Volatility

Volatility is a critical input necessary for pricing options, but it is not directly observable and (as we will see) it is not constant over time. Consequently, both theorists and practitioners are concerned with the behavior of volatility and the construction of option pricing models in which volatility can change. Hedging volatility risk is also an important issue for market-makers.

In this chapter we will discuss specific techniques for measuring volatility and also demonstrate how volatility models can be incorporated into the Black-Scholes pricing framework. The pricing models have the potential to explain observed prices better than the Black-Scholes formula, but the models are derived using the same pricing principles we discussed in Chapter 21.

The chapter is divided into four general topics:

- Implied volatility: what information do option prices provide about the stock price distribution?
- Volatility estimation: given the past history of stock returns, what can you say about volatility?

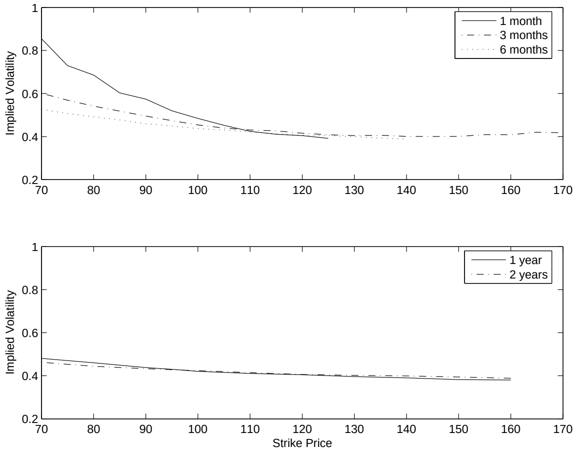
- Volatility hedging: what instruments are available hedge volatility risk?
- Pricing options when volatility is stochastic

You should keep in mind when reading this chapter that the distribution of asset prices remains an area in which there is significant ongoing research. Statistical techniques for measuring volatility continue to evolve, and there is continuing research into the question of which pricing models best explain observed option prices.

25.1 IMPLIED VOLATILITY

To provide a context for the discussion in this chapter, we begin by revisiting implied volatility, a concept we discussed in Section 12.1. Figure 25.1 depicts implied volatilities for exchange-traded IBM options on January 18, 2001, an arbitrarily-chosen date. The patterns are typical for equity options, with in-the-money (low strike) calls having higher implied volatilities than at-the-money and out-of-the-money calls. In addition, the implied volatility curve is flatter for options with longer time to maturity.

The pattern of implied volatilities generally is referred to as the volatility skew. However, specific patterns are frequently observed. If you use your imagination, the implied volatility plot in Figure 25.1 resembles a lopsided grin or a smirk. The pattern in the figure is sometimes called a *volatility smirk*. When the plot of implied volatility against strike prices looks like a smile, it is called



25.1 Implied volatilties for IBM call options, January 18, 2001. The top panel shows implied volatilities for options with 1, 3, and 6 months to expiration, while the bottom panel shows implied volatilities for options with 1 and 2 years to expiration. The IBM closing price on January 18 was $108.31. Source: Optionmetrics

or a **volatility smile.** Volatility *frowns* may also be observed.

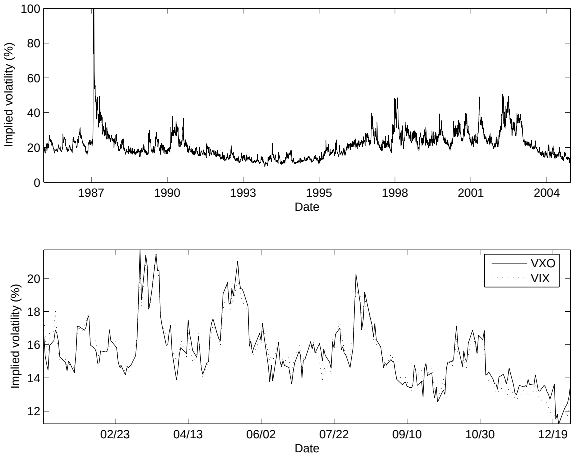
Implied volatility may seem like a natural way to measure the volatility that is expected to prevail over a future period of time. However, there are two problems. First, we typically measure implied volatility by inferring volatility from an option pricing formula that assumes volatility is constant. This raises the question of what an implied volatility is actually measuring. Second, as we saw in Chapter 12, there is no one measure of implied volatility; volatilities are typically different for different strike prices and maturities with the same underlying asset. This is evident in Figure 25.1. This again raises the question of how to interpret the implied volatility numbers: should we look at volatility at a particular "moneyness"? Is there some way to average the different volatilities? We will see later in this chapter that some theoretical pricing models are able to account for implied volatility patterns such as those in Figure 25.1.

