Investors Do Not Get Paid for Bearing Risk

Harry M. Markowitz

The relationship between the excess return of each security and its beta, where beta is defined as its regression against the return on the market portfolio, is linear in the Sharpe–Lintner ("S–L") Capital Asset Pricing Model ("CAPM"). This linear relationship is often interpreted to mean that CAPM investors are paid for bearing systematic risk. In this article, I will show that this is not a correct interpretation because two securities may have identical risk structures in terms of their covariances with other securities in the market, yet have different excess returns. In fact, if the parameters of the CAPM are generated in a natural way, then securities with the same risk structure almost surely will have different expected returns.

THE MOSSIN VERSION OF THE SHARPE–LINTNER MODEL

Although the premises and conclusions of the S–L CAPM were first presented in Sharpe (1964), I will use Mossin's (1966) systematic formulation of Sharpe's version of the S–L model in this analysis. Mossin states the premise of the Sharpe model explicitly and draws valid conclusions from it, where Sharpe's version is vague about the statement of his premises and deduction of his conclusions, and one of his conclusions is incorrect.[1]

Inputs to Mossin's version of the S–L CAPM model ("Mossin S–L model") include the following: the utility functions of many investors; the number of shares each investor first owns of each stock; and the expected returns per share (not per dollar) and covariances per share (not per dollar) upon which all investors agree. There is also an interest rate input which is only used in the excess return calculation.

Outputs of the Mossin S–L model include market clearing prices, expected returns and covariances per dollar, the composition of the market portfolio, and the regressions against the market portfolio. Along the way, Mossin uses Tobin and Sharpe to prove various things about the S–L model. For example, the Mossin S–L model involves equations that include cash, expected returns, and a Lagrangian multiplier. With a little algebra, the equations can be expressed in excess returns rather than expected returns.

Mossin also cites that Sharpe's version of the Tobin Separation Theorem holds in equilibrium (Sharpe's version of the Tobin Separation Theorem involves borrowing as well as lending, while Tobin's version involves lending only). In Sharpe's version of the Tobin Separation Theorem, each investor's efficient set consists of one particular portfolio of risky securities as well as the ability to borrow or lend. This much of the theorem holds for any mean–variance investor who can borrow without limit (even if no one else can do so), or has the same beliefs, or even seeks mean–variance efficiency. When we assume that all investors have the same beliefs and all seek mean–variance efficiency, then it follows that all investors mix the same portfolio of risky assets and this must be the market portfolio.

[1] In connection with Figure 6 of his paper, Sharpe asserts that as prices adjust some combinations of risky assets can become perfectly correlated with each other. This is associated with a linear portion of the resulting expected return, standard deviation efficient set considering risky securities only. Mossin assumes that the covariance matrix per share is nonsingular. It is not clear whether Sharpe assumes that this covariance matrix of returns is nonsingular, but he certainly permits this, so no matter what positive prices are present in the market, the covariance matrix of returns per dollar invested as well as the covariance matrix per share is nonsingular. This implies that no two distinct linear combinations of risky assets can be perfectly correlated and there can be no linear segment in the set of mean-standard deviation efficient combinations.

The Mossin S–L model is like a giant TV set: You set the dials (e.g., share availabilities) and the screen shows outputs (e.g., prices and expected returns per dollar invested). Like Mossin, I won't worry about the existence and uniqueness of the solutions to the equations involved. I assume that when I flip on the switch the TV works. If I set the dials the same way as yesterday, I get the same output as yesterday.

One noteworthy feature of the Mossin S–L TV is that you can tell how the inputs are set by looking at the outputs. Outputs include prices, and with prices you can figure the original shares available as well as means and covariances per share.

However, there are some conceivable outputs one will never see—namely outputs that could only happen if some assumed given shares were negative (or zero for all investors because such securities would be dropped from the analysis.) But, these would be the securities with zero or negative percent demanded in the market portfolio. We will only see output consistent with any positive vectors $X > 0$ in Equation (1), where ν and σ_{ij} are the excess return and covariance per dollar invested.

$$\begin{pmatrix} \sigma_{11} \\ \sigma_{21} \\ . \\ . \\ . \\ \sigma_{n1} \end{pmatrix} X_1 + \begin{pmatrix} \sigma_{12+} \\ \sigma_{22} \\ . \\ . \\ . \\ \sigma_{n2} \end{pmatrix} X_2 + ... \begin{pmatrix} \sigma_{1n} \\ \sigma_{2n} \\ . \\ . \\ . \\ \sigma_{nn} \end{pmatrix} X_n = h \begin{pmatrix} \nu_1 \\ \nu_2 \\ . \\ . \\ . \\ \nu_n \end{pmatrix} \tag{1}$$

So, how do we find all possible outputs that can occur on the Mossin S–L TV? The question practically answers itself if we write the basic optimization equations as in Equation (1).

