Ec2723, Asset Pricing I
Class Notes, Fall 2005

Static Portfolio Choice, the CAPM, and the APT

John Y. Campbell[1]

First draft: July 30, 2003
This version: September 26, 2005

[1]Department of Economics, Littauer Center, Harvard University, Cambridge MA 02138, USA. Email john_campbell@harvard.edu.

The principle of participation

We begin by considering a choice between one safe and one risky asset. We make only weak assumptions on preferences and the distribution of returns on the risky asset. An investor with initial wealth w can invest in a safe asset with return r, or a risky asset with return $r + \widetilde{x}$. Final wealth is

$$w(1 + r) + \theta\widetilde{x} = w_0 + \theta\widetilde{x},$$

where θ is the dollar amount (not the share of wealth) invested in the risky asset. The investor's problem is

$$\mathrm{Max}_\theta \; V(\theta) = \mathrm{E}u(w_0 + \theta\widetilde{x}).$$

The first-order condition is

$$V'(\theta^*) = \mathrm{E}\widetilde{x}u'(w_0 + \theta^*\widetilde{x}) = 0$$

and the second-order condition is

$$V''(\theta^*) = \mathrm{E}\widetilde{x}^2u''(w_0 + \theta^*\widetilde{x}) = 0,$$

which shows that the problem is well defined for a risk-averse investor.

We have

$$V'(0) = \mathrm{E}\widetilde{x}u'(w_0),$$

which has the same sign as $\mathrm{E}\widetilde{x}$. The investment in the risky asset should be positive if it has a positive expected return. This is true for any level of risk aversion. Thus we cannot explain non-participation in risky asset markets by risk aversion. We need fixed costs of participation or a kink in the utility function that generates "first-order risk aversion".

Portfolio choice with a small risk

We consider a small risk

$$\widetilde{x} = k\mu + \widetilde{y}$$

and assume $k > 0$. The first-order condition is

$$\mathrm{E}(k\mu + \widetilde{y})u'(w_0 + \theta^*(k)(k\mu + \widetilde{y})) = 0.$$

Differentiating w.r.t. k,

$$\mu \mathrm{E}u'(\widetilde{w}) + \theta^*(k)\mu \mathrm{E}(k\mu + \widetilde{y})u''(\widetilde{w}) + \theta^{*\prime}(k)\mathrm{E}(k\mu + \widetilde{y})^2 u''(\widetilde{w}) = 0.$$

Evaluating at $k = 0$,

$$\theta^{*\prime}(0) = \frac{\mu}{\mathrm{E}\widetilde{y}^2} \frac{1}{A(w_0)}.$$

Then a Taylor expansion for the investment in the risky asset gives

$$\theta^*(k) \approx \theta^*(0) + k\theta^{*\prime}(0) = \frac{\mathrm{E}\widetilde{x}}{\mathrm{E}(\widetilde{x} - \mathrm{E}\widetilde{x})^2} \frac{1}{A(w_0)}.$$

We can divide θ by wealth to find the share of wealth invested in the risky asset. Call this α. We find

$$\alpha^*(k) = \frac{\theta^*(k)}{w_0} \approx \frac{\mathrm{E}\widetilde{x}}{\mathrm{E}(\widetilde{x} - \mathrm{E}\widetilde{x})^2} \frac{1}{R(w_0)}.$$

Portfolio choice in the CARA-normal case

The above formula for dollars invested in the risky asset is exact when risk is normal, $\widetilde{x} \sim N(\mu, \sigma^2)$, and utility is CARA with risk aversion A. In this case the problem becomes

$$\mathrm{Max}_\theta \; V(\theta) = \mathrm{E}[-\exp(\,-A(w_0 + \theta\widetilde{x})).$$

Utility is lognormally distributed (its log is normally distributed) if $\widetilde{x}$ is normally distributed. For any lognormal random variable $\widetilde{z}$, we have the following extremely useful result:

$$\log \mathrm{E}(\widetilde{z}) = \mathrm{E}\log(\widetilde{z}) + \frac{1}{2}\mathrm{Var}\log(\widetilde{z}).$$

The portfolio choice problem is equivalent to

$$\mathrm{Min}\;\log\,\mathrm{E}[-\exp(\,-A(w_0 + \theta\widetilde{x})) = -A(w_0 + \theta\mu) + \frac{1}{2}A^2\theta^2\sigma^2,$$

which is equivalent to

$$\mathrm{Max}\;A(w_0 + \theta\mu) - \frac{1}{2}A^2\theta^2\sigma^2.$$

The solution is

$$\theta^* = \frac{\mu}{A\sigma^2},$$

independent of the initial level of wealth.

This framework is very tractable:

- It is easy to add multiple assets.
- It is easy to handle additive background risk, arising from random income or nontradable assets. Background risk does not affect the demand for tradable risky assets if it is uncorrelated with their returns.
- Equilibrium with heterogeneous agents is easy to calculate because the wealth distribution does not affect the demand for risky assets.

However there are also serious problems with this framework:

- Wealth does not affect the amount invested in a risky asset.
- Growth in consumption and wealth with multiplicative risks implies increasing absolute risks. CARA implies that this should generate an upward trend in risk premia which we have not seen historically.
- The assumption of normality cannot hold over more than one time interval. The compounding of returns over many periods converts a symmetric normal distribution into a right-skewed, non-normal distribution.

Portfolio choice in the CRRA-lognormal case

Assume that asset returns are lognormally distributed, and utility is power with relative risk aversion γ. The maximization problem is

$$\max \mathrm{E}_t W_{t+1}^{1-\gamma}/(1-\gamma).$$

Maximizing this expectation is equivalent to maximizing the log of the expectation. If $\gamma < 1$, then the scale factor $1/(1-\gamma) > 0$ and it can be omitted since it does not affect the solution. If $\gamma > 1$, then $1/(1-\gamma) < 0$ and we turn the problem into a minimization problem which turns out to have the same solution. Proceeding with the $\gamma < 1$ case, if next-period wealth is lognormal, we can rewrite the problem as

$$\max \log \mathrm{E}_t W_{t+1}^{1-\gamma} = (1-\gamma)\mathrm{E}_t w_{t+1} + \frac{1}{2}(1-\gamma)^2 \sigma_{wt}^2,$$

where $w_t = \log(W_t)$.

The budget constraint is

$$w_{t+1} = r_{p,t+1} + w_t,$$

where $r_{p,t+1} = \log(1 + R_{p,t+1})$ is the log return on the portfolio. So we can restate the problem as

$$\max \mathrm{E}_t r_{p,t+1} + \frac{1}{2}(1-\gamma)\sigma_{pt}^2,$$

where σ_{pt}^2 is the conditional variance of the log portfolio return.

