usually works quite well. As a result, for most applications it is reasonably safe to state that the OLS estimator is approximately normally distributed. More information about Monte Carlo experiments is provided in Davidson and MacKinnon (1993, Chapter 21), while a simple illustration is provided in Patterson (2000, Section 8.2).

2.7 Illustration: The Capital Asset Pricing Model

One of the most important models in finance is the Capital Asset Pricing Model (CAPM). The CAPM is an equilibrium model which assumes that all investors compose their asset portfolio on the basis of a trade-off between the expected return and the variance of the return on their portfolio. This implies that each investor holds a so-called **mean variance efficient** portfolio, a portfolio that gives maximum expected return for a given variance (level of risk). If all investors hold the same beliefs about expected returns and (co)variances of individual assets, and in the absence of transaction costs, taxes and trading restrictions of any kind, it is also the case that the aggregate of all individual portfolios, the **market portfolio**, is mean variance efficient. In this case, it can be shown that expected returns on individual assets are linearly related to the expected return on the market portfolio. In particular, in holds that[21]

$$E\{r_{jt} - r_f\} = \beta_j E\{r_{mt} - r_f\}, \tag{2.77}$$

where r_{jt} is the risky return on asset j in period t, r_{mt} the risky return on the market portfolio, and r_f denotes the riskless return, which we assume to be time-invariant for simplicity. The proportionality factor β_j is given by

$$\beta_j = \frac{\text{cov}\{r_{jt}, r_{mt}\}}{V\{r_{mt}\}} \tag{2.78}$$

and indicates how strong fluctuations in the returns on asset j are related to movements of the market as a whole. As such, it is a measure of systematic risk (or market risk). Because it is impossible to eliminate systematic risk through a diversification of one's portfolio without affecting the expected return, investors are compensated for bearing this source of risk through a risk premium $E\{r_{mt} - r_f\} > 0$.

In this section, we consider the CAPM and see how it can be rewritten as a linear regression model, which allows us to estimate and test it. A more extensive discussion of empirical issues related to the CAPM can be found in Berndt (1991) or, more technically, in Campbell, Lo and MacKinlay (1997, Chapter 5) and Gourieroux and Jasiak (2001, Section 4.2). More details on the CAPM can be found in finance textbooks, for example Elton and Gruber (2003).

2.7.1 The CAPM as a Regression Model

The relationship in (2.77) is an *ex ante* equality in terms of unobserved expectations. Ex post, we only observe realized returns on the different assets over a number of

[21] Because the data correspond to different time periods, we index the observations by t, $t = 1, 2, \dots, T$, rather than i.

periods. If, however, we make the usual assumption that expectations are rational, so that expectations of economic agents correspond to mathematical expectations, we can derive a relationship from (2.77) that involves actual returns. To see this, let us define the unexpected returns on asset j as

$$u_{jt} = r_{jt} - E\{r_{jt}\},$$

and the unexpected returns on the market portfolio as

$$u_{mt} = r_{mt} - E\{r_{mt}\}.$$

Then, it is possible to rewrite (2.77) as

$$r_{jt} - r_f = \beta_j(r_{mt} - r_f) + \varepsilon_{jt}, \tag{2.79}$$

where

$$\varepsilon_{jt} = u_{jt} - \beta_j u_{mt}.$$

Equation (2.79) is a regression model, without an intercept, where ε_{jt} is treated as an error term. This error term is not something that is just added to the model, but it has a meaning, being a function of unexpected returns. It is easy to show, however, that it satisfies some minimal requirements for a regression error term, as given in (A7). For example, it follows directly from the definitions of u_{mt} and u_{jt} that it is mean zero, i.e.

$$E\{\varepsilon_{jt}\} = E\{u_{jt}\} - \beta_j E\{u_{mt}\} = 0. \tag{2.80}$$

Furthermore, it is uncorrelated with the regressor $r_{mt} - r_f$. This follows from the definition of β_j, which can be written as

$$\beta_j = \frac{E\{u_{jt}u_{mt}\}}{V\{u_{mt}\}},$$

(note that r_f is not stochastic) and the result that

$$E\{\varepsilon_{jt}(r_{mt} - r_f)\} = E\{(u_{jt} - \beta_j u_{mt})u_{mt}\} = E\{u_{jt}u_{mt}\} - \beta_j E\{u_{mt}^2\}.$$

