PORTFOLIO THEORY AND THE CAPM: THEORY AND TESTS.

JIRO E. KONDO

Abstract. The purpose of this handout is two-fold. Firstly, it serves as a guide to understanding some of the important intuitions from portfolio theory and the capital asset pricing model (CAPM). This purpose is best served by gaining comfort with the assumptions, remarks, and figures in this handout. The second goal here is to provide interested students with an accurate sketch of the theory guiding the results from this section of the course. Although proofs are not provided in full detail, the work contained in the appendices should allow students to gain a sufficient understanding of the foundational theorems of portfolio theory and the CAPM.

As a special bonus, I have included a couple summaries of important empirical papers related to the CAPM. These write-ups were prepared for personal use and are not expected to be fully absorbed by everyone. However, since I’m sure some students have the technical background necessary to follow these summaries, I’ve decided to throw them in. I encourage students to at least read over the summaries of these sections.

1. Portfolio Theory I: General Discussion.

Remark 1.1 (Portfolio Theory). Portfolio theory is an economic theory of investor behavior. It postulates a framework for selecting optimal (efficient) portfolios. This framework is closely connected to the efficiency frontier because every investor chooses a portfolio on the upward sloping part of this curve. The importance of this topic became clearest when it was used to derive the CAPM (which is a model of asset pricing based on investor behavior described by the portfolio theory). In its essence, portfolio theory narrows in on one specific aspect of normative investment... short term diversification. Its results, taken at face value, are not accurate if we consider diversification over long horizons, asymmetric information/differences in opinion, transactions costs, nontradeability of certain assets, and other imperfections in financial markets. Regardless, portfolio theory remains an important component of finance theory for three reasons:

- Reason 1: CAPM.

- Reason 2: Diversification, even in the short term, is an extremely important component of optimal investment.

- Reason 3: Historically, portfolio theory represents the first major theoretical breakthrough in the theory of finance (with the exception of Bachelier’s work in 1900 which was a slightly flawed analysis of option pricing and wasn’t really ”discovered” until the late 50s). Portfolio theory was developed by Harry Markowitz (see Markowitz(1952)) who won the Nobel prize in economics for this contribution.
Assumption 1.2 (Distribution of Returns). There are $N$ assets with joint distribution $(\mu, \Sigma)$ where $\Sigma$ is an invertible variance-covariance matrix. $\mu$ and $\Sigma$ are known to every investor. There is only 1 time period (i.e. this is a static model).

Assumption 1.3 (Investor Behavior). Investors only care about the mean and variance of their portfolio returns ($w'\mu$ and $w'\Sigma w$ respectively). Specifically, they like higher means (return) and dislike higher variance (risk). As a result, every investor chooses a portfolio that solves the following constrained optimization problem

$$V^*[\mu^*] = \min_{w \in B(\mu^*)} w' \Sigma w$$

where

$$B(\mu^*) = \{ w : w'1 = 1 \text{ and } w'\mu = \mu^* \}$$

for a some value of $\mu^*$. $B(\mu^*)$ can be thought of as a budget set along with a specified target level of expected return, $\mu^*$. Mathematically, $B(\mu^*)$ is a $(N-2)$-dimensional linear subspace of $R^N$. Economically, all this optimization says is that investors will choose the portfolio that minimizes portfolio return variance, $V^*[\mu^*]$, given their chosen level of portfolio expected return, $\mu^*$. This behavioral assumption is often referred to as the mean-variance utility of wealth assumption (although it actually assumes a little more than this).

Remark 1.4 (About the Assumptions). The invertibility of $\Sigma$ assumption is equivalent to assuming that there is no riskless or redundant security among the $N$ assets. It turns out that neither assumption is critical. In fact, we will begin to assume that there is a riskless asset a little later and this will be important in deriving the CAPM. On the other hand, the assumption that every investor knows $\mu$ and $\Sigma$ is very unrealistic (it effectively implies that all investors have the same information, i.e. there’s no private information, and that they all process information the same way, i.e. there are no disagreements). Assuming a static world is also unreasonable especially if joint returns are not iid over time (e.g. no serial dependence in the distribution of returns, no stochastic time-variation in volatility, etc). Furthermore, the assumption of mean-variance utility is tough to reconcile with the usual approach used in economic theory (namely that agent’s maximize their expected utility of wealth). One notable case where the two approaches are equivalent occurs when we additionally assume that returns are distributed joint normally and wealth is exactly equal to portfolio value. These two additional assumptions are quite restrictive.

Theorem 1.5 (Efficient Portfolios). Given assumptions (1.1) and (1.2), the efficient portfolio that achieves the desired target level of return $\mu^*$ is given by

$$w^*(\mu^*) = w_p + w_a \mu^*$$

where $w_p$ is an efficient portfolio (the one with $\mu^* = 0$) and $w_a$ is a zero-investment portfolio (with an expected return of 1). Both portfolios only depend on $\mu$ and $\Sigma$.

Remark 1.6 (More on Efficient Portfolios). The following results can all be proven using the formula for efficient portfolios

- Result 1: All efficient portfolios can be written as a weighted average of any two other efficient portfolios. This means that all efficient portfolios lie on a line in portfolio space. This line is called the portfolio frontier.
- **Result 2:** A weighted average (i.e. portfolio) of any number of efficient portfolios yields an efficient portfolio (i.e. every point on the portfolio frontier is an efficient portfolio).

Result 2 is particularly important in deriving the results leading up to the CAPM. Figure 1.1 below illustrates Result 1 in the 3 asset case. The set of all possible portfolios is illustrated by the plane, \( w'1 = 1 \), while the efficient portfolios (which are the only portfolios that investors pick in our model) are located on a line (red) on this plane. Results 1 and 2 taken together are usually referred to as *two fund separation* (more on this later).

Theorem 1.7 (Efficiency Frontier). The efficient portfolios map out an optimal relationship between risk and return in mean-standard deviation space. This relationship is given by the hyperbola

\[
\sigma^*(\mu^*)^2 - \sigma_{aa}^2 [\mu^* + \frac{\sigma_{pa}^2}{\sigma_{aa}^2}]^2 = \frac{\sigma_{aa}^2 \sigma_{pp}^2 - \sigma_{pa}^4}{\sigma_{aa}^2}
\]

where \( \sigma_{ij}^2 \) is the covariance of the returns of portfolio \( i \) with the returns of portfolio \( j \) and \( \sigma^*(\mu^*) = V^*[\mu^*]^{1/2} \) (i.e. the standard deviation of the optimal portfolio with mean return \( \mu^* \)).

Remark 1.8 (Visualizing the Efficiency Frontier). Figure 1.2 gives a fairly generic sketch of the efficiency frontier. The red dots in the figure indicate the location of each individual asset in mean-standard deviation space. It’s important to notice that each asset lies on or within the efficiency frontier. That is, points outside the efficiency frontier cannot be achieved by any portfolio using the \( N \) assets.
Remark 1.9 (Upward Slopping Part of Efficiency Frontier). In general, we’re only interested in points on the upward sloping part of the efficiency frontier. The idea is that investors will only choose portfolios that lie on this part because they dislike portfolio variance. We can rewrite the relationship between $\mu^*$ and $\sigma^*(\mu^*)$ for this upward part as

\[
\sigma^*(\mu^*) = [\kappa_1 + \kappa_2(\mu^* - \kappa_3)]^{1/2}
\]

where $\kappa_1 = \frac{\sigma_{aa}^2 - \sigma_{pa}^4}{\sigma_{aa}^2}$, $\kappa_2 = \sigma_{aa}^2$, and $\kappa_3 = -\frac{\sigma_{pa}^2}{\sigma_{aa}^2}$ are all positive constants (notice that this means that the returns of portfolios $p$ and $a$ are negatively correlated). The derivative of $\sigma^*$ at $\mu^*$ with respect to $\mu^*$ is given by

\[
\frac{\partial \sigma^*(\mu^*)}{\partial \mu^*} = \frac{\kappa_2(\mu^* - \kappa_3)}{\sigma^*(\mu^*)}.
\]