To understand this, note that

$$\mathrm{E}_t r_{p,t+1} + \sigma_{pt}^2/2 = \log \mathrm{E}_t(1 + R_{p,t+1}).$$

Thus we can equivalently write

$$\max \log \mathrm{E}_t(1 + R_{p,t+1}) - \frac{\gamma}{2}\sigma_{pt}^2.$$

The investor trades off mean against variance in the portfolio return. The relevant mean return is the mean simple return, or arithmetic mean return, and the investor trades the log of this mean linearly against the variance of the log return.

When $\gamma = 1$, the investor has log utility and chooses the *growth-optimal portfolio* with the maximum log return. When $\gamma > 1$, the investor seeks a safer portfolio by penalizing the variance of log returns; when $\gamma < 1$, the investor actually seeks a riskier portfolio because a higher variance, with the same mean log return, corresponds to a higher mean simple return. The case $\gamma = 1$ is the boundary where these two opposing considerations cancel.

The growth-optimal portfolio has the property that as the investment horizon increases, it outperforms any other portfolio with increasing probability. To see this, note that the difference between the log return on the growth-optimal portfolio and the log return on any other portfolio is normally distributed with a positive mean. Assume that returns are iid over time. Then as the horizon increases, the mean and variance of the excess log return both grow linearly which means that the ratio of mean to standard deviation grows with the square root of the horizon. This ratio determines the probability that the excess return is positive, which therefore increases with the investment horizon. Markowitz and others have used this to argue that long-term investors should have log utility, but this claim has been strongly opposed by Samuelson and others.

Now we need to relate the log portfolio return to the log returns on underlying assets. The simple return on the portfolio is a linear combination of the simple returns on the risky and riskless assets. The log return on the portfolio is the log of this linear combination, which is not the same as a linear combination of logs.

Over short time intervals, however, we can use a Taylor approximation of the nonlinear function relating log individual-asset returns to log portfolio returns:

$$r_{p,t+1} - r_{f,t+1} \approx \alpha_t(r_{t+1} - r_{f,t+1}) + \frac{1}{2}\alpha_t(1-\alpha_t)\sigma_t^2.$$

The difference between the log portfolio return and a linear combination of log individual-asset returns is given by $\alpha_t(1-\alpha_t)\sigma_t^2/2$ which is zero if $\alpha_t = 0$ or 1. When $0 < \alpha_t < 1$, the portfolio is a weighted average of the individual assets and the term $\alpha_t(1-\alpha_t)\sigma_t^2/2$ is positive because the log of an average is greater than an average of logs..

Another way to understand this is to rewrite the equation as

$$r_{p,t+1} - r_{f,t+1} + \frac{\sigma_{pt}^2}{2} \approx \alpha_t\left(r_{t+1} - r_{f,t+1} + \frac{\sigma_t^2}{2}\right),$$

using the fact that $\sigma_{pt}^2 = \alpha_t^2\sigma_t^2$. This shows that the mean of the simple excess portfolio return is linearly related to the mean of the simple excess return on the risky asset.

Properties of this approximation:

- It becomes more accurate as the time interval shrinks. It is exact in continuous time with continuous paths for asset prices (then it follows from Itô's Lemma.).

- It rules out bankruptcy, even with a short position ($\alpha_t < 0$) or leverage ($\alpha_t > 1$).

With two assets, the mean excess portfolio return is $\mathrm{E}_t r_{p,t+1} - r_{f,t+1} = \alpha_t(\mathrm{E}_t r_{t+1} - r_{f,t+1}) + \frac{1}{2}\alpha_t(1-\alpha_t)\sigma_t^2$, while the variance of the portfolio return is $\alpha_t^2\sigma_t^2$. Substituting into the objective function, we get

$$\max \alpha_t(\mathrm{E}_t r_{t+1} - r_{f,t+1}) + \frac{1}{2}\alpha_t(1-\alpha_t)\sigma_t^2 + \frac{1}{2}(1-\gamma)\alpha_t^2\sigma_t^2.$$

The solution is

$$\alpha_t = \frac{\mathrm{E}_t r_{t+1} - r_{f,t+1} + \sigma_t^2/2}{\gamma\sigma_t^2},$$

which is an exact version of the result for a small risk.

Mean-variance analysis with two risky assets

Mean-variance analysis judges portfolios by their first two moments of returns. In a static (single-period) model, this can be justified by

- Quadratic utility of wealth, or
- Return distributions for which the first two moments are sufficient statistics. E.g.
  - Normal distribution
  - Lognormal distribution (with short time intervals so that portfolio returns and individual asset returns can both be lognormal)
  - Multivariate t distribution
  - Any of the above, plus an arbitrarily distributed common risk that does not affect portfolio choice.

Along with assumptions on return distributions, we may also use utility assumptions such as CARA with normal or CRRA with lognormal to get tractable closed-form portfolio rules. But these utility assumptions are not needed to justify mean-variance analysis.

We shall assume that short sales are permitted. Short sales restrictions introduce additional constraints that destroy the analytical simplicity of the basic mean-variance analysis.

With two risky assets, what combinations of mean and variance (or standard deviation) can you get? Start with the choice between two risky assets with returns R_1 and R_2.

$$R_p = w_1 R_1 + w_2 R_2.$$

$$\overline{R}_p = w_1 \overline{R}_1 + w_2 \overline{R}_2.$$

$$\begin{aligned}
\sigma_p^2 &= \mathrm{Var}(w_1 R_1 + w_2 R_2) \\
&= w_1^2 \mathrm{Var}(R_1) + w_2^2 \mathrm{Var}(R_2) + 2 w_1 w_2 \mathrm{Cov}(R_1, R_2) \\
&= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2 w_1 w_2 \sigma_{12} \\
&= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2 w_1 w_2 \sigma_1 \sigma_2 \rho_{12} \ ,
\end{aligned}$$

where $\rho_{12} \equiv \mathrm{Corr}(R_1, R_2)$.