From the previous section, it then follows that the OLS estimator provides a consistent estimator for β_j. If, in addition, we impose assumption (A8) that ε_{jt} is independent of $r_{mt} - r_f$ and assumptions (A3) and (A4) stating that ε_{jt} does not exhibit autocorrelation or heteroskedasticity, we can use the asymptotic result in (2.74) and the approximate distributional result in (2.76). This implies that routinely computed OLS estimates, standard errors and tests are appropriate, by virtue of the asymptotic approximation.

2.7.2 Estimating and Testing the CAPM

The CAPM describes the expected returns on any asset or portfolio of assets as a function of the (expected) return on the market portfolio. In this subsection, we consider the returns on three different industry portfolios, while approximating the return on the market portfolio by the return on a value-weighted stock market index. Returns for the period January 1960 to December 2002 (516 months) for the food, durables and construction industries were obtained from the Center for Research in Security Prices (CRSP).[22] The industry portfolios are value-weighted and are rebalanced once every year. While theoretically the market portfolio should include all tradeable assets, we shall assume that the CRSP value-weighted index is a good approximation. The riskless rate is approximated by the return on 1-month treasury bills. Although this return is time-varying, it is known to investors while making their decisions. All returns are expressed in percentage per month.

First, we estimate the CAPM relationship (2.79) for these three industry portfolios. That is, we regress excess returns on the industry portfolios (returns in excess of the riskless rate) upon excess returns on the market index proxy, not including an intercept. This produces the results presented in Table 2.3. The estimated beta coefficients indicate how sensitive the value of the industry portfolios are to general market movements. This sensitivity is relatively low for the food industry, but fairly high for durables and construction: an excess return on the market of, say, 10% corresponds to an expected excess return on the food, durables and construction portfolios of 7.9, 11.1 and 11.6% respectively. It is not surprising to see that the durables and construction industries are more sensitive to overall market movements than is the food industry. Assuming that the conditions required for the distributional results of the OLS estimator are satisfied, we can directly test the hypothesis (which has some economic interest) that $\beta_j = 1$ for each of the three industry portfolios. This results in t-values of -7.38, 3.89 and 6.21, respectively, so that we reject the null hypothesis for each of the three industry portfolios.

As the CAPM implies that the only relevant variable in the regression is the excess return on the market portfolio, any other variable (known to the investor when making his decisions) should have a zero coefficient. This also holds for a constant term. To check whether this is the case, we can re-estimate the above models while including an intercept term. This produces the results in Table 2.4. From these results, we can test the validity of the CAPM by testing whether the intercept term is zero. For food, the appropriate t-statistic is 2.66, which implies that we reject the validity of the CAPM at the 5% level. The point estimate of 0.339 implies that the food industry portfolio is expected to have a return that is 0.34% per month higher than the CAPM predicts. Note that the estimated beta coefficients are very similar to those in Table 2.3 and that the R^2s are close to the uncentred R^2s.

Table 2.3 CAPM regressions (without intercept)

Dependent variable: *excess industry portfolio returns*			
Industry	Food	Durables	Construction
excess market return	0.790	1.113	1.156
	(0.028)	(0.029)	(0.025)
uncentred R^2	0.601	0.741	0.804
s	2.902	2.959	2.570

Note: Standard errors in parentheses.

[22] The data for this illustration are available as CAPM2.

Table 2.4 CAPM regressions (with intercept)

Dependent variable: *excess industry portfolio returns*			
Industry	Food	Durables	Construction
constant	0.339	0.064	−0.053
	(0.128)	(0.131)	(0.114)
excess market return	0.783	1.111	1.157
	(0.028)	(0.029)	(0.025)
R^2	0.598	0.739	0.803
s	2.885	2.961	2.572

Note: Standard errors in parentheses.