In addition to examining the pattern of implied volatilities at a point in time, it is also possible to track implied volatilities over time. Since 1993 the Chicago Board Options Exchange (CBOE) has reported an index of implied volatility for near-term S&P 100 index options. This index is called the "VIX", after its ticker symbol. Using this measure, we can track changes over time in implied volatility. Originally, the CBOE computed implied volatility by extracting implied volatility from near-the-money options, much as we discussed in Section 12.1. **This section reference should resolve corectly to 12.5.** This index is called the "Old VIX", with ticker symbol VXO. However, beginning in 2003, the CBOE began computing implied volatility based on a new formula that we will describe later in this chapter.

Figure 25.2 plots the old VIX index from 1986 to 2004, and compares the new and old VIX for one year. The spike in the VIX in 1987 occurs on four days, October 19, 20, 22, and 26, when the VIX exceeded 100%. This period corresponds to the October 19, 1987 market crash, in which the Dow Jones index declined over 20% on one day.

The bottom panel in Figure 25.2 compares the new and old VIX during one year, showing that while the two measures are not identical, they are generally quite close. At the scale in the top panel, it would be difficult to detect any difference at all in the two series.

We will see in Section 25.3 below that the new VIX measure neatly finesses



25.2 The top panel depicts the old VIX from 1986 to 2004. The bottom panel compares the new and old VIX indices during 2004. Source: CBOE

the problem of which particular option implied volatility to use by considering prices of *all* out-of-the-money options with a given time to maturity.

25.2 MEASUREMENT AND BEHAVIOR OF VOLATILITY

In this section we examine different ways to characterize and measure the behavior of volatility using only historical information about the asset price. In our examples we will concentrate on stock price volatility.

We take as a starting point the standard lognormal model of stock prices, which is used in deriving the Black-Scholes formula:

$$S_{t+h} = S_t e^{(\alpha-\delta-0.5\sigma^2)h+\sigma\sqrt{h}Z(t)} \tag{25.62}$$

where α is the continuously-compounded expected return on the stock, δ is the continuously-compounded dividend yield, σ is the volatility, and $Z(t) \sim \mathcal{N}(0,1)$. If we observe a series of stock prices, we can compute continuously-compounded returns, ϵ, over a period of length h:

$$\epsilon_{t+h} = \ln(S_{t+h}/S_t)$$

We will assume throughout this section that h is small, and therefore that we can ignore the mean return.

Historical Volatility

The natural starting point for examining volatility is historical volatility, which we compute using past stock returns. Under the assumption that volatility is constant, we can estimate annual the annual variance of returns, σ^2, as

$$\hat{\sigma}^2_{\mathrm{H}} = \frac{1}{h(n-1)} \sum_{i=1}^{n} \epsilon_i^2 \tag{25.63}$$

This calculation differs from the usual formula for variance since it assumes the per-period mean is zero. This assumption makes little difference if h is small. We annualize the variance estimate by dividing by h.[21]

---

[21] Suppose r_i is computed daily. It is important to annualize *after* computing the standard deviation of the daily returns, as opposed to annualizing the daily returns and then computing the standard deviation. To see why, let $\{r_i\}$ be the set of non-annualized daily returns. The annualized standard deviation is

$$\sqrt{252} \times \sqrt{\mathrm{Var}(\{r_i\})}$$

The top panel of Figure 25.3 displays the historical 60-day volatility for IBM and the S&P 500 index from 1991 to 2002. Each day, the preceding 60 trading days are used to compute the standard deviation of the continuously compounded daily return. The resulting standard deviation is multiplied by $\sqrt{252}$ to produce an annualized standard deviation. It is common to annualize using 252 days instead of 365 days because the volatility on weekends and holidays is lower than on ordinary business days. If all days were the same, return volatility over a weekend should be $\sqrt{3}$ times the weekday volatility. Weekend volatility is significantly less.[22]

The use of overlapping 60-day intervals induces smoothness in the series, since each day's return affects next 60 days of volatility calculations. Even so, there is a great deal of variability in the standard deviation. For both series, volatility appears to have risen toward the end of the 1990s.