With only two assets, we have $w_2 = 1 - w_1$ and

$$\overline{R}_p = w_1 \overline{R}_1 + (1 - w_1) \overline{R}_2 = \overline{R}_2 + w_1 (\overline{R}_1 - \overline{R}_2) \ .$$

The mean portfolio return is a linear function of w_1. Hence a target mean return determines w_1. Given a target $\overline{R}_p$,

$$w_1 = \frac{\overline{R}_p - \overline{R}_2}{\overline{R}_1 - \overline{R}_2} \ .$$

The variance of the portfolio return is

$$\sigma_p^2 = w_1^2 \sigma_1^2 + (1 - w_1)^2 \sigma_2^2 + 2 w_1 (1 - w_1) \sigma_1 \sigma_2 \rho_{12} \ .$$

This is a quadratic function of w_1, and hence of $\overline{R}_p$. A plot of mean against variance is a parabola; a plot of mean against standard deviation is a hyperbola.

When the two assets are perfectly (positively or negatively) correlated, then one can complete the square. In this case the standard deviation of the portfolio return is linear in the portfolio weight, and one can set it to zero by choosing the appropriate weight. In between these extreme correlation values, $\sigma_p^2 > 0$ for $-1 < \rho_{12} < 1$. Since 1 is the largest possible correlation,

$$\sigma_p^2 \leq (w_1\sigma_1 + (1-w_1)\sigma_2)^2 \quad \text{if } 0 < w_1 < 1 .$$

This illustrates the power of diversification.

Also,

$$\begin{aligned} \frac{d\sigma_p^2}{dw_1} &= 2w_1\sigma_1^2 - 2(1-w_1)\sigma_2^2 + 2(1-2w_1)\sigma_{12} \\ &= 2w_1[\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}] - 2[\sigma_2^2 - \sigma_{12}] . \end{aligned}$$

This derivative is increasing in w_1.

To find the *global minimum-variance portfolio* with the smallest possible variance, set the derivative to zero to get

$$w_{G1} = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}} , \qquad w_{G2} = \frac{\sigma_1^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}} .$$

Special cases:

1. When $\sigma_{12} = 0$,

$$w_{G1} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} , \qquad w_{G2} = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} .$$

2. When $\sigma_1^2 = \sigma_2^2$,

$$w_{G1} = \frac{1}{2}\,, \qquad w_{G2} = \frac{1}{2}\,.$$

3. When $\sigma_{12} = \sigma_1\sigma_2$ (so $\rho_{12} = 1$), the portfolio variance can be set to zero by

$$w_{G1} = \frac{-\sigma_2}{\sigma_1 - \sigma_2}\,, \qquad w_{G2} = \frac{\sigma_1}{\sigma_1 - \sigma_2}\,.$$

4. When $\sigma_{12} = -\sigma_1\sigma_2$ (so $\rho_{12} = -1$), the portfolio variance can be set to zero by

$$w_{G1} = \frac{\sigma_2}{\sigma_1 + \sigma_2}\,, \qquad w_{G2} = \frac{\sigma_1}{\sigma_1 + \sigma_2}\,.$$

Mean-variance analysis with one risky and one safe asset

In this case $\sigma_2^2 = 0$ and $R_2 = R_f$, the riskless interest rate. We have $\overline{R}_p - R_f = w_1(\overline{R}_1 - R_f)$ and $\sigma_p^2 = w_1^2\sigma_1^2$ or $w_1 = \sigma_p/\sigma_1$. Hence

$$\overline{R}_p - R_f = \sigma_p\left(\frac{\overline{R}_1 - R_f}{\sigma_1}\right).$$

This defines a straight line, called the *capital allocation line (CAL)*, on a mean-standard deviation diagram. The slope

$$S_1 = \left(\frac{\overline{R}_1 - R_f}{\sigma_1}\right)$$

is called the *Sharpe ratio* of the risky asset. Any portfolio that combines a single risky asset with the riskless asset has the same Sharpe ratio as the risky asset itself.

The standard rule of myopic portfolio choice is

$$w_1 = \frac{\overline{R}_1 - R_f}{RRA\sigma_1^2} = \frac{S_1}{RRA\sigma_1}.$$

The optimal share in the risky asset is the risk premium divided by risk aversion times variance, or equivalently the Sharpe ratio of the risky asset divided by the coefficient of relative risk aversion times standard deviation.

Mean-variance analysis with N risky assets

In the 2-asset case, the mean return target uniquely defines the portfolio weights. This is no longer true when we have N assets. Now the problem is to find portfolios that have minimum variance for a given mean return. These are called "minimum-variance" portfolios and we say they lie on the "minimum-variance frontier".

Define $\overline{R}$ as the vector of mean returns, Σ as the variance-covariance matrix of returns, w as the vector of portfolio weights, and ι as a vector of ones. Write the problem as

$$\min_w \frac{1}{2} w' \Sigma w \text{ s.t.}$$
$$\begin{aligned} \overline{R}'w &= \overline{R}_p \\ \iota' w &= 1 \,. \end{aligned}$$

Set up the Lagrangian

$$\mathcal{L}(w, \lambda_1, \lambda_2) = \frac{1}{2} w' \Sigma w + \lambda_1(\overline{R}_p - \overline{R}'w) + \lambda_2(1 - \iota' w)$$

and get first-order conditions

$$\Sigma w = \lambda_1 \overline{R} + \lambda_2 \iota \,.$$

Premultiply both sides by Σ^{-1} to get:

$$w = \lambda_1 \Sigma^{-1} \overline{R} + \lambda_2 \Sigma^{-1} \iota \ .$$

λ_1 and λ_2 can be found by using the two constraints

$$\begin{aligned} \overline{R}_p &= \overline{R}' w = \lambda_1 \overline{R}' \Sigma^{-1} \overline{R} + \lambda_2 \overline{R}' \Sigma^{-1} \iota = \lambda_1 A + \lambda_2 B \\ 1 &= \iota' w = \lambda_1 \iota' \Sigma^{-1} \overline{R} + \lambda_2 \iota' \Sigma^{-1} \iota = \lambda_1 B + \lambda_2 C \ , \end{aligned}$$

where $A \equiv \overline{R}' \Sigma^{-1} \overline{R}$, $B \equiv \overline{R}' \Sigma^{-1} \iota = \iota' \Sigma^{-1} \overline{R}$, and $C \equiv \iota' \Sigma^{-1} \iota$. Note that A and C are mathematically guaranteed to be positive; B is not, but for economic reasons we normally expect it to be positive (see discussion of the global minimum-variance portfolio below).

Solving these equations, we get

$$\lambda_1 = \frac{C\overline{R}_p - B}{D} \ , \qquad \lambda_2 = \frac{A - B\overline{R}_p}{D} \ ,$$

where $D \equiv AC - B^2$.