We know that the covariance matrix *C* per dollar invested is nonsingular because the covariance matrix per share is nonsingular—the only thing that Mossin tells us about the latter matrix. Given any nonsingular covariance matrix *C*, consider the set of excess return vectors, ν, that result if *X* in Equation (1) is non-negative, $X \geq 0$ (later we consider only $X > 0$). These ν form the "cone", K_C, generated by *C* which is drawn as follows. As in Equation (1), look at each column $\sigma^{(i)}$ of *C* in turn as if it were a point in ν-space. Draw a ray from the origin through $\sigma^{(i)}$ for $i = 1,...n.$ Fill in the middle of the diagram by taking convex combinations of anything you already have. This gives you K_C, the "cone generated by *C*."

But we want $X > 0$ only. We get this by throwing away the exterior of K_C, keeping only the interior K_C^0. Even though we toss the exterior of K_C, there will be plenty of good stuff left as long as $|C| \neq 0$ as assumed.

I call K_C^0 the "compatible cone" (i.e., the cone full of ν-vectors which are compatible with the given *C* in the sense that only these will ever appear together on the Mossin S–L TV). In other words, *given any nonsingular C, pick any* ν-vector from K_C^0. *This combination of C and* ν *may appear on the Mossin S–L TV. These are the only C,* ν *combinations which can appear.* Because in this model *C* is all things risk and ν all things return, in an abstract way we now know the model's possible combinations of risk and return. But, we do not have a concrete picture of what these K_C^0 look like. For starters let's assume that returns are uncorrelated (this is not plausible, but easy to analyze). *C* is then diagonal:

$$C = \begin{pmatrix} V_1 & 0 & 0 & 0 \\ 0 & V_2 & 0 & 0 \\ 0 & 0 & V_3 & 0 \\ 0 & 0 & 0 & V_4 \end{pmatrix}$$

The first column, considered as a ν-vector, lies on the X_1-axis, the second on the X_2-axis, etc. Because the location of a ray from the origin through a point does not depend on which point on the ray we choose, we have

$$K_C = K_I$$

That is, K_C is the same as the cone generated by the identity matrix I, the entire positive orthantant, specifically the entire first quadrant, when $n = 2$. K_C^0 is the interior of this region. Thus, if returns are uncorrelated then any positive $V_1, V_2, ..., V_n$ and $\nu_1, ..., \nu_n$ are permitted. In particular, two securities can have identical V but different ν, thus they can have identical risk structures given our assumption of zero correlation and yet have differing expected returns, therefore differing excess returns.

From Equation (1) with diagonal C we see that when returns are uncorrelated, i.e., stocks with high means and low variances are a large part of the market.

$$X_i = k\nu_i / V_i \qquad (2)$$

Equation (2) may also be written as

$$X_i V_i = k\nu_i \qquad (3)$$

Because $\sigma_{ij} = 0$ for all $j \neq i$, the standard formula for the covariance between a security and a portfolio yields $X_i V_i$ as the covariance between the return on security i and the return on the investor's portfolio. Thus, in this case, Equation (3) says that the covariance of each security with its portfolio is proportional to the securities excess return. If you divide both sides of Equation (3) by the variance of the investor's portfolio you obtain

$$\beta_i = \bar{k}\nu_i \qquad (4)$$

where β_i is the regression against the portfolio. Here $\bar{k} = k / Var(R_p)$ and should be given no other interpretation. Equation (4) is a relationship between a security and one individual's portfolio. But if all investors are essentially the same, then Equation (4) carries over to the market as well.

It is useful to see where Equation (1)—and therefore Equation (4)—come from. For this purpose let us further simplify the Mossin S–L by assuming the following form of utility function

$$E - cV \qquad (5)$$

where c may vary from one investor to the next. To maximize Equation (5), take partial derivatives and the set result to zero to get

$$\frac{\partial V}{\partial X_1} = (1/c)\frac{\partial E}{\partial X_i} \qquad i = 1, ..., n \qquad (6)$$

In other words, each investor is advised to push each security into the investor's portfolio to the point where each security has the same ratio of marginal effect on portfolio variance to marginal effect on portfolio mean. This aggregates up to the portfolio level. Divide by market variance and the well-known CAPM relationship is created.