Remark 1.10 (Adding Assets). Taking $\mu$ and $\Sigma$ for the $N$ initial assets as constant, adding assets available to investors when they form portfolios can only make them better off. To see this, note that they can still form the most preferred portfolio among the $N$ assets when they choose among $N + K$ assets (i.e. $K$ assets added). Of course, since they have more freedom in forming portfolio (i.e. more choice of assets), they can be better off. This is reflected by a "shift to the left" of the frontier (to be specific, this isn’t exactly a shift because the shape of the frontier is likely to change as well - though it will still be a hyperbola). Figure 1.3 illustrates this effect where the red dots are the initial assets that generate the bold efficiency frontier while the turquoise dot represents the new asset which, together with the initial assets, generates the dotted efficiency frontier. Given these shifts, one might wonder why investors wouldn’t hold as many assets as possible in their portfolios. This might occur for reasons completely outside this model, however, even portfolio theory has something to say about this. Although, adding assets can’t hurt investors, they don’t necessarily help. As a general rule of thumb, it is useful to think along the following lines. Adding assets when $N$ is small is probably very helpful, but not so helpful when $N$ is large. Practitioners often refer to a number
between 20 and 40 as the threshold where adding assets isn’t so useful. To the extent that managing portfolios with a larger number of assets is more costly (e.g. higher transactions costs, more effort for research needed), this philosophy should be kept in mind. However, it can also be very misleading. Notably, the shifts in the efficiency frontier are mainly due to covariances among assets, therefore adding assets that correlate in the right way with the other assets as a whole (i.e. low or negative correlation) is what really helps. This statement holds regardless of the size of \( N \).

\[
\begin{align*}
\mu^* & \quad (\mu^*) \\
\sigma^*[\mu^*] & \quad \sigma^*[\mu^*]
\end{align*}
\]

**Figure 1.3:** Adding Assets.

2. **Portfolio Theory II: Two Assets.**

**Assumption 2.1** (\( N=2 \)). The joint distribution of asset returns has mean and variance given by

\[
\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix}
\]

where \( \sigma_i \) is the standard deviation of the returns of asset \( i \), \( \rho \) is the coefficient of correlation between the returns of assets 1 and 2, and \( \mu_1 \neq \mu_2 \).

**Remark 2.2** (What’s Special About Two Assets?). The difference between the case where \( N = 2 \) and \( N \geq 2 \) is that when \( N = 2 \), there is only one portfolio that achieves the target level of expected return \( \mu^* \). This is due to the fact that \( B(\mu^*) \) is just a point for every value of \( \mu^* \) (a 0-dimensional linear subspace of \( R^2 \) is just a point). This greatly simplifies the analysis because we’ve effectively taken out the portfolio choice component of the investor’s problem (i.e. all we’re left with is the tradeoff between risk and return).

**Theorem 2.3** (Optimal Portfolio and Variance). The efficient portfolio with target expected return \( \mu^* \) is given by

\[
\begin{align*}
w_1^*(\mu^*) &= \frac{\mu^* - \mu_2}{\mu_1 - \mu_2} \quad \text{and} \quad w_2^*(\mu^*) = \frac{\mu_1 - \mu^*}{\mu_1 - \mu_2}
\end{align*}
\]
which can be written as

\( w^*(\mu^*) = \left[ \begin{array}{c} -\frac{\mu_2}{\mu_1} \\ \frac{1}{\mu_1-\mu_2} \end{array} \right] + \left[ \begin{array}{c} \frac{\mu_1-\mu_2}{\mu_1-\mu_2} \\ \frac{\mu_1-\mu_2}{\mu_1-\mu_2} \end{array} \right] \mu^* \)

and the standard deviation of this portfolio’s return is given by

\( \sigma^*(\mu^*) = [w_1^2(\mu^*)\sigma_1^2 + 2\rho w_1^2(\mu^*)w_2^2(\mu^*)\sigma_1\sigma_2 + w_2^2(\mu^*)\sigma_2^2]^{1/2} \)

and the derivative of \( \sigma^* \) with respect to \( \mu^* \) is

\( \frac{\partial \sigma^*(\mu^*)}{\partial \mu^*} = \frac{C_1 \mu^* - C_2}{\sigma^*(\mu^*)} \)

where \( C_1 = \frac{\sigma_1^2 - 2\rho \sigma_1 \sigma_2 + \sigma_2^2}{(\mu_1 - \mu_2)^2} \) and \( C_2 = \frac{\mu_2 \sigma_1^2 - \rho (\mu_1 + \mu_2) \sigma_1 \sigma_2 + \mu_1 \sigma_2^2}{(\mu_1 - \mu_2)^2} \).

Remark 2.4 (Simplifying (2.5) for an Important Case). It turns out that expression (2.5) can be reduced to the much nicer formula when evaluated at the point \((\sigma_2, \mu_2)\)

\( \frac{\partial \sigma^*(\mu^*)}{\partial \mu^*} = \frac{\rho \sigma_1 \sigma_2 - \sigma_2^2}{\mu_1 - \mu_2} \)

Likewise, when evaluated at the point \((\sigma_1, \mu_1)\), we get

\( \frac{\partial \sigma^*(\mu^*)}{\partial \mu^*} = \frac{\rho \sigma_1 \sigma_2 - \sigma_1^2}{\mu_2 - \mu_1} \).

Remark 2.5 (Visualizing the Efficiency Frontier for 2 Assets). When there are only two assets as described above, both assets lie on the efficiency frontier because each is the unique portfolio that achieves \( \mu_1 \) and \( \mu_2 \) respectively. The exact shape of the frontier is easy to derive for the case where \( \rho = 1 \) and \( \rho = -1 \) (this only involves completing a square and taking a square root). In both cases, the efficiency frontier is linear. In all other cases (i.e. \(-1 < \rho < 1\)), the efficiency frontier is a hyperbola. Figure 2.1 illustrates these properties.

![Figure 2.1: Two Asset Efficiency Frontier.](image)
Remark 2.6 (One of the Assets is Riskless). When one of the assets is riskless, say asset 2, it immediately follows that $\rho$ and $\sigma_2$ are both equal to zero. In this case, the variance of the efficient portfolio with target return $\mu^*$ is given by

\[
\sigma^*(\mu^*) = \frac{\sigma_1}{\mu_1 - r_f}(\mu^* - r_f)
\]

where $r_f$ is the return on the riskless asset (which we used to refer as $\mu_2$ when asset 2 was risky). The inverse of the slope of this function, $\frac{\mu_1 - r_f}{\sigma_1}$, is referred to as the Sharpe ratio of asset 1. More generally, the Sharpe ratio of an asset or portfolio is given by its excess return divided by its standard deviation. This relationship is shown in Figure 2.2.

Figure 2.2: With a Riskless Asset.


Remark 3.1 (A Two-Step Problem). When the investor selects his portfolio by choosing among $N$ risky assets and a riskless asset, he effectively is selecting two things: 1) an optimal mix of risky assets, 2) an allocation his funds between this optimal mix and the riskless asset. It turns out that these two steps are independent in the sense all agents will select the same portfolio in step 1 regardless of their allocation decision in step 2. This can be seen from the fact that the investors problem from section I becomes

\[
\max_{w \in B(\mu^*)} w'\Sigma w
\]

where

\[
B(\mu^*) = \{w : w'\mu + (1 - w'1)r_f = \mu^*\}
\]

where $w$ is the holding of risky assets and $(1 - w'1)$ is the holding of the riskless asset implied by the budget constraint (i.e. the one constraint in (3.2) incorporates both the budget and the target expected return constraints described in (1.2)). $\mu$ and $\Sigma$ continue to denote the mean and variance of the joint distribution of risky assets.
Theorem 3.2 (The Optimal Mix of Risky Assets). Given a target level of return \( \mu^* \), the weight on the risky assets is

\[
w^*(\mu^*) = A(\mu^* - r_f)
\]

where

\[
A = \Sigma^{-1}(\mu - r_f)\Sigma^{-1}(\mu - r_f)'.
\]

That is, all efficient portfolios lie on a line and there is two-fund separation.

Remark 3.3 (Optimal Mix and Maximal Sharpe Ratio). It can be shown (using only matrix algebra and calculus) that the portfolio of risky assets with maximum Sharpe ratio is the optimal mix of risky assets for some target level of expected returns \( \mu^* \). This result is important because it follows from the definition of two-fund separation that any efficient portfolio is a weighted average (i.e., portfolio) of the portfolio with maximal Sharpe ratio and the riskless asset. This gives us a special case of the two asset case with a riskless asset from section II. Of course, the asset with maximal Sharpe ratio is the tangency portfolio \( t \) since the slope of the line through the riskless asset and any portfolio in mean-standard deviation space is, by definition, the Sharpe ratio of the portfolio. As a result, we know that the efficiency frontier with a riskless asset is a line as shown in Figure 3.1. This line is called the capital market line (CML). Note that this implies that the CML is really the efficiency frontier generated by the \( N \) risky assets and the riskless asset.