The minimized variance is

$$\begin{aligned} \sigma_p^2 &= w' \Sigma w = w' \Sigma (\lambda_1 \Sigma^{-1} \overline{R} + \lambda_2 \Sigma^{-1} \iota) \\ &= \lambda_1 w' \overline{R} + \lambda_2 w' \iota = \lambda_1 \overline{R}_p + \lambda_2 \\ &= \frac{A - 2B\overline{R}_p + C\overline{R}_p^2}{D}. \end{aligned}$$

Thus $d\sigma_p^2 / d\overline{R}_p = \lambda_1$. λ_1 measures the variance cost of a higher mean return target, and it is increasing in $\overline{R}_p$.

Now look at the *global minimum-variance portfolio*. The first constraint (on the mean) is dropped from the problem; this is equivalent to setting $\lambda_1 = 0$. We get

$$w_G = \lambda_2 \Sigma^{-1} \iota \ ,$$

$$1 = \iota' w_G = \lambda_2 \iota' \Sigma^{-1} \iota \ .$$

So $\lambda_2 = 1/(\iota'\Sigma^{-1}\iota) = 1/C$, and

$$w_G = \frac{\Sigma^{-1}\iota}{\iota'\Sigma^{-1}\iota} \ .$$

In the 2×2 case,

$$w_{G1} = \frac{\text{sum of top row of } \Sigma^{-1}}{\text{sum of all elements of } \Sigma^{-1}} \ ,$$

$$w_{G2} = \frac{\text{sum of bottom row of } \Sigma^{-1}}{\text{sum of all elements of } \Sigma^{-1}} \ ,$$

and we can use the standard formula for the inverse of a 2×2 matrix,

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix} ,$$

to show that the portfolio weights of the global minimum-variance portfolio are the same that we got before.

The mean return on the global minimum-variance portfolio is

$$w_G'\overline{R} = \left(\frac{\iota'\Sigma^{-1}\overline{R}}{\iota'\Sigma^{-1}\iota} \right) = \frac{B}{C}.$$

We expect the mean return on the global minimum-variance portfolio to be positive, and thus we expect B to be positive.

In the general model with an arbitrary mean return constraint, we can verify that when $\overline{R}_p > B/C$, then the Lagrange multiplier for the mean constraint, $\lambda_1 > 0$.

The set of minimum-variance portfolios that satisfy this condition is called the *mean-variance efficient set.*

The variance of the return on the global minimum-variance portfolio is

$$w_G' \Sigma w_G = \frac{\iota' \Sigma^{-1} \Sigma \Sigma^{-1} \iota}{(\iota' \Sigma^{-1} \iota)^2} = \frac{1}{(\iota' \Sigma^{-1} \iota)}.$$

This simplifies in the case where all assets are symmetrical, having the *same* variance and the *same* correlation ρ with each other. Then the global minimum-variance portfolio is equally weighted, $w_G = \iota / N$, and

$$\begin{aligned} w_G' \Sigma w_G &= \frac{\iota' \Sigma \iota}{N^2} = \frac{N^2 \rho \sigma^2}{N^2} + \frac{N(1-\rho)\sigma^2}{N^2} \\ &= \rho \sigma^2 + \frac{(1-\rho)\sigma^2}{N}. \end{aligned}$$

When assets are uncorrelated with one another, $\rho = 0$, this further simplifies to

$$w_G' \Sigma w_G = \frac{\sigma^2}{N}.$$

The mutual fund theorem

The mutual fund theorem says that all minimum-variance portfolios can be obtained by mixing just two minimum-variance portfolios in different proportions. Thus if all investors hold minimum-variance portfolios, all investors hold combinations of just two underlying portfolios or "mutual funds".

To show this, rewrite the equation for portfolio weights as

$$\begin{aligned} w &= \lambda_1 \iota' \Sigma^{-1} \overline{R} \left(\frac{\Sigma^{-1} \overline{R}}{\iota' \Sigma^{-1} \overline{R}} \right) + \lambda_2 \iota' \Sigma^{-1} \iota \left(\frac{\Sigma^{-1} \iota}{\iota' \Sigma^{-1} \iota} \right) \\ &= \lambda_1 B \left(\frac{\Sigma^{-1} \overline{R}}{\iota' \Sigma^{-1} \overline{R}} \right) + \lambda_2 C w_G, \end{aligned}$$

and note that $\lambda_1 B + \lambda_2 C = 1$. Thus the optimal portfolio is a combination of two portfolios, the second of which is the global minimum-variance portfolio, and the first of which invests more heavily in assets with high mean returns.

Mean-variance analysis with a riskless asset and N risky assets.

Write the riskless asset return as R_f. Rewrite the problem as one of choosing weights w in the risky assets, where the portfolio is completed by lending or borrowing at the riskless rate R_f. Thus we no longer require $\iota' w = 1$. Drop this constraint and write the problem as

$$\min_w \frac{1}{2} w' \Sigma w \quad \text{s.t.} \quad (\overline{R} - R_f \iota)' w = (\overline{R}_p - R_f) \ .$$

Set up the Lagrangian

$$\mathcal{L}(w_1, w_2, \lambda_1) = \frac{1}{2}(w' \Sigma w) + \lambda_1 (\overline{R}_p - R_f - (\overline{R} - R_f \iota)' w) \ ,$$

and get first-order conditions

$$\frac{\partial \mathcal{L}}{\partial w} = \Sigma w - \lambda_1 (\overline{R} - R_f \iota) = 0$$

so

$$w = \lambda_1 \Sigma^{-1} (\overline{R} - R_f \iota) \ .$$

$$\overline{R}_p - R_f = (\overline{R} - R_f \iota)' w = \lambda_1 (\overline{R} - R_f \iota)' \Sigma^{-1} (\overline{R} - R_f \iota) = \lambda_1 E \ ,$$

where $E \equiv (\overline{R} - R_f \iota)' \Sigma^{-1} (\overline{R} - R_f \iota)$. Thus

$$\lambda_1 = \frac{\overline{R}_p - R_f}{E} \ .$$

Also,

$$\sigma_p^2 = w'\Sigma w = \lambda_1^2(\overline{R} - R_f\iota)'\Sigma^{-1}\Sigma\Sigma^{-1}(\overline{R} - R_f\iota) = \lambda_1^2 E \ .$$

Thus

$$\sigma_p^2 = \frac{(\overline{R}_p - R_f)^2}{E} \ ,$$

and

$$| \ \overline{R}_p - R_f \ | = \sqrt{E}\sigma_p \ .$$

This gives a straight line on a plot of mean against standard deviation. The straight line is a tangency line from the riskless asset return on the vertical axis, to the minimum-variance set obtainable if one can only hold the risky assets and not the riskless asset. The slope of this line, $\sqrt{E}$, is the Sharpe ratio of the tangency portfolio. The tangency portfolio has the highest Sharpe ratio of any risky asset or portfolio of risky assets.