If returns are correlated, compatible cones are derived by plotting the columns of C as if they were ν vectors, drawing straight lines from the origin through them, and then taking convex combinations of these lines using and the interior of the resulting cone. This is easiest to do for $n = 2$. Note that the cone tends to close up as correlation increases. If the two securities have a correlation of 1, then the two lines are identical. This is the law of one price. Identically distributed securities must have proportionate expected returns. As long as the C matrix is nonsingular, therefore K_C^O is not empty, two different points can be found in K_C^O not on the same ray through the origin. In this case two securities can have identical risk structures C but differing excess returns. If the initial endowments of investors are drawn randomly from some continuous distribution, then the probability distribution of (C, μ) cases will be continuous. Thus, there is a zero probability that two securities with the same covariances will have the same μ.

AFTERTHOUGHTS

1. Rather than use

$$(\sigma_{1j}, \sigma_{2j}, ..., \sigma_{nj})'$$

as a point in ν-space to help draw a ray from the origin, we could use any other point on the ray, like

$$(\sigma_{1j} / \sigma_{jj}, \sigma_{2j} / \sigma_{jj}, ..., \sigma_{nj} / \sigma_{jj})'$$

Recall that $\sigma_{ij} / \sigma_{jj}$ is the regression coefficient—the beta if you will—of the return of security *i* against that of *j*.

2. Do not expect compatible cones to go away in multiperiod discrete or continuous-time analysis. In particular, there is not much difference between a one-period optimization and a many-period optimization. Bellman (1957) tells us that to optimize a many-period problem you just optimize a sequence of one-period problems. It's a matter of getting the objectives right for the problem-within-a-problem. But often problems-within-a-problem are not all that different from the genuine one-period problem. Therefore, if the genuine one-period model has compatible cones there should also be some lurking around the model's many-period version.
3. The continuous-time model is a special case of the discrete time model. If you do not believe that it is because you are stuck with an old-fashioned numbering system. The Greeks thought you could not squeeze any new numbers between the fractions (a.k.a. the rationals because the Greeks thought that the new numbers were irrational). In fact, in some sense there are more irrationals than there are rationals.

   Now we have infinitesimals between each real (rational or irrational) number. Gottfried Leibniz (1646–1716) thought he had a handle on the infinitesimals, but it had to wait for Abraham Robinson to figure it out rigorously in the mid-20th century. In a dynamic analysis, rather than have time travel along the real continuum, we can have it walk step-by-step along the "hyperfinite time line." The latter is as simple as a garden path, but with infinitesimal distances between successive stepping stones. With the hyperfinite timeline, Brownian motion is what comes out at a macro level if at each infinitesimal time increment you flip a coin and record an infinitesimal gain or loss.

   Thus, it is a small step to imagine a single-period Mossin S–L CAPM where the period is an infinitesimal tick long. After all, where does Mossin say how long the period has to be in a single-period CAPM? A day? A year? A microsecond? An infinitesimal increment? Because the analysis is independent of the time increment, it applis to infinitesimal steps and, therefore, to continuous models. Our conclusion is the same as in the discrete time model: that possible combinations of excess returns which can go with a given covariance matrix will probably form some kind of cone and that two securities with the same risk structure will probably have different expected returns. Therefore, one clearly cannot say that the CAPM investor is paid for bearing risk in either in the single-period, multiple-period, or continuous case.

WHERE DOES THIS LEAVE CAPM?

The points made in this paper do not change the major conclusions of CAPM. Given the assumptions of CAPM, the market is an efficient portfolio and there is a linear relationship between the expected return of each security and its regression against the market. But, we must not interpret this as the bearing of risk. Insofar as the world works like CAPM at an aggregate level, this linear relationship is useful. Someone who wished to issue a new security could estimate its beta, which would indicate the expected return that the market would assign to this new security.

REFERENCES

Albeverio,Sergio, Jens Erik Fenstad, Raphael Høegh-Krohn, and Tom Lindstrøm. 1986. *Nonstandard Methods in Stochastic Analysis and Mathematical Physics.* Academic Press, Inc.: Orlando, FL.

Bellman, Richard E. 1957. *Dynamic Programming.* Princeton University Press: Princeton, NJ.

Mossin, J. 1966. "*Equilibrium* in a Capital Asset Market." *Econometrica,* vol. 34, no. 4 (October):768–783.

Robinson, Abraham. 1966. *Nonstandard Analysis*. North Holland Publ.: Amsterdam.

Sharpe, William F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *The Journal of Finance,* vol. 19, no. 3 (September):425–442.