Remark 4.1 (CAPM). Portfolio theory, as described in sections I through III, is a partial equilibrium model because it only describes the behavior of individuals
and takes the prices (i.e. return distributions) of assets as given. The capital asset pricing model (CAPM) is a general equilibrium extension of portfolio theory where prices are simultaneously determined with investor behavior (however, investors in this model act as though prices are given - this is reasonable so long as every investor is sufficiently small and can’t affect prices through his portfolio decision). The essence of the CAPM is that the market portfolio is an efficient portfolio. From this statement, a little mathematics (somewhat related to Sharpe ratios) gets us the famous formula that relates expected excess returns of individual assets to the expected excess returns of the market portfolio. Namely, this relationship is found to be linear with an intercept of zero. As a description of the real world, the CAPM suffers from the same problems as portfolio theory, however, it also benefits from the generally held belief among academics that its assumptions are fairly acceptable approximations and its conclusions are, at worst, informative for the purpose of investment and pricing. The CAPM was developed by William Sharpe (see Sharpe (1964)) and related work by Lintner and Mossin. Sharpe also won the Nobel prize in economics for this contribution. It should also be noted that a key step in moving from portfolio theory to the CAPM, namely the two-fund separation result, was first illustrated Tobin (1958).

**Assumption 4.2 (Portfolio Theory with a Riskless Asset).** The CAPM makes the same assumptions as portfolio theory and additionally assumes that there is a riskless asset with return \( r_f \).

**Definition 4.3 (The Market Portfolio).** The market portfolio is defined as the portfolio that consists of all assets in the economy. In the context of this model, the market portfolio is the aggregate of all bonds and stocks owned by investors. The weight on each asset in the market portfolio is equal to the fraction of the asset’s value to the market’s value. We can call this a value-weighted portfolio of all the stocks and bonds in the market. The value of an asset is defined as the number of the stocks of this asset outstanding times the price of each stock.

**Remark 4.4 (The Market Portfolio is an Efficient Portfolio).** It turns out that the market portfolio is an efficient portfolio. This result comes directly from the realization that portfolio theory with a riskless asset exhibits two-fund separation. To see this, note that each investor selects an efficient portfolio (i.e. a portfolio on the CML) and, by definition, the market portfolio is an aggregate (i.e. it’s weights are a weighted average of each investor’s portfolio weights). From this, just apply result 2 from 1.6 (which applies to the case with the riskless asset because there is two-fund separation). Furthermore, if we assume that there is zero net supply of the riskless asset, the market portfolio is also on the portfolio frontier of the risky assets. Since the CML and this frontier intersect at only one point, we can exactly locate the market portfolio. Go back to Figure 3.1 for an illustration. We will make the assumption of zero net supply of the riskless asset throughout this section for simplicity. Not assuming this does not change any of the CAPM’s important result (i.e. market is efficient and the pricing relationship). Furthermore, to the extent that we believe riskless assets are only generated by lender/borrower relationships, this assumption is valid (i.e. one investor’s borrowing is another investors lending - they cancel out). An implication of this result is that all investors should only hold the market portfolio and some riskless assets. The is due to the fact that all efficient portfolios are a weighted average of the market portfolio and the riskless.
asset. This conclusion is often used to argue in favor of index fund investing where it is presumed that the index fund holds the market (i.e. is the market portfolio).

**Theorem 4.5** (Slope of the CML). It follows immediately from the efficiency of the market portfolio that the slope of the CML is given by

\[ \theta_m = \frac{\mu_m - r_f}{\sigma_m} \]

where \( \theta_m \) is the Sharpe ratio of the market portfolio and \( \mu_m \) and \( \sigma_m \) are the mean and standard deviation of the returns of the market portfolio.

**Remark 4.6** (Making Use of the Two Asset Case). An important mathematical fact is the efficiency frontier generated by any asset \( i \) and the market portfolio \( m \) is tangent to the CML at the point \( m \). This is due to the fact that the frontier is a hyperbola (hence, it’s upward sloping part is continuously differentiable). To see this, notice that if the two slopes were not equal at \( m \), then the efficiency frontier generated by \( m \) and \( i \) would cross the CML indicating that there is a portfolio that only contains \( m \) and \( i \) that does better than any combination of the \( N \) risky assets and the riskless asset. Of course, this isn’t possible since investors can’t be better off when additional constraints are placed on their investment behavior, therefore the slopes must be equal. This observation is illustrated in Figure 4.1.

Economically, we can interpret the equality of the two slopes as conditions (one for each asset) saying that no investor wants to change the configuration of the risky part of their portfolio (which is the optimal mix - market/tangency portfolio), not even a little bit. Surprisingly, a little manipulation of this constraint on the equality of the slopes gives us the CAPM pricing formula for individual assets described in the following theorem.

\[ \mu_i - r_f = \beta_i (\mu_m - r_f) \]
where $\beta_i = \frac{\rho_{im} \sigma_i}{\sigma_m} = \frac{\text{Cov}(r_i, r_m)}{\text{Var}(r_m)}$ is called the beta of asset $i$ on the market and $(\mu_m - r_f)$ is the excess return on the market portfolio. This relationship (between risk and return) is linear with an intercept equal to zero. This formula is usually referred to as the Sharpe-Lintner CAPM.

**Remark 4.8 (About the Last Two Remarks).** The last two remarks describe the two BIG tricks that are used to derive the CAPM. Given the mathematics of the portfolio frontier, these tricks reduce the derivation of the CAPM formula to a triviality (plus a little algebra). I find these tricks very neat and innovative, both economically (1st trick) and mathematically (2nd trick).

**Remark 4.9 (Alpha).** Forgetting CAPM for a second, the intercept, $\alpha_i$, of a risk-return relationship for asset $i$

$$\mu_i - r_f = \alpha_i + \beta_i(\mu_m - r_f) \quad (4.3)$$

is called the alpha of the asset. The CAPM remarks that, if certain assumptions hold, the alphas for every asset must be zero.

**Remark 4.10 (Beta).** Taking $\mu_m$, $\sigma_m$, and $r_f$ as given, the CAPM pricing formula say that any variation in the expected return of an asset, $\mu_i$, is due to variation in its beta, $\beta_i$, or, more specifically, variation in the covariance of the asset’s return with the return on the market portfolio. *This tells us that an asset’s beta fully summarizes its risk.* Although an investor’s risk (i.e. the dislikable characteristic of his portfolio’s returns) is defined as the variance in his portfolio’s return, the idea that more volatility must imply higher expected returns only holds true for efficient portfolios - not necessarily for individual assets. That is, it *can* be the case that an asset has lower expected return than another asset despite the fact that it has a much higher variance. Of course, this implies that this asset (which may seem like a bad deal to a naive investor) has a lower beta than the other asset. That is, its return covaries less with that of the market (i.e. has a significantly lower coefficient of correlation with the market). This is precisely the point, when considering an assets risk, you should only care about the component of variance that is undiversifiable, in other words, the asset’s *contribution to the risk of the portfolio.*

**Remark 4.11 (Security Market Line).** A plot of the relationship between an asset’s expected return and its beta is easy to obtain and is illustrated below in Figure 4.2. As mentioned earlier, this relationship is a line. It no longer has a zero intercept because we are talking about expected returns, not excess returns (this means the intercept is equal to the riskless rate). This line is called the *security market line* (SML). It is important to keep in mind that *every* asset and portfolio lies on the SML.

**Remark 4.12 (More General CAPM Pricing Formula).** If you look carefully at the proof of theorem 4.7, you’ll notice that we didn’t actually need to assume that $m$ was the market portfolio. All we assumed is that $m$ is on the CML and isn’t the riskless asset. Therefore, we can extend the CAPM in the following way:

- **Extension:** For every portfolio $q$ on the CML and any asset $i$, we have that

$$\mu_i - r_f = \beta_{iq}(\mu_q - r_f) \quad (4.4)$$
where $\beta_{1q} = \frac{\text{Cov}(r_i, r_q)}{\text{V}(r_q)}$.

We can also prove a version of the CAPM formula when there is no riskless asset. The derivation of this version also makes extensive use of the two asset case from section II and the fact that there is two-fund separation on the efficiency frontier generated by only the risky assets. It is due to Fisher Black (see Black(1972)).

![Figure 4.2: Security Market Line (SML).](image)

5. TESTS OF THE CAPM: GENERAL GUIDELINES.

Remark 5.1 (What Do Econometricians Look at for Tests?). Tests of the CAPM often focus on its three most prominent conclusions: 1) the alpha for every asset equals zero, 2) there is a linear relationship between an asset’s beta and its return, and 3) beta completely summarizes expected returns (i.e. completely describes risk). Some tests also look to see if the market risk premium is positive (which is also a prediction of the CAPM since investors dislike risk - this is consistent with the claim that the market portfolio lies on the upward sloping part of the efficiency frontier).