In the presence of a riskless asset, the mutual fund theorem says that all investors, regardless of their risk aversion, should hold risky assets in the same proportion. This contradicts conventional investment advice, and this contradiction has been called the "asset allocation puzzle" by Canner, Mankiw, and Weil (1997).

Practical difficulties with mean-variance analysis

Is mean-variance analysis a practical solution to the problem of portfolio choice? Perhaps not:

- Estimates of means are imprecise over short periods.
- Means may not be constant over long periods.
- The variance-covariance matrix Σ has $N(N+1)/2$ variances and covariances that have to be estimated. This can be a very large number!

- If $N \geq T$, the historical variance-covariance matrix is always singular. This means that it cannot be inverted. There will appear to be riskless combinations of risky assets in the data.
- Even if $N < T$, if N is large the data will suggest that some combinations of risky assets are almost riskless. This can lead to a highly leveraged portfolio.

These difficulties have motivated a search for shortcut methods to find optimal portfolios:

- Capital Asset Pricing Model (CAPM).
- Multifactor models.

These models also have broader implications:

- Testable restrictions on asset returns.
- Capital budgeting (what discount rate to use in evaluating investment projects).
- Mutual fund performance evaluation (how large a return should one expect given the risk that a fund manager is taking).

The Capital Asset Pricing Model (CAPM)

Assumptions of the CAPM:

- All investors are price-takers.
- All investors care about returns measured over one period.
- There are no nontraded assets.
- Investors can borrow or lend at a given riskfree interest rate (Sharpe-Lintner version of the CAPM - this is relaxed in the Black version).

- Investors pay no taxes or transaction costs.
- All investors are mean-variance optimizers.
- All investors perceive the same means, variances, and covariances for returns.

These assumptions imply that:

- All investors work with the same mean-standard deviation diagram.
- All investors hold a mean-variance efficient portfolio.
- Since all mean-variance efficient portfolios combine the riskless asset with a fixed portfolio of risky assets, all investors hold risky assets in the same proportions to one another.
- These proportions must be those of the *market portfolio* or *value-weighted index* that contains all risky assets in proportion to their market value.
- Thus *the market portfolio is mean-variance efficient.*

The implication for portfolio choice is that a mean-variance investor need not actually perform the mean-variance analysis! The investor can "free-ride" on the analyses of other investors, and use the market portfolio (in practice, a broad index fund) as the optimal mutual fund of risky assets (tangency portfolio). The optimal capital allocation line (CAL) is just the *capital market line (CML)* connecting the riskfree asset to the market portfolio.

Asset pricing implications of the CAPM

We begin by looking at *covariance properties of efficient portfolios.*

An increase in portfolio weight w_i, financed by a decrease in the weight on the riskless asset, affects the mean and variance of the portfolio return as follows:

$$\frac{d\overline{R}_p}{dw_i} = \overline{R}_i - R_f \ .$$

$$\frac{d\text{Var}(R_p)}{dw_i} = 2\text{Cov}(R_i, R_p),$$

because the terms in $\text{Var}(R_p)$ that involve w_i are

$$2w_1 w_i \text{Cov}(R_1, R_i) + \ldots + w_i^2 \text{Var}(R_i) + \ldots + 2w_N w_i \text{Cov}(R_1, R_N),$$

and so the derivative

$$\begin{aligned} \frac{d\text{Var}(R_p)}{dw_i} &= 2w_1 \text{Cov}(R_1, R_i) + \ldots + 2w_i \text{Var}(R_i) \\ &\quad + \ldots + 2w_N \text{Cov}(R_1, R_N) = 2\text{Cov}(R_i, R_p). \end{aligned}$$

The ratio of the effects on mean and on variance is

$$\frac{d\overline{R}_p/dw_i}{d\text{Var}(R_p)/dw_i} = \frac{\overline{R}_i - R_f}{2\text{Cov}(R_i, R_p)} \ .$$

If portfolio p is efficient, this ratio should be the same for all assets. Why? Consider adjusting two different portfolio weights, w_i and w_j. The effects on the mean and variance of R_p are

$$d\overline{R}_p = (\overline{R}_i - R_f)dw_i + (\overline{R}_j - R_f)dw_j.$$

$$d\text{Var}(R_p) = 2\text{Cov}(R_i, R_p)dw_i + 2\text{Cov}(R_j, R_p)dw_j.$$

Now consider setting dw_j so that the mean portfolio return is unchanged, $d\overline{R}_p = 0$:

$$dw_j = -\frac{(\overline{R}_i - R_f)}{(\overline{R}_j - R_f)} dw_i.$$

Then the portfolio variance must also be unchanged, because otherwise one could achieve a lower variance with the same mean, which would contradict the assumption that the portfolio is efficient. We have

$$d\text{Var}(R_p) = \left[2\text{Cov}(R_i, R_p) - 2\text{Cov}(R_j, R_p) \frac{(\overline{R}_i - R_f)}{(\overline{R}_j - R_f)} \right] dw_i = 0.$$

This requires

$$\frac{\overline{R}_i - R_f}{2\text{Cov}(R_i, R_p)} = \frac{\overline{R}_j - R_f}{2\text{Cov}(R_j, R_p)}.$$

This equation must hold for all assets j, including the original portfolio itself. Setting $j = p$, we get

$$\frac{\overline{R}_i - R_f}{2\text{Cov}(R_i, R_p)} = \frac{\overline{R}_p - R_f}{2\text{Var}(R_p)} ,$$

$$\overline{R}_i - R_f = \frac{\text{Cov}(R_i, R_p)}{\text{Var}(R_p)} (\overline{R}_p - R_f) = \beta_{ip}(\overline{R}_p - R_f),$$

where $\beta_{ip} \equiv \text{Cov}(R_i, R_p)/\text{Var}(R_p)$ is the regression coefficient of asset i on portfolio p.

Since the CAPM implies that the market portfolio m is efficient, this equation describes the market portfolio:

$$\overline{R}_i - R_f = \beta_{im}(\overline{R}_m - R_f) ,$$

where $\beta_{im} \equiv \text{Cov}(R_i, R_m)/\text{Var}(R_m)$ is the regression coefficient of asset i on the market portfolio m. β_{im} can also be interpreted as the weight on the market of the portfolio of market and riskless asset that most closely replicates the return on asset i.

