for a single parameter vector [$f_0$ in a single time period], with one or more parameters in that vector being stochastic [here, $f_{\beta t}$ is presumed to include the stochastic $\varepsilon_{Mt}$], is identical to the ordinary least squares estimator that ignores stochastic variation. It is when a number of cases having different stochastic parameter vectors are combined that the stochastic variation becomes important.) Since the second term of $V^a$ makes no contribution, and the first term is diagonal, it turns out that there is no heteroscedasticity correction at all, contrary to the authors' intention. The reason given by BS for choosing this simplification, instead of the full variance-covariance matrix, is: "We used this method of estimating the covariance matrix because the estimate [sic] errors of the elements of the estimated covariance matrix itself turned out to be extremely high."
carrying out the regressions for \( \hat{\beta}_1 \) and \( \hat{\beta}_5 \) as if they were unrelated, thus providing yet another justification for the name attached to this problem and violating Requirement 2 of III.3. BS arbitrarily weight each time period equally, thus violating Requirement 3. Both of these omissions sacrifice statistical efficiency and, by ignoring heteroscedasticity, generate biased estimates of the standard errors of the estimated coefficients and, therefore, biased hypothesis tests.

The time-series analysis of \( \hat{\beta}_B \) is not carried out, and potentially valuable information concerning the CAPM model is thereby lost. Also, the possibility of compensation for specific risk is assumed away, and logarithmic returns are not analyzed, two factors that have been found to influence profoundly the estimated compensation to yield.

In view of the numerous differences between the BS methodology and that of the present authors, it is difficult to attribute the differences in results to any one source, despite the fact that the data base is virtually identical. We attempted to replicate the BS results, intending to illustrate the progressive improvements in precision from more careful treatment of residual variance. However, we were unable to reproduce the portfolio weights reported in BS, Table 2 (p. 15).¹ An

¹For the five beta classes within each yield quintile, the portfolio weights for estimation of reward for yield, \( \omega_{it} \), generally increase as beta increases. A reversal can be said to occur if, within any yield class, for two beta classes \( i \) and \( j \), where \( \beta_{it} > \beta_{jt} \), \( \omega_{it} < \omega_{jt} \). BS report seven sets of portfolio weights for seven years, to two decimal places. For the 175 reported weights, there is only one reversal. By contrast, according to our computation of the weights, in each of the
important parameter, from the point of view of investment policy, is the increase in total expected return for a one-unit increase in cash yield. Based upon their entire sample, from 1936 to 1966, the BS estimate is approximately .23, with a reported standard error of .24. Our estimate, from Table 14, is .565, with a standard error of .232. The similarity in the standard errors is accidental; in the context of BS, where the compensation to total risk is set to zero, the standard error would be only .2132.

V.3 The Pioneering Application of the Multiple-Factor Model

In "Risk, Return, and Equilibrium: Empirical Tests," Fama and MacBeth (1973) are concerned with the compensations in expected return for systematic risk, specific risk, and a possible nonlinearity term, β². A prior grouping procedure is used to construct descriptors for these variables; since the grouping is done by n-tiles of beta alone, orthogonal variance in the descriptors between the groups is relatively small, and the explanatory power of the resulting prediction rules is minimized. Furthermore, descriptors are only revised every four calendar years (rather than each year, as in all other studies), reducing the explanatory

years (and, indeed, in each of the thirty-one years), there is at least one reversal that would show up if the weights were reported to two decimal places. In all, there are sixteen reversals in our computed weights for the seven years in which BS report one reversal. We have no way of accounting for the greater regularity of the BS results, and we find that regularity to be quite implausible in view of the definition of the estimator. Accordingly, we leave this discrepancy as an unexplained mystery.
power of the descriptors by increasing the average time lag. These two aspects reduce statistical efficiency. More important, Fama and MacBeth (FM) identify these descriptors or historical instrumental estimates of the return parameters with the expectations, as is appropriate for naive expectations of Type 2. Inefficiency in these instrumental estimates will therefore reduce the estimated compensation to the return parameters, if the expectations process actually corresponds to a rational prediction rule of Type 3 or 4.