Remark 5.2 (Looking at Alpha). We can think of alpha as the expected abnormal return on an asset. If this value is positive (negative), the CAPM would say this is a good (bad) deal. Of course, in equilibrium, there shouldn’t be good or bad deals, only fair deals. In order to test that we’re in a CAPM-style equilibrium, tests generally estimate realized abnormal returns for each asset using an estimate for the asset’s beta. For each asset, this gives a sequence of realized abnormal returns over the sample (this is called a time series - and there is one for each asset). Tests of the zero alpha property (i.e. all alphas are zero) look at this time series data and can roughly be split into three categories: 1) univariate tests, 2) joint tests, 3) trading strategy tests. Trading strategy tests of the CAPM remark that it shouldn’t be possible to make a risk-adjusted profit over time (using a CAPM implied adjustment for risk) and could, for instance, use time series data on realized risk-adjusted profits of the strategy. We won’t discuss trading strategy tests here.
Remark 5.3 (Looking at the Market Premium). We can approach the question of whether the market premium, $\mu_m - r_f$, is positive, in two easy ways. The most straightforward approach is to take a proxy for the market (e.g. S&P500) and subtract from it a proxy for the riskless rate (e.g. T-bill rate) which yields a time series of estimates of realized excess returns on the market portfolio. Averaging these estimates gives us a rough approximation of the market premium and statistical techniques (e.g. under certain assumption, we can use a t-test) can be used to test whether the market premium is positive. Of course, there are problems with using proxies to generate data used in tests (see section 7). Another approach can be used if we know the betas of each asset. At every point in time, we can take an estimate of the realized market return by computing implied market premia on each asset (using the CAPM formula) and taking a cross-sectional average of these implied values. Again, this average should be “close” to the true realized market return at that time. Doing this at every point in time gives a time series of estimates of the realized market return. Again, using this series, we can do statistical inference (e.g. once more, under certain assumption, we can use a t-test). If we don’t know the betas, we can still do this, however, we must either correct for the measurement error in our betas or use techniques that hopefully sidestep this problem (see section 6 for an example of the latter approach).

Remark 5.4 (Looking at Linearity in Beta). If we know the values of the betas, we can check for nonlinearities by running the following regression over all assets (can do this at each point in time):

\[
rt - rf = at + bt\beta_i + ct(f(\beta_i)) + \epsilon_{it}
\]

where $f(.)$ is a pre-specified nonlinear function (e.g. $f(x) = x^2$) and $\epsilon_{it}$ is the usual stochastic disturbance term. This type of regression is called a cross-sectional regression. If CAPM holds, we know that $ct = 0$ in population for every $t$. We can test this by analyzing the time series estimates of $ct$ by making use of some statistical results (again, see section 6 for an example). Notice that we’ve assumed here that the beta of the asset is constant over our time series sample period. If the sample is large enough (i.e. spans a large enough time period), there isn’t much reason to expect that the betas should stay constant (e.g. companies’ business configurations change over time - expand to new industries, change of focus, etc).

Remark 5.5 (Looking At Explanatory Power of Other Risk Measures). If beta completely summarizes the risk of an asset, adding other “risk measures” (e.g. price-to-earnings ratio, dividend yield, market-to-book ratio, idiosyncratic risk, etc) should not help explain the expected return of assets. We can test this claim by running the regression

\[
r_{it} - r_{ft} = b_i\beta_i + c_t f_{it} + \epsilon_{it}
\]

where $f_{it}$ is the value at time $t$ of an additional ”risk factor” for firm $i$. If this additional risk measure does not explain expected returns, the time-series of $c_t$ should indicate that each $c_t$ is indistinguishable from zero on average. Again, under certain assumptions, there are some statistical results that help us verify this.

Remark 5.6 (Importance of Each Test). If the amount of research done regarding each test category is an indication of relative importance, we would have to conclude that analysis of the zero alpha property and the sufficiency of beta in explaining returns are of highest concern. A great deal of academic work has been directed
on various versions of these tests. The popularity of zero alpha tests may be due to the fact that alpha maps into abnormal returns (recall that abnormal returns are of central importance when thinking about market efficiency). On the other hand, it is also the case that testing this property of expected returns is much more straightforward than the test of nonlinearity, so the focus may be due more to tractibility. Tests looking at the explanatory power of other risk measures are important because they have documented many anomalies of asset pricing. Some academics though have argued that these tests are difficult to interpret because of data mining issues and the lack of theoretical motivation for the use of some of these risk factors. Finally, the test of a positive market premium may be less popular mainly because most people can’t imagine this property not holding, even if the CAPM doesn’t hold).

The Point 5.7 (What Do We Conclude From The Tests?). If these tests seem to indicate that: 1) there is a nonzero alpha, 2) the market premium is less than or equal to zero, 3) factors other than the market premium explain expected returns, or 4) expected excess returns are nonlinear in beta, then this is evidence that the CAPM pricing formula is false and that we don’t live in a CAPM world. Intuitively, we might expect to reject at least one of the conclusions of the CAPM since we clearly don’t live in the world described by the assumptions of the model, yet early work on tests (see Baker, Jensen, and Scholes (1972) or Fama and MacBeth (1973)) were not able to reject the CAPM pricing formula. It wasn’t until the late 70s that evidence against the CAPM began to be discovered and not till much later that this evidence became compelling to most academics. Many initially attributed these anomalies to data mining or measurement error (e.g. of the betas).

Disclaimer 5.8 (Important Test That I’ve Left Out). There is a very important approach to testing that I haven’t mentioned directly in this section: the event study. In some sense, event studies can fall in any of the categories mentioned above where each sample is determined using a pre-specified rule (e.g. the sample consists of returns starting 30 days before the event and ending 60 days after the event). We will not be discussing this test in detail here.


Summary 6.1 (Early Test of the CAPM Implications). Along with Black, Jensen, and Scholes (1972), Fama and MacBeth (1973), FM, was one of the pioneering studies looking at the accuracy of the CAPM in explaining expected returns across assets. It was an ambitious study at the time because it looked systematically at all three aforementioned properties (see section 5). Furthermore, its conclusions were encouraging because it presented support for the CAPM. Yet, possibly with the benefit of 30 or so years of discovery in statistical theory, it is now recognized that the tests employed were not optimal.

Remark 6.2 (CAPM). Recall that the CAPM formula says that

\[ \mu_t - r_f = (\mu_m - r_f)\beta_t \quad \text{or} \quad \mu_t = a + b\beta_t \]

where \( a = r_f \) and \( b = (\mu_m - r_f) \). This can be rewritten in realized return form (as opposed to expected return form):

\[ r_{it} = a_t + b_t\beta_{it} + \epsilon_{it} \]

where \( E[\epsilon_{it}] = 0 \).
Remark 6.3 (An Alternative to the CAPM). In testing the CAPM, FM make use of alternatives that fit within the following general class of expected return models:

\[ r_{it} = \]


Summary 7.1 (Implications of an Incorrect Proxy). When estimating and testing the CAPM, empiricists and investors use a proxy for the market portfolio (e.g. S&P500 index, CRSP EWI, CRSP VWI, etc). This proxy is not stochastically equivalent to the true market portfolio and Roll (1977), R, shows that this can taint conclusions derived from these tests. This is known as the Roll Critique.

Remark 7.2 (The Joint Test). As mentioned in section V, when trying to estimate the CAPM pricing formula or test the validity of the CAPM's conclusions, we are faced with the problem of estimating the parameters of joint returns (i.e. \( \mu \) and \( \Sigma \) - which also determine betas). Various statistics theorems (i.e. versions of the law of large numbers) tell us that there exists unbiased and consistent estimates of these parameters under certain assumptions (need to assume something close to independence in sampling errors though). In this sense, this estimation problem is not a symptom of the theory of finance but more or less a statistical nuisance that vanishes as our sample size approaches infinity. However, R points out another difficulty regarding the CAPM that cannot be resolved with the use of large samples: the unobservability of returns for the market portfolio. More generally, the critique regards the unobservability of the tangency portfolio (among risky assets) because the CAPM pricing formula mathematically holds for the tangency portfolio \( q \) without making any assumptions on investor behavior (remember that the assumptions on investor behavior made by the CAPM imply that the market portfolio is on the CML. It also equals the tangency portfolio if the riskless asset is in zero net supply). This unobservability adds error in the estimation and tests of the CAPM because these procedures rely on the assumption that the proxy used equals the market portfolio. In other words, these tests are really joint tests of the CAPM and the validity of the proxy.