Thus, if we consider the regression of excess returns on the market excess return,

$$R_{it} - R_{ft} = \alpha_i + \beta_{im}(R_{mt} - R_{ft}) + \varepsilon_{it},$$

where $\alpha_i \equiv \overline{R}_i - R_f - \beta_{im}(\overline{R}_m - R_f)$, α_i should be zero for all assets. α_i is called *Jensen's alpha* and is used to try to find assets that are mispriced relative to the CAPM. The relationship

$$\overline{R}_i = R_f + \beta_{im}(\overline{R}_m - R_f)$$

is called the *security market line (SML)*, and α_i measures deviations from this line.

What if we do not allow riskless borrowing and lending? Then all investors choose combinations of the same two mutual funds. The market portfolio must be a combination of these mutual funds, and must therefore be efficient. We still have that the market portfolio is efficient. Our analysis of the covariance properties goes through as before, except that we replace the riskless asset with an efficient portfolio z that is uncorrelated with the market portfolio. We get

$$\overline{R}_i - \overline{R}_z = \beta_{im}(\overline{R}_m - \overline{R}_z) \ ,$$

where β_{im} is defined as before. This version of the CAPM is due to Fischer Black.

Testing the CAPM with a group of assets

In practice the CAPM will never hold exactly. We need a statistical test to tell whether deviations from the model (mean-variance inefficiency of the market portfolio, or equivalently nonzero alphas) are statistically significant.

The two leading approaches are time-series and cross-sectional. At a deep level, they are much more similar than they appear to be at first.

The *time-series approach* starts from the regression

$$R^e_{it} = \alpha_i + \beta_{im} R^e_{mt} + \varepsilon_{it},$$

where $R^e_{it} = R_{it} - R_{ft}$ and $R^e_{mt} = R_{mt} - R_{ft}$. The null hypothesis is that $\alpha_i = 0$. This is a simple parameter restriction for any one asset; the challenge is to test it jointly for a set of N assets.

An asymptotic test is as follows. Define α as the N-vector of intercepts α_i, and Σ as the variance-covariance matrix of the regression residuals ε_{it}. (Note that this is different than the notation we used to do mean-variance analysis, where Σ was the variance-covariance matrix of the raw returns rather than the residuals.) Then as the sample size T increases, asymptotically

$$T \left[1 + \left(\frac{\overline{R}^e_m}{\sigma(R^e_{mt})} \right)^2 \right] \widehat{\alpha}' \widehat{\Sigma}^{-1} \widehat{\alpha} \sim \chi^2_N.$$

To see the intuition, suppose there were no market return in the model. Then the vector α would be a vector of sample mean excess returns, with variance-covariance matrix $(1/T)\Sigma$. Thus the quadratic form $\widehat{\alpha}' \widehat{\Sigma}^{-1} \widehat{\alpha}$ is a sum of squared intercepts, divided by its variance-covariance matrix, which has a χ^2_N distribution. The term in square brackets is a correction for the presence of the market return in the model. Uncertainty about the betas affects the alphas, and more so when the market has a high expected return relative to its variance.

A finite-sample test makes a further correction for the fact that the variance-covariance matrix Σ must be estimated. Under the assumption that the ε_{it} are serially uncorrelated, homoskedastic, and normal, we have

$$\left(\frac{T-N-1}{N} \right) \left[1 + \left(\frac{\overline{R}^e_m}{\sigma(R^e_{mt})} \right)^2 \right] \widehat{\alpha}' \widehat{\Sigma}^{-1} \widehat{\alpha} \sim F_{N,T-N-1}.$$

Gibbons, Ross, and Shanken (1989) give a nice geometric interpretation that relates the test statistic to mean-variance analysis. Define $\hat{S}^2_m$ as the estimated squared Sharpe ratio of the market. Define $\hat{S}^2_q$ as the squared Sharpe ratio of the estimated tangency portfolio (the highest squared Sharpe ratio available from the set of test assets together with the market). GRS show that the previous test statistic is equivalent to

$$\left(\frac{T-N-1}{N}\right)\left(\frac{\hat{S}_q^2-\hat{S}_m^2}{1+\hat{S}_m^2}\right).$$

Thus we are measuring how far the market is from the ex-post tangency portfolio in a mean-standard deviation diagram.

The *cross-sectional approach* first estimates betas from a time-series regression, then runs a cross-sectional regression

$$\overline{R}_i^e = \lambda\beta_{im} + \alpha_i,$$

where there is no intercept in the regression, λ is the cross-sectional reward for bearing market risk, and the alphas are now regression residuals. (We can alternatively include an intercept in the regression and test that it is zero, or allow it to be free in the Black version of the CAPM.) We are in the strange position of wanting to test that the true variation in the regression residuals is zero, that is that the regression has a perfect fit. We can do this only because the time-series regressions tell us something about the variability in the average returns.

One problem with the cross-sectional regression is that the residuals will not generally be uncorrelated with one another. (This makes it difficult to interpret plots of cross-sectional regressions, since many of the points may be generated by highly correlated underlying assets that are bound to plot near one another.) As before, we know that $\mathrm{E}(\alpha\alpha') = (1/T)\Sigma$. To correct for this, we can run Generalized Least Squares and get

$$\widehat{\lambda}_{GLS} = (\beta'\Sigma^{-1}\beta)^{-1}\beta'\Sigma^{-1}\overline{R}^e,$$

$$\widehat{\alpha}_{GLS} = \overline{R}^e - \widehat{\lambda}_{GLS}\beta.$$

An asymptotic test statistic based on the GLS cross-sectional regression, and correcting for the fact that the betas are not known but estimated from prior time-series regressions, is

$$T\left[1+\left(\frac{\widehat{\lambda}_{GLS}}{\sigma(R_{mt}^e)}\right)^2\right]\widehat{\alpha}_{GLS}'\widehat{\Sigma}^{-1}\widehat{\alpha}_{GLS} \sim \chi_{N-1}^2.$$

This is almost exactly the same as the asymptotic time-series test statistic, but we have lost one degree of freedom by estimating the reward for beta in the cross-section rather than from the average excess market return. We can earn that degree of freedom back by adding the market to the set of assets in the cross-sectional GLS regression. Then the regression puts all the weight on the market in estimating the reward for market risk (since other assets are just the market plus noise, so the GLS regression knows they are less informative about this parameter). The result is a test statistic that is exactly the same as the time-series test statistic.

The cross-sectional approach has the advantage that it can be implemented even when the factor is not the return on a traded portfolio. In that case we need to use the cross-section to estimate the reward for bearing factor risk.