Exponential Weighted Moving Average

Because volatility in Figure 25.3 appears to be changing over time, it is natural to try to take this variation into account when estimating volatility. We

---

If instead we first annualize the returns by multiplying them by 252, and then compute the standard deviation, we obtain

$$\sqrt{\text{var}(252 \times \{r_i\})} = 252 \times \sqrt{\text{Var}(\{r_i\})}$$

Annualizing returns before computing the standard deviation creates a return series that has too much volatility.

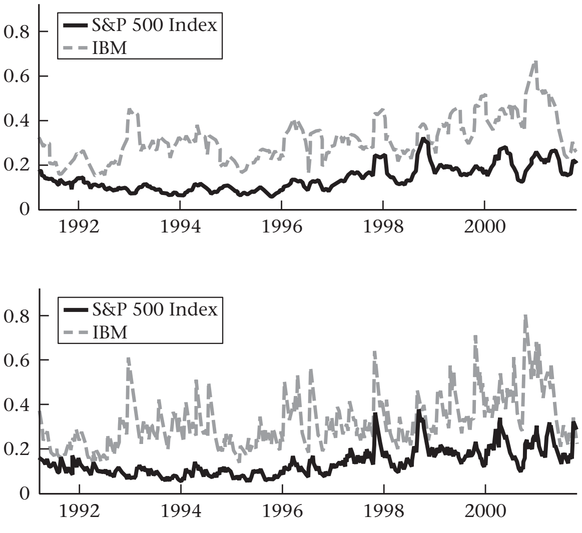
[22] French and Roll (1986) first showed that returns over weekends were less volatile than would be expected for a three-day return.

might reason that if volatility is changing, we want to emphasize more recent observations at the expense of more distant observations. One way to do this is to compute an *exponential weighted moving average* (EWMA) of the squared stock returns.

The EWMA formula computes volatility at time t as a weighted average of the time $t-1$ EWMA estimate, $\hat{\sigma}^2_{\mathrm{EWMA},t-1}$, and the time $t-1$ squared stock price change, ϵ^2_{t-1}. Thus, we have

$$\hat{\sigma}^2_{\mathrm{EWMA},t} = (1-b)\epsilon^2_{t-1} + b\hat{\sigma}^2_{\mathrm{EWMA},t-1} \tag{25.64}$$

where b is the weight applied to the previous EWMA estimate. We can lag equation (25.64) and substitute the resulting expression for $\hat{\sigma}^2_{\mathrm{EWMA},t-1}$ into the



25.3 Sixty-day volatility estimates for IBM and the S&P 500 index. The top panel shows volatility estimates using equation (25.63), whereas the bottom panel uses equation (25.66), with $b = 0.94$.

right hand side of equation (25.64). Continuing in this way, we obtain the EWMA estimator as a weighted average of past squared returns:

$$\hat{\sigma}^2_{\mathrm{EWMA},t} = \sum_{i=0}^{\infty} \left[(1-b)b^i\right] \epsilon^2_{t-1-i} \tag{25.65}$$

The term in square brackets in equation (25.66) is the weight applied to historical returns. The weights decline at the rate b, with the most recent return receiving the greatest weight. Because $\sum_{i=0}^{\infty}(1-b)b^i = 1$, the weights on past squared returns sum to one.

It is also possible to use a moving window when estimating EWMA volatility. For example, we might use only the previous n days of data. In this case, equation (25.66) becomes

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{EWMA},t} &= \sum_{i=0}^{n} \left[\frac{(1-b)b^i}{\sum_{j=1}^{n}(1-b)b^{j-1}}\right] \epsilon^2_{t-1-i} \\ &= \sum_{i=0}^{n} \left[\frac{(1-b)b^i}{1-b^{n+1}}\right] \epsilon^2_{t-1-i} \qquad (25.66)\end{aligned}$$

Because $\sum_{i=0}^{n}(1-b)b^i = 1-b^{n+1}$, the weights again sum to one.