Remark 7.3 (General Discussion: Using the Wrong Proxy). The CAPM formula is obtained directly by the equality of slopes property (see remark 4.6 and proof of theorem 4.7). Of course, the equality rests entirely on the fact that \( m \) lies on the CML. If we use a proxy \( p \) that does not lie on the CML (therefore, isn’t the market portfolio), then the slopes of the efficiency frontier generated by the proxy and an individual asset is no longer constant at the point \( (\sigma_p, \mu_p) \). Figure 7.1 illustrates this claim. The black bold line is the CML while the dotted bold lines are the efficiency frontiers generated using only the proxy and the respective asset (1 or 2). Meanwhile, the purple point is the location of the proxy in mean-standard deviation space. The green bold lines are the efficiency frontiers generated using only the proxy and the respective asset (1 or 2). Clearly, the slopes are not equal. As a result, the CAPM formula does not hold when we use the proxy in place of the market portfolio even if the CAPM formula is correct. The relationship between expected return and the beta with respect to the proxy need not be linear and probably won’t have a zero intercept. This happens only because \( p \) is not on the CML.
Figure 7.1: Violation of Equality of Slopes Property.

The Point 7.4 (Why is This Important?). Remember that most tests of the CAPM look at abnormal returns of assets implied the pricing formula. The idea is that if some stocks have significantly positive or negative abnormal returns, then CAPM fails. Yet, this approach is only valid if we correctly estimate abnormal returns. Since there is bias in our estimation if we use an incorrect proxy, our tests rely centrally on using the correct proxy.

Remark 7.5 (What’s the Market Portfolio?). Given our assumptions of investor preferences (i.e. mean-variance utility over wealth with no nontradeability constraints on assets that enter the wealth equation), the market portfolio must include all assets that contribute to an investor’s future wealth (in value-weighted proportions). Clearly this set of assets includes domestic stocks. In this sense, the commonly used proxies may look adequate. However, the set also includes international stocks, bonds, real estate, human capital, and many more entries. There is no reason to think that the usual proxies lie on the CML once we include all these assets.

Remark 7.6 (Explaining Returns for Subsets of Assets and the CAPM). From the last remark, one might think that using the proxy might at least hold for a pricing formula for domestic stocks. This could be true, but only if the proxy equals the tangency portfolio generated by the domestic stocks. In fact, mathematically, there is a portfolio of domestic stocks that will explain the returns of domestic stocks... but using this proxy isn’t a test of the CAPM. All the conclusions (zero alpha, linearity with beta) of the CAPM rely on the efficiency of the market portfolio (if and only if relationship) and are not independently testable.

Disclaimer 7.7. Though the results mentioned in this summary are accurate (i.e. I have derived them), they may not fully summarize or replicate the findings of R. This is due to the fact that I haven’t carefully read this paper.

Summary 8.1 (Testing Whether \( p \) is an Ex-Ante Efficient Portfolio). This paper focuses on the fact that, if a given portfolio \( p \) is an ex-ante efficient portfolio and there is a riskless asset (i.e. \( p \) is on the CML), \( \alpha_{ip} = 0 \) for every security and portfolio of assets that is considered in computing the efficiency frontier. This is a mathematical fact and does not rely on any assumptions of investor behavior (investor behavior assumptions are only needed to arrive to the claim that the market portfolio lies on the CML). Prior tests of this zero alpha property were univariate (i.e. they only looked to see if \( \alpha_{ip} = 0 \) for a particular asset or portfolio \( i \)). Gibbons-Ross-Shanken (1989), GRS, show that this can lead to misleading results (due to relative inefficiency of univariate tests compared to multivariate ones) and provide a test statistic that incorporates information about the joint distribution of returns and has a known finite sample distribution under both the null hypothesis and a broad class of alternative (non-zero vector of alphas). This test, known as the GRS \( F \)-test, is also attractive because of its economic interpretation (related to Sharpe ratios). Implications for test design (maximize power of test) and some light empirical results are presented in later sections of this paper as well.

Assumption 8.2 (Distribution of Returns). Excess returns of assets (or portfolios) \( R_{it} - r_f \) are contemporaneously related to the excess return on a given portfolio \( R_{pt} - r_f \) through
\[
R_{it} - r_f = \alpha_{ip} + \beta_{ip}(R_{pt} - r_f) + \epsilon_t \quad \text{equiv.} \quad R_{t} - r_f 1_N = \alpha_p + \beta_p(R_{pt} - r_f) + \epsilon_t
\]
where \( \epsilon_t \) is iid and distributed according to the multivariate normal distribution \( N(0, \Sigma) \) (with \( \Sigma \) invertible) and \( (R_{pt} - r_f) \) and \( \epsilon_t \) are uncorrelated.

Remark 8.3 (About the Assumptions). The only really restrictive assumptions made regard the joint distributions of \( \epsilon = [\epsilon_1, ..., \epsilon_T]' \) (notably, iid-ness and joint normality). The invertibility of \( \Sigma \) is achieved by only considering non-redundant assets (including the given portfolio \( p \)) while the assumption of no correlation between the regressor and the disturbance arises from a linear projection. Nonetheless, these assumptions are very restrictive (unfortunately, we need them to have any hope of characterizing our sampling distributions).

Fact 8.4 (Distribution of Sample Estimates). The distribution of the sample intercept (alphas) is given by
\[
\hat{\alpha}_p \sim N(\alpha_p, \frac{1}{T} \hat{\theta}_p ^2 \Sigma)
\]
where \( \hat{\theta}_p = \tau_p/s_p \) is \( p \)'s ex-post Sharpe ratio. Furthermore, the distributions of the sample betas (betas of \( i \) on portfolio \( p \)) and sample covariance matrix are given by
\[
\hat{\beta}_p \sim N(\beta_p, \frac{1}{T s^2_p} \Sigma)
\]
\[
T \hat{\Sigma} \sim W_N(T - 2, \Sigma)
\]
where \( W_N(T - 2, \Sigma) \) is a Wishart distribution (multivariate generalization of \( \chi^2 \) distribution) with \( (T - 2) \) degrees of freedom an covariance matrix \( \Sigma \). \( \hat{\alpha}_p \) and \( T \hat{\Sigma} \) are independent.
Fact 8.5 (Noncentral F-Distribution). If a vector $x$ and matrix $\Omega$ are independent and distributed according to

$$x \sim N(\mu, \Sigma) \quad \text{and} \quad \Omega \sim W_N(M, \Sigma)$$

where $M \geq N$, then it follows that

$$\frac{M - N + 1}{N} x' \Omega^{-1} x \sim F(N, M - N + 1, \lambda)$$

where $\lambda$, the noncentrality parameter of the F-Distribution, is equal to

$$\lambda = \mu' \Sigma^{-1} \mu.$$

Corollary 8.6 (GRS F-Distribution). It follows from the previous two facts that

$$\frac{T - N - 1}{N} \begin{pmatrix} \hat{\alpha}_p' \hat{\Sigma}^{-1} \hat{\alpha}_p \\ 1 + \hat{\theta}_p^2 \end{pmatrix}_{\text{GRS}} \sim F(N, T - N - 1, \lambda)$$

where

$$\lambda = T \frac{\alpha_p' \Sigma^{-1} \alpha_p}{1 + \theta_p^2}.$$

Remark 8.7 (GRS F-test). The GRS F-Distribution provides an exact finite sample distribution for the statistic $\text{GRS}$. We are interested in using this distribution to test whether $\alpha_p = 0$ (the null hypothesis, $H_0$) at a given significance level (say the 5% level). This is easily done since, under $H_0$, $\text{GRS}$ has a central F-distribution and we need only perform the usual F-test.

Remark 8.8 (Alternative Tests). In addition to the univariate tests and the GRS F-test, we could run tests using asymptotic distribution theory (e.g. Wald, LR, and LM tests). Of course, using the parameters of an asymptotic distribution to calculate finite sample test statistics yields only an approximation. GRS discuss this and find that p-values using asymptotic approximations can deviate in significant ways from their exact p-value (found using GRS F-distribution). Notably, the Wald test tends to dramatically underestimate the p-value (i.e. reject too often), especially when the number of return observations, $T$, is small. See Figure 1 from the paper for more details. It has also been pointed out by MacKinlay (1985) that the F-distribution is fairly robust to reasonable deviations from the normality assumption (more so than the asymptotic distributions). Therefore, even if we are willing to ignore the problems associated with using approximations in running tests, to the extent that we’re worried that joint normality of excess returns is not accurate (and there is plenty of evidence that it isn’t), the GRS F-test seems to be preferable to the other tests mentioned here.