A variant of the cross-sectional regression approach is the *Fama-MacBeth* approach. This first estimates betas, then runs a series of period-by-period cross-sectional regressions,

$$R_{it}^e = \lambda_t \beta_{im} + \alpha_{it}.$$

Here the observations in each regression run from $i = 1...N$, and the regressions are run separately for each $t = 1...T$. The coefficients and residuals are then averaged over time to estimate the average reward for beta exposure $\widehat{\lambda}$ and the average alphas $\widehat{\alpha}_i$. We use the variability of the coefficients and residuals over time to estimate the standard errors of these averages and construct test statistics for the model.

When the explanatory variables in the regression do not vary over time, the Fama-MacBeth approach is equivalent to cross-sectional OLS using the entire sample average, or to a pooled time-series cross-sectional OLS regression, with standard errors corrected for cross-sectional correlation of residuals. When the explanatory variables do vary over time, Fama-MacBeth is different because it gives equal weight to each time period, even if the explanatory variables are more dispersed in one period than another.

The basic Fama-MacBeth approach does not adjust for the fact that betas are not known but must be estimated from time-series regressions. However it does easily allow for changing betas over time. Thus it is more appropriate as a method to estimate the rewards to observable characteristics of firms (which could include their lagged historical betas), than as a method to test the CAPM.

All of these test approaches have important common features:

- Looking at more assets (increasing N) will tend to find larger deviations from the model.
- But increasing N also increases the size of the deviations you *need* to find to reject the model statistically.
- Thus, to get a powerful test you need to group assets into a few portfolios that summarize the behavior of a larger set of assets.
- However it is cheating to do this after looking at the average returns on the full set of assets and picking portfolios based on this information. This *data-snooping* leads to spurious rejections of the model.
- Much of the debate about the empirical validity of the CAPM centers on this issue.

Empirical evidence on the CAPM

- Early work looked at stocks grouped by their estimated betas. High-beta stocks do tend to have higher returns than low-beta stocks, but the relationship is weak and the reward for beta is lower than the equity premium.
- Small stocks tend to have higher returns than large stocks. Although small stocks also have higher betas, the return difference is too high to be consistent with the CAPM.
- In fact, the size effect seems to explain a large part of the high average returns of high-beta stocks. If you group stocks by both size and beta, size predicts returns and beta does not.
- However the size effect has been very weak since the early 1980's when it was first documented.
- Stocks with high book-market, dividend-price, or earnings-price ratios ("value" stocks) have high returns. In recent decades this is a striking violation of the CAPM because these stocks have relatively low betas. (But this was not true in the Great Depression or early postwar period, when value stocks had relatively high betas.)

- Stocks that have done well over the past year ("momentum" stocks) also have high returns and relatively low betas.

- Initial public offerings (IPO's) do poorly relative to the CAPM prediction. This may be another manifestation of the value effect, since IPO's tend to have extremely low book-market ratios.

Alternative reactions to these results:

- The anomalies are just the result of data-snooping.

- The CAPM is not testable because we do not know the composition of the true market portfolio, and so we use a broad stock index as an imperfect proxy. A rejection of the CAPM just tells us that our proxy is inadequate. This is the *Roll critique* of Roll (1977). It has particular force when we consider nontradable assets, most importantly human capital.

- The anomalies are concentrated in small, illiquid stocks that may have high average returns to compensate for their illiquidity.

- The CAPM test assumes that betas and the market risk premium are constant over time. If this is not the case, then the CAPM could hold conditionally at each point in time, but could fail unconditionally. Cochrane calls this the *Hansen-Richard critique* after Hansen and Richard (1987).

- A more general model of risk and return is needed, for example an intertemporal CAPM of the sort proposed by Merton (1973).

- Markets are inefficient so no model of risk and return will fully resolve the anomalies.

The conditional CAPM vs. the unconditional CAPM

Suppose that the CAPM holds conditionally:

$$\mathrm{E}_t R^e_{i,t+1} = \beta_{imt} \mathrm{E}_t R^e_{m,t+1}.$$

Taking unconditional expectations,

$$\mathrm{E} R^e_{i,t+1} = \overline{\beta}_{im} \mathrm{E} R^e_{m,t+1} + \mathrm{Cov}(\beta_{imt}, E_t R^e_{m,t+1}).$$

An asset can have a higher unconditional average return than predicted by the unconditional CAPM, if its beta moves with the market risk premium.

One way to test a conditional model is to parameterize the variables that shift betas over time. For example, we might write

$$\beta_{imt} = \beta_{i0} + \beta_{i1} z_t.$$

The conditional model can then be written as

$$\mathrm{E}_t R^e_{i,t+1} = \beta_{i0} \mathrm{E}_t R^e_{m,t+1} + \beta_{i1} \mathrm{E}_t z_t R^e_{m,t+1}$$

and now we can take unconditional expectations to get

$$\mathrm{E} R^e_{i,t+1} = \beta_{i0} \mathrm{E} R^e_{m,t+1} + \beta_{i1} \mathrm{E} z_t R^e_{m,t+1}.$$

This is a multifactor model, where the factors are the market and the market scaled by the state variable z_t, and it can be tested in the usual way using time-series or cross-sectional regressions. If a cross-sectional regression is used, it is important to include the market itself in the set of test assets. This ensures that the cross-sectional estimate of the reward for the z_t-scaled market factor is reasonable, given the time-series behavior of the market return.

Empirically, it seems that value stocks have higher betas when the expected market return is high, but the effect is not large enough to explain much of the value effect. It also seems that the conditional approach makes a bigger difference in tests of the consumption CAPM than in tests of the traditional CAPM.

Arbitrage pricing in a single-factor model

Suppose that we run the regression

$$R_{it} - R_f = \alpha_i + \beta_{im}(R_{mt} - R_f) + \epsilon_{it}.$$

This relationship is called the *market model*. It is the leading example of a *single-factor model* with a single common factor moving stock returns.

Suppose that the errors in this equation are uncorrelated across stocks:

$$\mathrm{E}[\epsilon_{it}\epsilon_{jt}] = 0$$

for $i \neq j$. Then the residual risk in any stock is *idiosyncratic*, unrelated to the residual risk in any other stock.

Implications:

- Covariances are easy to estimate for mean-variance analysis because

$$\mathrm{Cov}(R_{it}, R_{jt}) = \beta_{im}\beta_{jm}\sigma_m^2.$$

- If many assets are available, we should expect α_i typically to be very small in absolute value.