There is also a simple updating formula, analogous to equation (25.64), in the case of a moving window estimate. Each period we add the latest observation and drop the oldest observation. Equation (25.66) is equivalent to

$$\hat{\sigma}^2_{\mathrm{EWMA},t} = \frac{1-b}{1-b^{n+1}}\epsilon^2_{t-1} + b\hat{\sigma}^2_{\mathrm{EWMA},t-1} - \frac{1-b}{1-b^{n+1}}b^{n+1}\epsilon^2_{t-1-n} \tag{25.67}$$

Example 25.1 Suppose $b = 0.94$ and $n = 60$. We have $1 - b^{n+1} = 0.977$. The first term in equation (25.67) is then

$$\frac{(1-0.94)}{0.977} = 0.0614$$

This compares with a weight of $1/60 = 0.0167$ for each observation in the equal-weighted estimator in equation 25.63. Subsequent (earlier) observations have weights of 0.05778, 0.0543, 0.0510, etc.

The bottom panel in Figure 25.3 displays the EWMA estimate for $b = 0.94$ and $n = 60$ days. Note that the EWMA estimator is considerably more variable than the standard historical volatility estimate. This additional variability occurs because the most recent observation has four times the weight in the EWMA estimator as in the standard estimator. Thus a particularly large return will create a large effect on the estimate. This effect will then decay at the rate b. ■

There are two problems with the EWMA estimator, one practical and one conceptual. First, if we use the EWMA estimator in equation (25.64) to forecast future volatility, we obtain a constant expected volatility at any horizon. The reason is that the forecast of ϵ_t^2 is $\hat{\sigma}_{t-1}^2$, so that all forecasts of future volatility would equal $\hat{\sigma}_{t-1}^2$. Thus, the EWMA estimator does not forecast patterns in future volatility. Second, the EWMA estimator is not derived from a formal statistical model in which volatility can vary over time. We have not assumed a distribution for the ϵ_t^2, so we cannot conduct hypothesis tests using the model. ARCH and GARCH, which we discuss next, address both problems.

Time-Varying Volatility: ARCH and GARCH

A casual examination of data, such as looking at historical volatilities (Figure 25.3), or looking the behavior over time of implied volatilities (Figure 25.2) suggests that volatilities are not constant.[23] What do we do once we formally accept that volatilities change over time? Ideally we would have a statistical model that permits volatility changes to occur. Such a model could serve both to provide better estimates of volatility and also to provide a building block for better pricing models.

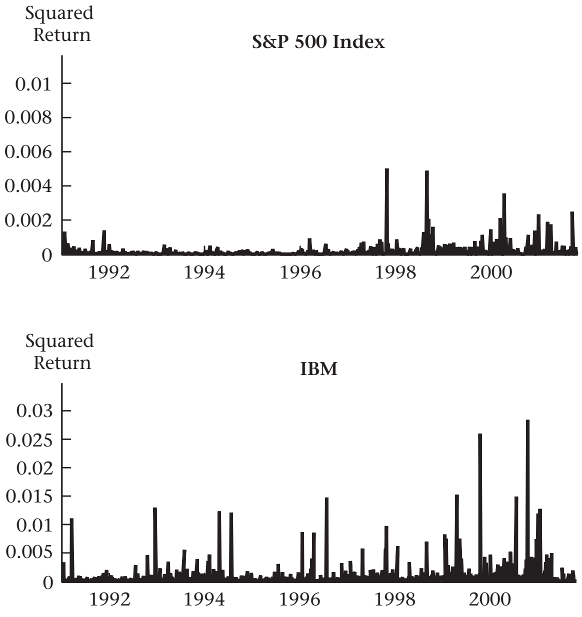
Research on the behavior of volatility shows that for many assets, there are periods of turbulence and periods of calm: high volatility tends to be followed by high volatility and low volatility by low volatility. Put differently, during a period when measured volatility is high, the typical day tends to exhibit high volatility. (High volatility could in principle also arise from an increased chance of large but infrequent price moves.) Figure 25.4 displays squared daily returns for the S&P 500 index and IBM. At a casual level, this figure exhibits this effect,

---

[23] We can perform a back-of-the-envelope calculation by assuming that continuously compounded returns are normally distributed. In that case, the ratio of variances drawn from independent time periods has the the F distribution. If we estimate two annual volatilities using 252 observations, the ratio of the two estimated variances is distributed $F(\alpha, 251, 251)$. The 99.5% and 0.005% confidence levels are obtained from $F^{-1}(p, 251, 251)$, where $p = 0.995$ or $p = 0.005$. At a 1% significance level, the two bounds for the ratio of estimated variances are 1.386 and 0.722, corresponding to volatility ratios of 1.177 and 0.849. Thus, if volatility in one year is 15%, a subsequent measured volatility outside the range 12.74% – 17.66% rejects the hypothesis of constant variance at a 1% significance level.

with periods in which many of the daily squared returns are large, and periods when many are small. This is called **volatility clustering**.