Remark 8.9 (Economic Interpretation). Another nice feature of the GRS F-distribution is that $\hat{\alpha}_p' \hat{\Sigma}^{-1} \hat{\alpha}_p$ and $\alpha_p' \Sigma^{-1} \alpha_p$ can be written as function of ex-post and ex-ante Sharpe ratios respectively. GRS show in the Appendix of their paper that

$$\hat{\alpha}_p' \hat{\Sigma}^{-1} \hat{\alpha}_p = \hat{\theta}_q^2 - \hat{\theta}_p^2 \quad \text{and} \quad \alpha_p' \Sigma^{-1} \alpha_p = \theta_q^2 - \theta_p^2$$

where $\hat{\theta}_q$ is the ex-post tangency portfolio’s Sharpe ratio and $\theta_q$ is the ex-ante tangency portfolio’s Sharpe ratio (these two portfolios almost surely have different compositions). In both cases, the tangency portfolio is taken with respect to the
efficiency frontier generated by the $N$ individual assets, who’s returns we are trying to explain, plus the portfolio $p$ (hence, it is generated by $(N + 1)$ points in mean-variance space). The economic content of these relationships is quite meaningful, although probably not surprising. The idea is that the larger the difference between $p$’s ex-post Sharpe ratio and the maximal ex-post Sharpe ratio, the weaker the evidence that $p$ is the ex-ante tangency portfolio. This interpretation is illustrated in Figure 9.1 which sketches the ex-post frontier of risky assets. $q$ is the ex-post tangency portfolio and $p$ and $p'$ are two portfolios whose ex-ante tangency we wish to test. Clearly, we are more likely to accept that $p$ is the ex-ante tangency portfolio since it’s Sharpe ratio (slope of dotted line passing through it) is significantly closer to that of $q$.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure8.1.png}
\caption{GRS Interpretation.}
\end{figure}

The Point 8.10 (Dealing With the Difference Between Ex-Ante and Ex-Post Frontiers). Since investors choose portfolios before they actually observe returns, they are interested in the ex-ante efficiency frontier. This is the frontier generated by population values for the joint mean, $\mu$, and variance, $\Sigma$, of the distribution of returns. In portfolio theory and the CAPM, it is assumed that these values are known to investors. In reality, this clearly isn’t the case and investors (and econometricians!) must estimate the moments of the joint distribution. If investors don’t have any a priori or interim information about returns (e.g. due to fundamental analysis), a good candidate for this estimate is the respective sample mean, $\hat{\mu}$, and sample variance $\hat{\Sigma}$. These parameters are then used to calculate the ex-post efficiency frontier. Of course, a portfolio that is ex-ante efficient will not necessarily be ex-post efficient. This is due to sampling error and this should be taken into account when trying to decide whether or not a given portfolio is ex-ante efficient using realized return data (ex-post stuff). GRS show that even though an ex-ante efficient portfolio isn’t generally on the ex-post frontier, it should be sufficiently close in that this portfolio’s ex-post Sharpe ratio shouldn’t be much smaller than the ex-post maximal Sharpe ratio.
Remark 8.11 (Power the of GRS F-Test). In addition to its significance, we’re also interested in the power of the GRS F-test with respect to interesting and relevant alternatives. The alternatives GRS consider are $\alpha_p \neq 0$. From (5.9), we see that this changes the distribution of the GRS statistic by producing a nonzero value for the noncentrality parameter $\lambda$. It is precisely when the alternative implies a high value of $\lambda$ that this test has high power. It immediately follows from (5.10) that the power of the test increases as the ex-ante inefficiency of $p$ increases. Figure 2 from their paper illustrates this idea for various values of $N$ and $T$. As an application of this analysis, we would like to design our test (e.g. choose $N$ given $T$) such that the power is maximized. GRS tackle this under a special case and find that choosing $N$ equal to roughly 1/3 to 1/2 of $T$ produced approximately optimal tests.

Remark 8.12 (Light Empirical Implementation). GRS present a few implementations of their test:

- **Test 1**: 10 beta portfolios. Test ex-ante tangency of CRSP EWI. Similar to findings by Black, Jensen, and Scholes (1972), it is found that the efficiency of CRSP EWI cannot be rejected.

- **Test 2**: 12 industry portfolios. Test ex-ante tangency of CRSP VWI. Reject efficiency of CRSP VWI at the 1 percent significance level. Contrast with fact that each of the 12 basic univariate tests fails to reject at even the 5 percent level.

- **Test 3**: 10 size portfolios. Test ex-ante tangency of CRSP VWI. Fail to reject efficiency of CRSP VWI. Contrast with fact that univariate tests tend to reject.

These tests highlight some of the problems associated with univariate tests. The problem is related to the efficiency of tests, univariate tests completely ignore relevant covariances in the joint distribution of returns. Figure 4 from the paper demonstrates that the relevant covariances for test 2 and 3 exhibit particular patterns that provide important information that can (and should) be used in testing.

Conclusion 8.13 (Recap). At the general level, GRS show us that making use of all information available in developing tests is important and highly relevant. In applying the multivariate methodology to a test of the ex-ante tangency of a given portfolio $p$, they provide a useful tool to dealing with the Roll Critique (Roll, 1977). Further, they demonstrate that implementation of this test is, broadly speaking, economically intuitive in that it is closely related to the difference in Sharpe ratio of $p$ with the ex-post tangency portfolio $q$. Correspondingly, the power of this test is increased as the alternative postulates that $p$ is more inefficient.


Summary 9.1. Shanken (1987), S, proposes a measure of how good or bad a set $P$ of proxies are for an assumed expected return factor $f$. This measure is quite intuitive (i.e. it is just the multiple correlation of $P$ and $f$) and has a significant impact on the levels of expected abnormal returns we can observe (notably, it allows us to place an upper bound on a specific quadratic form of these expected abnormal returns). This result is especially useful when $P$ are returns on portfolios. This fact is used to test hypotheses about the correlation between a specific proxy (CRSP EWI) and expected return factor.
**Assumption 9.2 (Expected Return Model).** It is assumed that expected return is linear in the covariance of itself and a particular scalar factor $f$. That is,

$$E[R] = r_f 1_N + \text{cov}(R, f).$$

A special case of this assumption is the CAPM where

$$f = \theta_m$$

with $\theta_m$ being the Sharpe ratio of the market portfolio.

**Remark 9.3 (What’s Really Being Assumed?).** Expression (9.1) is not an assumption unless we clearly specify the identity of $f$. For instance, it is a mathematical fact that any portfolio on the efficiency frontier generated by $(r, R)$ will satisfy this expected return relationship (in particular, the tangency portfolio $t$ will as well). This means that there always exists a factor that satisfies (9.1), we just don’t necessarily know what it is. Of course, (9.1) along with (9.2) (i.e. the CAPM) requires assumptions because we specify the identity of the factor. Notice that this was the main point of the CAPM... with the factor being identified, presumably we could estimate expected returns. However, Roll pointed out that the factor in question, $\theta_m$, was not observable and estimation (or tests) were not possible. Like the Gibbons, Ross, and Shanken paper, $S$ can be viewed as a response which tries to deal with this critique.

**Fact 9.4 (A Statistical Bound).** Consider the observable regression of the multiple proxy returns $P$ on $R$

$$R = a + BP + \epsilon$$

and the unobservable regression of the

$$f = a_f + b_f P + e_f$$

where $\epsilon \sim (0, \Sigma)$, $\rho$ is the multiple correlation between $P$ and $f$, and $\sigma_f^2$ is the variance of $f$. It can be shown using the projection theorem that

$$\text{cov}(\epsilon, e_f)'\Sigma^{-1}\text{cov}(\epsilon, e_f) \leq \sigma_f^2(1 - \rho^2)$$

where equality holds if and only if there is a linear relationship between $e_f$ and $\epsilon$, that is,

$$e_f = \alpha + \beta \epsilon.$$

**Remark 9.5 (Remark on This Bound).** It is important to notice that neither $\text{cov}(\epsilon, e_f)$, $\sigma_f^2$, or $\rho$ are observable, therefore, at this point, it is not possible to estimate them. We will deal with some of this unobservability later, yet it should be noted that some academics believe that the usual proxies (e.g. CRSP EWI) are have fairly high $\rho$ with the factor $f = \theta_m$ (i.e. market portfolio).

**Definition 9.6 (Multiple Correlation).** As usual, the vector correlation of $P$ and $f$ is given by

$$\rho(P, f) = V[P]^{-1/2}\text{cov}(P, f)V[f]^{-1/2}.$$  

Meanwhile, the multiple correlation, which is given by

$$\rho = \sqrt{\rho(P, f)'\rho(P, f)},$$

is just a scalar representation of $\rho(P, f)$. 

**Fact 9.7** (Asset Pricing Bound). Given assumption 9.2, it follows from fact 9.4, that

\[ d' \Sigma^{-1} d \leq \sigma_f^2 (1 - \rho^2) \]

where \( d = E[R] - r_f 1_N - B \text{cov}(P, f) \).

**Assumption 9.8** (Proxies are Portfolios). Assume that the set of proxies are all portfolios and therefore can be explained using the expected return model from assumption 9.2. That is,

\[ P = R_P \quad \text{and} \quad E[R_P] = r_f 1_K + \text{cov}(R_P, f) \]

where \( K \) denotes the fact that we are using \( K \) proxies.