The R^2s in these regressions have an interesting economic interpretation. Equation (2.79) allows us to write that

$$V\{r_{jt}\} = \beta_j^2 V\{r_{jt}\} + V\{\varepsilon_{jt}\},$$

which shows that the variance of the return on a stock (portfolio) consists of two parts: a part related to the variance of the market index and an idiosyncratic part. In economic terms, this says that total risk equals market risk plus idiosyncratic risk. Market risk is determined by β_j and is rewarded: stocks with a higher β_j provide higher expected returns because of (2.77). Idiosyncratic risk is not rewarded because it can be eliminated by diversification: if we construct a portfolio that is well diversified, it will consist of a large number of assets, with different characteristics, so that most of the idiosyncratic risk will cancel out and mainly market risk matters. The R^2, being the proportion of explained variation in total variation, is an estimate of the relative importance of market risk for each of the industry portfolios. For example, it is estimated that 59.8% of the risk (variance) of the food industry portfolio is due to the market as a whole, while 40.2% is idiosyncratic (industry-specific) risk. The durables and construction industries appear to be better diversified.

Finally, we consider one deviation from the CAPM that is often found in empirical work: the existence of a January effect. There is some evidence that, ceteris paribus, returns in January are higher than in any of the other months. We can test this within the CAPM framework by including a dummy in the model for January and testing

Table 2.5 CAPM regressions (with intercept and January dummy)

Dependent variable: *excess industry portfolio returns*			
Industry	Food	Durables	Construction
constant	0.417	0.069	−0.094
	(0.133)	(0.137)	(0.118)
January dummy	−0.956	−0.063	0.498
	(0.456)	(0.473)	(0.411)
excess market return	0.788	1.112	1.155
	(0.028)	(0.029)	(0.025)
R^2	0.601	0.739	0.804
s	2.876	2.964	2.571

Note: Standard errors in parentheses.

whether it is significant. By doing this, we obtain the results in Table 2.5. Computing the t-statistics corresponding to the January dummy shows that for two of the three industry portfolios we do not reject the absence of a January effect at the 5% level. For the food industry, however, the January effect appears to be negative and statistically significant. Consequently, the results do not provide support for the existence of a positive January effect.

2.8 Multicollinearity

In general, there is nothing wrong with including variables in your model that are correlated. In an individual wage equation, for example, we may want to include both age and experience, although it can be expected that older persons, on average, have more experience. However, if the correlation between two variables is too high, this may lead to problems. Technically, the problem is that the matrix $X'X$ is close to being not invertible. This may lead to unreliable estimates with high standard errors and of unexpected sign or magnitude. Intuitively, the problem is also clear. If age and experience are highly correlated it may be hard for the model to identify the *individual* impact of these two variables, which is exactly what we are trying to do. In such a case, a large number of observations with sufficient variation in both age and experience may help us to get sensible answers. If this is not the case and we do get poor estimates (for example: t-tests show that neither age nor experience are individually significant), we can only conclude that there is insufficient information in the sample to identify the effects we would like to identify. In the wage equation, we are trying to identify the effect of age, keeping experience and the other included variables constant, as well as the effect of experience, keeping age and the other variables constant (the ceteris paribus condition). It is clear that in the extreme case that people with the same age would have the same level of experience we would not be able to identify these effects. In the case where age and experience are highly but not perfectly correlated, the estimated effects are likely to be highly inaccurate.

In general, the term **multicollinearity** is used to describe the problem when an approximate linear relationship among the explanatory variables leads to unreliable regression estimates. This approximate relationship is not restricted to two variables but can involve more or even all regressors. In the wage equation, for example, the problems may be aggravated if we include years of schooling in addition to age and years of experience. In the extreme case, one explanatory variable is an exact linear combination of one or more other explanatory variables (including the intercept). This is usually referred to as **exact multicollinearity**, in which case the OLS estimator is not uniquely defined from the first order conditions of the least squares problem (the matrix $X'X$ is not invertible).

The use of too many dummy variables (which are either zero or one) is a typical cause for exact multicollinearity. Consider the case where we would like to include a dummy for males ($male_i$), a dummy for females ($female_i$) as well as a constant. Because $male_i + female_i = 1$ for each observation (and 1 is included as the constant), the $X'X$ matrix becomes singular. Exact multicollinearity is easily solved by excluding one of the variables from the model and estimating the model including either $male_i$ and a constant, $female_i$ and a constant, or both $male_i$ and $female_i$ but no constant. The latter