If volatility is persistent, a volatility measure should weight recent returns more heavily than more distant returns. This difference in weighting is exactly how an EWMA volatility estimate differs from the ordinary equally weighted volatility measure. ARCH and GARCH models that also give weight to recent returns.



25.4 Squared daily returns on the S&P 500 index (top panel) and IBM (bottom panel) from January 2, 1991, to October 24, 2001.

ARCH

The ARCH (Autoregressive Conditional Heteroskedasticity) model of Engle (1982) and the subsequent GARCH (Generalized ARCH) model of Bollerslev (1986) are important and widely used volatility models that attempt to capture statistically the ebb and flow of volatility.[24] Engle in fact won the 2003 Nobel Prize in economics for his work in this area—see the box on p. 410. The basic idea motivating ARCH is that if volatility is high today it is likelier than average to be high tomorrow. Engle (1982) provided a statistical framework for modeling this effect.

We have been assuming that a statistical model for asset prices would take the form

$$\ln(S_t/S_{t-h}) = (\alpha - \delta - 0.5\sigma^2)h + \epsilon_t \tag{25.68}$$

In this specification, the error term would have variance

$$\text{var}(\epsilon_t) = \sigma^2 h \tag{25.69}$$

If σ^2 is constant over time, we say the error term, ϵ_t, is **homoskedastic**. Based on Figure 25.4, however, a more reasonable specification would be to assume that the variance of ϵ_t varies over time, in which case it is **heteroskedastic**.

---

[24] Bollerslev *et al.* (1994) surveys the literature on ARCH and its variants. Two recent accessible introductions were written as a result of the Nobel prize: Diebold (2004) and Royal Swedish Academy of Sciences (2003).

A Nobel Prize for Volatility

The 2003 Nobel Prize in economics was awarded to Robert F. Engle and Clive Granger for their work in statistical methods in economics. Engle was cited for his work in studying the behavior of volatility. This quote, from the Royal Swedish Academy of Science press release announcing the award of the 2003 economics prize, describes Engle's contribution:

> ...random fluctuations over time—volatility—are particularly significant because the value of shares, options and other financial instruments depends on their risk. Fluctuations can vary considerably over time; turbulent periods with large fluctuations are followed by calmer periods with small fluctuations. Despite such time-varying volatility, in want of a better alternative, researchers used to work with statistical methods that presuppose constant volatility. Robert Engles discovery was therefore a major breakthrough. He found that the concept of autoregressive conditional heteroskedasticity (ARCH) accurately captures the properties of many time series and developed methods for statistical modeling of time-varying volatility. His ARCH models have become indispensable tools not only for researchers, but also for analysts on financial markets, who use them in asset pricing and in evaluating portfolio risk.

Granger was cited for his work in cointegration, a statistical method important for studying the long-run behavior of economic time series.

If the time interval in equation (25.68) is short, then (as we saw in Chapter 20), the mean, $(\alpha - \delta - 0.5\sigma^2)h$ is small, and ϵ_t^2 is essentially the squared return. We will continue to assume that h is short enough so that we can ignore the drift in equation (25.68).

Let Ψ_t denote the information that is available up to and including time t, and therefore information that we have available at time t. The idea behind Engle's ARCH model is that squared returns have a variance that changes over time according to a statistical model. Specifically, let q_t be the conditional (upon information available at time $t-1$) value of the return variance, i.e.,

$$q_t \equiv E(\epsilon_t^2|\Psi_{t-1})$$

The ARCH model supposes that we can write

$$q_t = a_0 + \sum_{i=1}^{m} a_i \epsilon_{t-i}^2 \tag{25.70}$$

where $a_0 > 0, a_i \geq 0, i = 1, ..., m$. Equation (25.70) is referred to as an ARCH(m) model, signifying that there are m lagged terms. In order for volatility to be well-behaved, we must have $\sum_{i=1}^{n} a_i < 1$. This model states that volatility at a point in time depends upon recent observed volatility.