To understand why α_i should be small in absolute value, consider forming a portfolio of N assets i. The portfolio return will be

$$R_{pt} - R_f = \alpha_p + \beta_{pm}(R_{mt} - R_f) + \epsilon_{pt},$$

where $\alpha_p = \sum_{j=1}^N w_j\alpha_j$, $\beta_{pm} = \sum_{j=1}^N w_j\beta_{jm}$, and $\epsilon_{pt} = \sum_{j=1}^N w_j\epsilon_{jt}$.

The variance of ϵ_{pt} will be

$$\text{Var}(\epsilon_{pt}) = \sum_{j=1}^{N} w_j^2 \text{Var}(\epsilon_{jt}),$$

which will shrink rapidly with N provided that no single weight w_j is too large. In the benchmark case where the portfolio is equally weighted ($w_j = 1/N$) and all the stocks have the same idiosyncratic variance ($\text{Var}(\epsilon_{jt}) = \sigma^2$), we get

$$\text{Var}(\epsilon_{pt}) = \frac{\sigma^2}{N}.$$

Now suppose that the portfolio has enough stocks, with a small enough weight in each one, that the residual risk $\text{Var}(\epsilon_{pt})$ is negligible. We say that the portfolio is *well diversified*. For such a portfolio, we can neglect ϵ_{pt} and write the return as

$$R_{pt} - R_f = \alpha_p + \beta_{pm}(R_{mt} - R_f).$$

But then we must have $\alpha_p = 0$. If not, there is an arbitrage opportunity: short β_{pm} units of the market, long one unit of the portfolio. This gives a riskless excess return of α_p.

The *arbitrage pricing theory* of Ross (1976) builds on this result. The idea is to show that $\alpha_p = 0$ for all well diversified portfolios implies that "almost all" individual assets have α_i very close to zero. Intuitively, a few assets can be mispriced (have nonzero α_i), but if too many assets are mispriced then one can group them into a well diversified portfolio and create an arbitrage opportunity. Technically, the result is that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \alpha_i^2 = 0.$$

To derive this result, consider forming a portfolio by investing

$$\frac{\alpha_i}{\sqrt{N \sum_{i=1}^{N} \alpha_i^2}}$$

in each asset $i = 1...N$. Because $\sum_{i=1}^{N} \alpha_i = 0$, this portfolio is an *arbitrage portfolio* whose initial dollar investment is zero. The return on the portfolio is therefore undefined, but we can calculate the payoff as

$$\frac{\sum_{i=1}^{N} \alpha_i R_{it}}{\sqrt{N \sum_{i=1}^{N} \alpha_i^2}}.$$

The expected payoff is

$$\frac{\sum_{i=1}^{N} \alpha_i^2}{\sqrt{N \sum_{i=1}^{N} \alpha_i^2}} = \sqrt{\frac{\sum_{i=1}^{N} \alpha_i^2}{N}},$$

because α_i is cross-sectionally uncorrelated with beta. For the same reason the variance of the payoff is

$$\frac{Var(\sum_{i=1}^{N} \alpha_i \varepsilon_{it})}{N \sum_{i=1}^{N} \alpha_i^2} = \frac{\sum_{i=1}^{N} \alpha_i^2 \sigma_i^2}{N \sum_{i=1}^{N} \alpha_i^2}.$$

If the idiosyncratic variances σ_i^2 are bounded, then the variance of the payoff shrinks to zero as $N \longrightarrow \infty$. To avoid an *asymptotic arbitrage opportunity*, the expected payoff must shrink to zero, which implies that the square of the expected payoff shrinks to zero. This is the desired result.

We have derived the beta pricing equation of the CAPM, without using any of the apparatus of mean-variance analysis. The key assumption is the assumption that residual risk from the market model is uncorrelated across stocks. An advantage of this approach is that the "market" in the model can be any broadly diversified portfolio that produces uncorrelated residual risk; we do not have to worry about measuring the true market portfolio of all wealth, and thus we sidestep the Roll critique.

In practice, however, no single factor accounts for all the correlations among individual stock returns; there are many other sources of common variation in stocks, besides the common influence of the market return. For example:

- Industry effects
- Cyclical stocks are particularly sensitive to changes in industrial production

- Highly leveraged firms are sensitive to changes in interest rates
- Small firms seem to move together.

Multifactor models

To handle this, we generalize to a multifactor model. If there are K portfolios capturing the common influence of K underlying sources of risk, then we have

$$R_{it} - R_f = \alpha_i + \sum_{k=1}^{K} \beta_{ik}(R_{kt} - R_f) + \epsilon_{it}.$$

We assume that the residual is uncorrelated across stocks. The prediction of the model is that $\alpha_i = 0$ for almost all stocks. This is restrictive if $K << N$.

Alternatively, if we measure the factors directly as mean-zero shocks (for example, innovations to macroeconomic variables), then we have

$$R_{it} - R_f = \mu_i + \sum_{k=1}^{K} \beta_{ik} F_{kt} + \epsilon_{it},$$

and the prediction of the model is that

$$\mu_i = \sum_{k=1}^{K} \beta_{ik} \lambda_k,$$

where λ_k is the *price of risk* of the k'th factor. (In the version of the model with traded portfolios as factors, the risk prices are captured by the mean returns on the traded portfolios, which is why we do not see them explicitly.)

The multifactor model can also be interpreted in terms of mean-variance analysis. It says that the full mean-variance-efficient frontier can be constructed from the K factor portfolios, so a mean-variance investor should always hold some combination of these portfolios. The dimension of mean-variance analysis is greatly reduced.

Ways to pick factors for multifactor models:

- The first factor is almost always chosen to be the return on a broad market index.
- Early research studied the covariance matrix of asset returns and tried to find the most important common factors. Unfortunately this approach tends to find more and more factors as you increase the number of assets.
- Macroeconomic factors (inflation, interest rates, industrial production).
- Portfolios of stocks with common characteristics (size, book-to-market, momentum). This leads to a data-snooping problem if the characteristics are chosen to produce high average returns in the sample.

The generality of arbitrage pricing theory is also a weakness.

- We know that some portfolio is always ex post mean-variance efficient. Thus we know we can always get a 1-factor model to fit the data. A fortiori, we can always get a K-factor model to fit the data. What does this tell us about the world?
- The theory does not determine the signs or magnitudes of the risk prices. Being a common factor is a necessary, but not a sufficient condition to be a priced factor that helps to determine the cross-section of asset returns. Much recent work on general equilibrium asset pricing aims to pin down the risk prices more precisely from theoretical considerations.