**Fact 9.9** (Workable Asset Pricing Bound). Given assumption 9.8, it follows from fact 9.7, that

\[ d' \Sigma^{-1} d \leq \theta_P^2 (\rho_t^2 - 1) = \theta_t^2 - \theta_P^2 \]

where \( d = E[R] - r_f 1_N - B \text{cov}(P, f) \), \( \theta_t \) is the Sharpe ratio of the tangency portfolio generated by \((r, R, P)\), \( \theta_P \) is the maximal Sharpe ratio achievable by \( P \), and \( \rho_t \) is the multiple correlation of \( P \) and \( t \).

**The Point 9.10** (Why Is Only The Last Bound Workable?). Even if we are willing to assume some lower bound estimate of the multiple correlation between the proxies and the factor (i.e. \( \rho \)), it is difficult to imagine that in general we can estimate the variance of the factor (i.e. \( \sigma_f^2 \)) or the covariance vector (i.e. \( \text{cov}(P, f) \)). To this extent, fact 9.9 is workable because \( d \) can be estimated with observables, as can \( \theta_P \) and \( \theta_t \). In essence, this bound allows us to do two things: 1) assuming that we 'know' \( \rho \), we can test the expected return model by seeing if the workable bound is estimated to hold, or 2) assuming that we 'know' the expected return model is true, we can test whether \( \rho \) is greater than a certain value. We can think of the later test as a joint test of whether the expected return model is true and \( \rho \) is above a certain value. This is certainly an improvement over Roll’s view of testing a model (e.g. CAPM).

**The Point 9.11** (Interpretation of \( d \)). Clearly, \( d \) is just a generalization of \( \alpha \) from Gibbons, Ross, and Shanken. It is simply the vector of excess returns implied by the proxies used in estimation (can think of this as the vector of deviations - hence \( d \) - of the model). Mathematically, an estimate of \( d \) is the intercept of the multivariate regression of \((R - r_f 1_N)\) on \((R_P - r_f 1_K)\). Therefore, it shouldn’t be a surprise that we obtain the RHS expression of the inequality in (11). Furthermore, from GRS, we know that

\[ \frac{(T - N - K)T}{N(T - K - 1)} \cdot \frac{d' \Sigma^{-1} d}{1 + \hat{\theta}_P^2} \sim F(N, T - N - K, \lambda) \]

where

\[ \lambda = T \frac{d' \Sigma^{-1} d}{1 - \hat{\theta}_P^2} \]

**Remark 9.12** (The Shanken Test). The workable bound implies that an upper bound on the noncentrality parameter of the \( F \)-distribution is

\[ \lambda \leq T \frac{\theta_t^2 - \theta_P^2}{1 - \theta_P^2} \]
This can be tested rather easily using data.

Remark 9.13 (Empirical Implementation). S presents a few implementations of his test:

- **Test 1**: 20 size portfolios \((N = 20)\), test on equal length subsamples of 2/53 to 11/83 but delete January data \((T = 68\) for each subsample), use CRSP EWI as proxy, and use T-bill data for \(r_f\). Test if proxy is perfect assuming CAPM is true. Reject with an aggregate p-value of 0.02. See Table 1.

- **Test 2**: Same setup as test 1. Test if \(\rho \geq 0.7\) (again, under the assumption that the CAPM is true). Problem: need to know true value of \(\theta_P^2\) (called *nuissance parameter*). With assumptions, estimate using bias corrected MLE to get a value of 0.023 with standard deviation 0.019. Using different approach ranging from simply using the estimate, to using a conservative estimate (i.e. \(\theta_P^2 = 1\)), and a three-point posterior of \(\theta_P^2\) (see Table 4), they still reject at around the 0.06 level (except in conservative case where they get a p-value of 0.11 - still very low). See Table 3 for details. These results indicate that the CRSP EWI accounts for less than half of the variation in the factor (in this case, the return on the true market portfolio). Some sensitivity analysis is performed (i.e. try using slightly different values of the moments of the estimate of \(\theta_P^2\), and see what happens - Table 5).

- **Test 3**: Same general setup and procedure as in tests 1 and 2, but use two proxies simultaneously: CRSP EWI and a long-term bond index. Still reject perfect correlation and \(\rho \geq 0.7\) at around the 0.02 and 0.1 levels, respectively. See Tables 7 and 8.

Initially, one might test these tests highlight the problem of using common proxies for the market portfolio in tests of the CAPM. This is not exactly true. What these tests do show is that one of the following is probably true: 1) the CAPM is *invalid*, or 2) the usual proxies account for at *most* 1/2 or 2/3 of variation in the market return. In many ways though, the findings are stronger if we’re willing to move away from tests of economic content (i.e. the tests do indicate that the usual proxies are problematic substitutes for the actual expected return factor). Of course, we’re really mainly interested in this economic content.

Conclusion 9.14 (Recap). The point of this paper is that lower bounds on the multiple correlation of \(P\) and \(f\) (i.e. how much variation in \(f\) is explained by \(P\)) allow us to place upper bounds on the population deviations from a multifactor regression of returns (with intercept set to \(r_f\) and explanatory variables as excess portfolio returns). This yields tests in sample that have known finite sample distributions (like GRS F-test). These tests can be used to draw inference on \(\rho\) under the assumption that the expected return model is true (e.g. CAPM).

10. Fama-French (1993): **Empirically Motivated Alternative to the CAPM.**

11. MacKinlay (1995): **Sources of Deviations From CAPM.**

Summary 11.1 (Why Is CAPM Wrong?). Let’s drop the Roll Critique for now and take the evidence of non-negative alphas as given. An interesting question
one might ask is what’s the source of deviation from the CAPM. The two main stories are based on risk and non-risk. MacKinlay (1995), M, focuses on the risk story while arguing that these two stories differ in an important way, namely that Sharpe ratios can be bounded in the risk case and may, in theory be unbounded (e.g. because of arbitrage opportunities), in the nonrisk case. This has an implication on the GRS type test by placing an upper bound on the noncentrality parameter in the risk case. Furthermore, M remarks that: 1) the usual GRS type test will have difficulty distinguishing between the CAPM and other risk-based stories, and 2) evidence suggests that non-risk stories better explain the deviations from the CAPM.

Remark 11.2 (The Usual Suspects). It is useful to begin with a description of the prominent explanations academics have enumerated for deviations from the CAPM:

- **Risk Stories:**
  1) **Misidentification of Factors**... this is the Roll Critique (not considered by M), and 2) **Missing Factors**... left out factors such that when these factors are included along with the market proxy, the CAPM will hold (with a linear combination of the multiple proxies being stochastically equivalent to the market portfolio).

- **Non-risk Stories:**
  1) **Empirical Difficulties**... for example, data snooping or bid-ask bounce bias, 2) **Market Frictions**... for example, transactions costs or liquidity effects, and 3) **Investor Irrationality**... for example, investor overreaction or over-extrapolation.

Figure 11.1 lists these explanations.

**Figure 11.1:** Violations of the CAPM.

Remark 11.3 (A More General Perspective). The methodology developed in M is not confined to explaining deviations from the CAPM. It can be used in the analysis of any multifactor model. Hence, throughout the theory part of his paper, he deal
with a baseline multifactor model (i.e., the null hypothesis) given by

\[(R_t - r_f) = \alpha + B(R_{pt} - r_f) + \epsilon_t\]

where there are \(N\) assets, \(K\) factors. For convenience, we rewrite this as

\[(11.2) z_t = \alpha + Bz_{pt} + \epsilon_t\]

and let \(\epsilon_t \sim (0, \Sigma), z_{pt} \sim (\mu_p, \Omega_p), \text{ and } \text{cov}(z_{pt}, \epsilon_t) = 0.\)

**Remark 11.4 (The Optimal Orthogonal Portfolio).** We know that \(\alpha = 0\) if the tangency portfolio \(t\) is a linear combination of the factors. \(M\) shows that if this isn’t the case, there exists a single portfolio \(h\) which is orthogonal to the factor portfolios such that the tangency portfolio \(t\) is a linear combination of the factors and \(h\). It follows that

\[(11.3) z_t = Bz_{pt} + \beta_h z_{ht} + u_t\]

where \(u_t \sim (0, \Phi), z_{ht} \sim (\mu_h, \sigma_h^2), \text{ and } \text{cov}(z_{pt}, u_t) = \text{cov}(z_{ht}, u_t) = 0.\)

**Fact 11.5 (Relationship Between \(\Sigma\) and \(\alpha\)).** It follows that

\[(11.4) \alpha = \beta_h \mu_h\]

and

\[(11.5) \Sigma = \alpha \Sigma_h + \Phi.\]

That is, deviations from the model, \(\alpha\), must be accompanied by a common component in residual variance.