At this point we can understand the meaning of "autoregressive conditional heteroskedasticty". *Autoregressive* means that the value at a point in time depends on past values. *Heteroskedasticity* means that variances are not equal. The unconditional variance is the variance estimated over a long period of time. The *conditional* variance is the variance estimated at a point in time, estimated taking into account ("conditional upon") recent volatility. Thus, *autoregressive conditional heteroskedasticity* essentially means that the level of variance depends on recent past levels of variance. This is the behavior captured by equation (25.70).

An important practical question is how many lags we need in order to estimate equation (25.70). To better understand the behavior of an ARCH model, let's consider a single lag, where we set $a_1 > 0$ and $a_i = 0, i > 1$. The volatility equation is then

$$E(\epsilon_t^2|\Psi_{t-1}) = a_0 + a_1\epsilon_{t-1}^2 \tag{25.71}$$

where $a_0 > 0$ and $a_1 < 1$. Equation (25.71) is an ARCH(1) model.

Suppose we forecast volatility at time $t+1$, $t+2$, etc., using only the information we have at time t. Equation (25.71) implies that for a one-period ahead forecast of q_t we have

$$\begin{aligned} E(\epsilon_{t+1}^2|\Psi_{t-1}) &= a_0 + a_1 E(\epsilon_t^2|\Psi_{t-1}) \\ &= a_0 + a_1\left(a_0 + a_1\epsilon_{t-1}^2\right) \end{aligned}$$

Similarly, for a two-period ahead forecast,

$$\begin{aligned} E(\epsilon_{t+2}^2|\Psi_{t-1}) &= a_0 + a_1 E(\epsilon_{t+1}^2|\Psi_{t-1}) \\ &= a_0 + a_1\left[a_0 + a_1\left(a_0 + a_1\epsilon_{t-1}^2\right)\right] \end{aligned}$$

Continuing in this way, for an n-period-ahead forecast, we have

$$E(\epsilon_{t+n}^2|\Psi_{t-1}) = a_0\left(1 + \sum_{i=1}^{n} a_i^i\right) + a_1^n\epsilon_{t-1}^2 \tag{25.72}$$

This predicted pattern of volatility persistence is very specific and inflexible. A large squared return today implies larger squared returns at all future dates, but the effect decays per period by the factor a_1. Shocks to volatility are expected to die off at a constant rate.

Equation (25.72) implies that unconditional volatility (the value we would estimate as a long-run average) is

$$\bar{\sigma}^2 = a_0(1 + a_1 + a_1^2 + \ldots) = \frac{a_0}{1 - a_1} \tag{25.73}$$

Thus, with estimates of a_0 and a_1 we can compute the unconditional volatility.

In practice, if markets become more turbulent, they may remain more turbulent for a period of time. Equation (25.70) with a single lag cannot account for a period of *sustained* high volatility. As you might guess, more than one lag—generally many lags—are necessary for ARCH to fit the data.

GARCH

The GARCH model, due to Bollerslev (1986), is a variant of ARCH that allows for infinite lags, yet can be estimated with a small number of parameters. The GARCH model has the form

$$q_t = a_0 + \sum_{i=1}^{m} a_i \epsilon_{t-i}^2 + \sum_{j=1}^{n} b_i q_{t-i} \tag{25.74}$$

where $a_0 > 0$, $a_i \geq 0, i = 1, ..., m$, $b_i \geq 0, i = 1, ..., n$, and $\sum_{i=1}^{m} a_i + \sum_{i=1}^{n} b_i < 1$. This model states that volatility at a point in time depends upon recent volatility as well as recent squared returns. Equation (25.74) is referred to as a GARCH(m,n) model.