FM did introduce an important innovation in applying the approach to estimation of the multiple-factor model, described above in the discussion in Section III.4 leading to (58). Variances $F_{jj}$ of the factors associated with the four descriptors are estimated. (These are the total variances, rather than residual variances as reported in the present study.) It is important to note that these estimated variances are unbiased only if the sampling variances of the estimated factors are known. Any misspecification of heteroscedasticity in the cross-sectional regression would tend to lead to a downward bias in the estimated standard errors, and hence to an upward bias in the estimated variance of the factors. Such bias was almost certainly present, for the specific returns were assumed to be homoscedastic.
VI. CONCLUSION

This paper has attempted to contribute to the methodology of tests of Capital Asset Pricing Hypotheses and to extend our understanding of the empirical distributions of security returns. The empirical results may be summarized as follows:

1) The increase in expected monthly logarithmic return, per unit of systematic risk or beta, is estimated as .0067 (with standard error of .0024) in comparison with the mean logarithmic, equal-weighted market return, in excess of the ninety-day treasury bill rate, of .0103.

2) The increase in expected monthly logarithmic return, per unit of total variance of logarithmic return, is estimated as -.295 (with standard error of .109) in contrast to the theoretical value of -.5 that would be expected in the absence of aversion to specific risk.

3) The increase in expected monthly logarithmic return, per unit of monthly cash yield, is estimated at .565 (with standard error of .232), a value that would be consistent with indifference between cash yield and capital gains at about a 60 percent marginal tax rate.

4) Predictable heteroscedasticity in specific and total return is conclusively demonstrated, and the predicted fluctuations in variance account for most of the kurtosis in returns.

5) The appropriateness of the multiple-factor model of security returns, with loadings equal to predetermined descriptors, as opposed to a single-factor or market model, is conclusively demonstrated, and substantial factor variance is associated with the descriptors.
6) Sequential variation in the zero-beta intercept and the prediction rule for beta are found.

7) The lognormal approximation to security returns dominates the normal approximation.

These empirical results are interesting but should only be viewed as indicative of the results to be expected from further analysis. At least three major improvements in the study, which should be undertaken before the results can be considered definitive, are: respecification of the market rate of return to reflect a value-weighted market; extension of the data sample from 1966 through to the present; and improvement of the specification of the prediction rules for systematic risk, total risk, and cash yield.

A two-step regression procedure, whose asymptotic efficiency has been demonstrated by Amemiya, has been utilized to estimate the prediction rules for residual and total risk. Maximum likelihood methods have been used to estimate the variance components in a multiple-factor model of security returns, and a robust test of misspecification has been provided. Maximum likelihood methods have been used to estimate the degree of sequential random variation in the parameters of return, and to obtain smoothed time-series estimates of these varying components. The increasingly sophisticated estimates of the variance components in security returns improve efficiency in Generalized Least Squares estimates of the prediction rules for expected return and systematic risk.

These efficient GLS estimators may be viewed as a standard of comparison for statistical methodology in tests on these return parameters.
It is therefore helpful to show that three apparently different testing procedures are actually equivalent to GLS: the "market-conditional" regression, the "two-step, cross-section, time-series" regression, and the direct test on the moments of the estimated factors. We have couched our approach in terms of the former procedure, which seems to us to be the most straightforward. Previous tests in the literature have been variants of the second and third approaches, and specific inefficiencies in these variants are considered. We have also examined the doubtful virtues of the grouping procedures that have been customary in the literature.

We implement the actual tests of the CAP hypothesis by maximum likelihood over the system of estimated prediction rules for the relevant return parameters, subject to the constraints imposed by the CAP hypotheses. The approach allows easy generalization to any number of return parameters. It does not require the absence of measurement error in the predictions or in the estimated prediction rules.

The properties of the hypothesis tests, as progressive improvements in the treatment of components of variance are introduced, are illustrated in a series of tables. It is encouraging that the estimates of the rational prediction rules for the return parameters are quite robust against misspecification of the components of variance, although the apparent statistical significance of the associated hypothesis tests is increased by misspecification.

We find that the conclusions are importantly influenced by three aspects of the problem formulation: whether the process of expectations
formation is naive or rational; whether logarithmic returns or rates of return are the subject of analysis; and whether all relevant return parameters—systematic risk, total risk, and yield—are admissible for compensation, or whether compensation for some of these is fixed to zero. We prefer rational expectations, logarithmic returns, and the admissibility of as many potentially relevant parameters of return as possible.