**Remark 11.6 (An Upper Bound).** Recall that

\[(11.6) GRS = T - N - K \cdot \frac{T \hat{\theta}_t^2 - \hat{\theta}_p^2}{1 + \hat{\theta}_p^2} \sim F(N, T - N - K, \lambda)\]

where

\[(11.7) \lambda = T \frac{\hat{\theta}_t^2 - \hat{\theta}_p^2}{1 + \hat{\theta}_p^2}.\]

It is rather trivial to show that

\[(11.8) \theta_h^2 = \theta_t^2 - \theta_p^2\]

hence we can rewrite

\[(11.9) \lambda = T \frac{\theta_h^2}{1 + \theta_p^2} < T \theta_h^2\]

which places an upper bound on the noncentrality parameter. We can think of this as a limit on the difference between the null hypothesis (i.e., multifactor model is true, \(\alpha = 0\)) and the risk-based alternatives. There is no reason to assume that this bound exists in non-risk stories.

**Remark 11.7 (Empirical Implementation).** Prior to thinking of tests, we need an estimate of \(\theta_h^2\) to make this upper bound workable. \(M\) does the following: 1) he assumes that \(\theta_t^2 = 0.031\) which corresponds to an excess return of 10% and a standard deviation of 16%, and 2) using the CRSP VWI as his proxy, he assume that \(\theta_p^2 = \hat{\theta}_p^2.\) I think both assumptions are quite bad since we don’t really know where the first one comes from and the second assumption is likely to overstate \(\theta_p^2\) and
therefore underestimate \( \theta^2_h \) which screws everything up. Nevertheless, this yields an estimate of \( \theta^2_h = 0.021 \) which can be used to determine the noncentrality parameter of the risk based alternative. See Table 1 for details. M also considers two non-risk based alternatives which assume that \( \alpha \) is normally distributed with mean 0 and standard deviation 0.0007 and 0.001 respectively (i.e. annual spread of 3.4 and 4.8 percent respectively). See p.16 for details. For his tests, these assumptions give noncentrality parameters of 0 (under the null), 7.1 (alternative risk-based), 39.4 (non-risk 0.0007), and 80.3 (non-risk 0.001).

- **Test 1:** Monthly data from 6/63-12/91 \((T = 342)\). Use \( N = 32 \) portfolios (formed on what? like FF1993?). Finds that null and risk case overlap quite a bit. Therefore, it’s difficult to distinguish between these two cases. Non-risk cases have higher noncentrality and fit the data better. M argues this is evidence for non-risk stories. See Figure 1.

- **Test 2:** Weekly data over the sample \((T = 1591)\). Motivation is market frictions like bid-ask bounce. Similar results as test 1. See Figure 2.

Overall, I find these tests very unconvincing. The noncentrality parameters of the non-risk alternatives seem way too sensitive to conjectures of \( \sigma_\alpha \) while the noncentrality parameter of the risk based alternative is also sensitive to assumptions that really aren’t supported by evidence.

**Conclusion** 11.8 (Recap). M sets out on an ambitious path. Pinpointing the sources of deviations from a multifactor pricing models is very important and challenging. However, with the exception of pointing out a few basic facts (existence of optimal orthogonal portfolio and relationship between \( \Sigma \) and \( \alpha \)), I just don’t feel much is accomplished in this paper. It is true that \( h \) can be used in debates about FF1993’s results, but the counterargument in favor on non-risk based alternatives strikes me as silly. If anything, a little more motivation regarding the assumptions would have been useful to add credibility to the findings in this paper. The paper has a good introduction though...

12. **Appendix.**

**Proof** 12.1 (Proof of Theorem 1.4). The constrained optimization described in (1.1) and (1.2) is equivalent to the unconstrained optimization

\[
\max_{w, \lambda_1, \lambda_2} w' \Sigma w - \lambda_1 (w' 1 - 1) - \lambda_2 (w' \mu - \mu^*)
\]

which has first-order (necessary) conditions

\[
\begin{align*}
2 \Sigma w - \lambda_1 1 - \lambda_2 \mu &= 0 \\
w' 1 &= 0 \\
w' \mu &= 0
\end{align*}
\]

which, solving for \( w \) yields

\[
w^*(\mu^*) = \frac{1}{\Gamma_1 \Gamma_3 - \Gamma_2^2} \Sigma^{-1}(\Gamma_3 1 - \Gamma_2 \mu) + \frac{1}{\Gamma_1 \Gamma_3 - \Gamma_2^2} \Sigma^{-1}(\Gamma_1 \mu - \Gamma_2 1) \mu^*.
\]
where $\Gamma_1 = 1'\Sigma^{-1}1$, $\Gamma_2 = 1'\Sigma^{-1}\mu = \mu'\Sigma^{-1}1$, and $\Gamma_3 = \mu'\Sigma^{-1}\mu$. It can be verified using matrix algebra that this can be written as
\begin{equation}
(12.6) \quad w^*(\mu^*) = w_p + w_a \mu^*
\end{equation}
where $\theta_p$ is a portfolio (the efficient portfolio with expected return equal to 0) and $\theta_a$ is a zero-investment portfolio (i.e. $\theta_a'1 = 0$). Second-order conditions can easily be checked since $\Sigma$ is positive definite (i.e. variances are always positive).

**Proof 12.2 (Proof of Theorem 1.6).** Since $\theta^*(\mu^*) = \theta_p + \theta_a \mu^*$, we have the following relationship between the mean and standard deviation of returns of efficient portfolios
\begin{equation}
(12.7) \quad (\theta_p + \theta_a \mu^*)' \Sigma (\theta_p + \theta_a \mu^*) = \sigma^*(\mu^*)
\end{equation}
From this point, algebraic manipulations yield the stated result.

**Proof 12.3 (Proof of Theorem 2.3).** A portfolio that achieves an expected return of $\mu^*$ must satisfy the system of equations
\begin{align}
(12.8) \quad w_1 \mu_1 + w_2 \mu_2 &= \mu^*
(12.9) \quad w_1 + w_2 &= 1
\end{align}
which has a unique solution given by (2.2) which can be rewritten as (2.3). (2.4) and (2.5) follow from the definition of standard deviation and taking a derivative (have to use the chain rule), respectively.

**Proof 12.4 (Proof of Theorem 3.2).** Proof is along the lines of the proof of theorem 1.4 except that there is now only one constraint to deal with (hence, only one Lagrange multiplier). Everything else is just algebra.

**Proof 12.5 (Proof of Theorem 4.5).** It suffices to observe that both the riskless asset and the market portfolio are efficient. Therefore, it follows that the points $(0, r_f)$ and $(\sigma_m, \mu_m)$ lie on the CML. Calculating the slope implied by these two points gives (4.1).

**Proof 12.6 (Proof of Theorem 4.7).** From (2.5) and (4.1), it follows that
\begin{equation}
(12.10) \quad \frac{\mu_m - r_f}{\sigma_m} = \frac{\sigma_m}{C_1 \mu_m - C_2}
\end{equation}
where
\begin{equation}
(12.11) \quad C_1 = \frac{\sigma_i^2 - 2 \rho_{im} \sigma_i \sigma_m + \sigma_m^2}{(\mu_i - \mu_m)^2} \quad \text{and} \quad C_2 = \frac{\mu_m \sigma_i^2 - 2 \rho_{im} (\mu_m + \mu_i) \sigma_i \sigma_m + \mu_i \sigma_m^2}{(\mu_i - \mu_m)^2}
\end{equation}
since we’re evaluating (2.5) at the point $(\sigma_m, \mu_m)$. Notice that it follows that
\begin{equation}
(12.12) \quad C_1 \mu_m - C_2 = \frac{\rho_{im} \sigma_i \sigma_m - \sigma_m^2}{\mu_i - \mu_m}
\end{equation}
or
\begin{equation}
(12.13) \quad \frac{\mu_m - r_f}{\sigma_m} = \frac{\sigma_m (\mu_i - \mu_m)}{\rho_{im} \sigma_i \sigma_m - \sigma_m^2}.
\end{equation}
From this point, it is straightforward to work out that (4.2) holds. Notice that
\begin{equation}
(12.14) \quad \frac{\rho_{im} \sigma_i}{\sigma_m} = \frac{\rho_{im} \sigma_i \sigma_m}{\sigma_m^2} = \frac{Cov(r_i, r_m)}{V(r_m)}
\end{equation}
which the usual interpretation of the beta of an asset.
References


Sloan School of Management. Do not send, copy, or distribute without the prior consent of the author.
E-mail address: jekondo@mit.edu