GARCH(1,1) is frequently used in practice. The GARCH(1,1) model is

$$q_t = a_0 + a_1 \epsilon_{t-1}^2 + b_1 q_{t-1} \tag{25.75}$$

It is instructive to compare the GARCH(1,1) model to the ARCH(1) model, equation (25.71). To do this, we can rewrite equation (25.75) to eliminate q_{t-1} on the right-hand side. Lagging equation (25.75) and substituting the result for q_{t-1} on the right-hand side of equation (25.75), we obtain

$$q_t = a_0 + a_1\epsilon^2_{t-1} + b_1\left(a_0 + a_1\epsilon^2_{t-2} + b_1 q_{t-2}\right)$$

Continuing in this way, we obtain

$$\begin{aligned} q_t &= a_0\sum_{i=0}^{\infty} b_1^i + a_1\sum_{i=0}^{\infty} b_1^i\epsilon^2_{t-1-i} \\ &= \frac{a_0}{1-b_1} + \frac{a_1}{1-b_1}\sum_{i=0}^{\infty}(1-b_1)b_1^i\epsilon^2_{t-1-i} \qquad (25.76) \end{aligned}$$

A GARCH(1,1) model is therefore equivalent to an ARCH(∞) model in which the lag coefficients decline at the rate b_1. Notice that the last term in equation (25.76) can be rewritten in terms of an EWMA volatility estimator (25.66):

$$q_t = \frac{a_0}{1-b_1} + \frac{a_1}{1-b_1}\hat{\sigma}^2_{\text{EWMA}} \qquad (25.77)$$

It is important to note that the parameter b in the EWMA expression in equation (25.77) is not arbitrarily chosen, as in equation (25.64), but is estimated as part of the GARCH estimation procedure.

Maximum Likelihood Estimation of a GARCH Model

Given the assumption that continuously-compounded returns are normally distributed with variance q_t and mean zero, we can estimate a GARCH model using maximum likelihood.[25] The probability density for ϵ_t, conditional on q_t, is

$$f(\epsilon_t; q_t) = \frac{1}{\sqrt{2\pi q_t}} e^{-0.5e_t^2/q_t} \tag{25.78}$$

Since the ϵ_t are independent, the probability of observing the particular set of n returns is the product of the probabilities, which gives us the likelihood function:

$$\prod_{i=1}^{n} f(\epsilon_i | q_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi q_i}} e^{-0.5\epsilon_i^2/q_i}$$

For a GARCH(1,1), q_i is a function of a_0, a_1, and b_1. The maximum likelihood estimate is the set of parameters—a_0, a_1, and b_1—that maximizes the probability of observing the returns we actually observed. Typically it is easiest to maximize the log of the likelihood function, in which case maximizing the likelihood is the same as maximizing

$$\sum_{i=1}^{n} -0.5\ln(q_i) - 0.5\epsilon_i^2/q_i \tag{25.79}$$

We omit the term $-0.5\ln(2\pi)$ since it does not affect the solution. The maximization of equation (25.79) can be performed in statistical packages or even using Solver in Excel (see the Chapter appendix).

Volatility Forecasts

We can forecast volatility in the GARCH(1,1) model as we did in the

---

[25] Alexander (2001) discusses the estimation of GARCH models and is replete with examples.

ARCH(1) model. To understand the calculation, recognize that since $q_t = E(\epsilon_t^2|\Psi_{t-1})$, then we have $E(q_t|\Psi_{t-j}) = E(\epsilon_t^2|\Psi_{t-j})$ for $j \geq 1$. Thus, using equation (25.75), we have

$$\begin{aligned} E(q_{t+1}|\Psi_{t-1}) &= a_0 + a_1 E(\epsilon_t^2|\Psi_{t-1}) + b_1 E(q_t|\Psi_{t-1}) \\ &= a_0 + (a_1 + b_1) E(q_t|\Psi_{t-1}) \\ &= a_0 + (a_1 + b_1)(a_0 + a_1 \epsilon_{t-1}^2 + b_1 q_{t-1}) \end{aligned}$$

The goal in this calculation is to express the forecasted value of q_{t+1} in terms of what we can observe at time t, namely ϵ_{t-1} and q_{t-1}. Following the same procedure, we obtain

$$E(q_{t+2}|\Psi_{t-1}) = a_0(1 + a_1 + b_1) + (a_1 + b_1)^2 (a_0 + a_1 \epsilon_{t-1}^2 + b_1 q_{t-1})$$

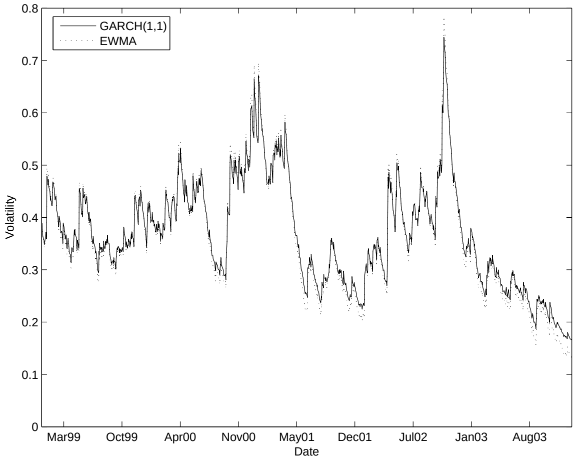
For a k step-ahead forecast, we have

$$E(q_{t+k}|\Psi_{t-1}) = a_0 \sum_{i=0}^{k-1} (a_1 + b_1)^i + (a_1 + b_1)^k (a_1 \epsilon_{t-1}^2 + b_1 q_{t-1})$$

As we let k go to infinity, we obtain an estimate of unconditional volatility in the GARCH(1,1) model:

$$\bar{\sigma}^2 = a_0 \sum_{i=0}^{\infty} (a_1 + b_1)^i = \frac{a_0}{1 - a_1 - b_1} \tag{25.80}$$

Using equations (25.77) and (25.80), we can express the GARCH(1,1) equa-



25.5 Comparison of GARCH(1,1) volatility and EWMA volatility for IBM, January 1999 to December 2003. The EWMA volatility estimate sets $b = 0.94$, while the GARCH parameters are estimated.

tion in terms of the EWMA estimate of volatility:

$$q_t = \alpha\bar{\sigma}^2 + (1-\alpha)\hat{\sigma}^2_{\text{EWMA}} \tag{25.81}$$

where $\alpha = (1 - a_1 - b_1)/(1 - b_1)$. Thus, the GARCH(1,1) expected volatility at a point in time is a weighted average of the unconditional variance, $\bar{\sigma}^2$, and the current estimated EWMA volatility, $\hat{\sigma}^2_{\text{EWMA}}$.

Example 25.2 Estimating a GARCH(1,1) model for IBM using daily return data from January 1999 to December 2003 yields the GARCH volatility estimate

$$q_t = 0.000001305 + 0.0446\epsilon^2_{t-1} + 0.9552q_{t-1}$$

The implied estimate of the unconditional annualized volatility is

$$\sqrt{\frac{0.000001305}{1 - 0.0446 - 0.9552} \times 252} = 1.5318$$

The historical volatility during this period was 39.85%. An estimated unconditional volatility of 153% suggests that the GARCH(1,1) model has trouble fitting the data. In this case, it turns out that the problem is caused by large returns on days during which IBM announced earnings.

During the 1999-2003 period there were four days on which the absolute one-day return exceeded 12%. On each of these days (April 21, 1999; October 20, 1999; July 19, 2000; and October 17, 2000), IBM announced earnings. The 153% volatility illustrates the GARCH model's difficulty in explaining these large magnitude returns under the assumption that returns are normally distributed. If we omit these four days from the sample, the estimated GARCH model is

$$q_t = 0.000002203 + 0.0507\epsilon_{t-1}^2 + 0.9462q_{t-1}$$

These parameters imply an unconditional volatility of

$$\sqrt{\frac{0.000002203}{1 - 0.0507 - 0.9462} \times 252} = 0.4229$$

The other parameters do not change much, and this unconditional volatility estimate of 42.29% is more reasonable. This example illustrates that a stochastic volatility model can be sensitive to extreme data points.

Figure 25.5 compares the GARCH volatility estimate with an EWMA estimate where $b = 0.94$. The two are not dramatically different, although the EWMA estimate exhibits more extreme behavior. An EWMA estimate with $b = 0.9462$ would be even closer to the GARCH estimate, because of the relation between GARCH and EWMA in equation (25.81). ■

Realized Quadratic Variation

We saw in Chapter 20 that the quadratic variation (the sum of squared increments) of a Brownian motion from 0 to T is T. That is, if you sample a Brownian process, $Z(t)$, very frequently (with h small), then if $n = T/h$,

$$\sum_{i=1}^{n} \left(Z[(i+1)h] - Z[ih]\right)^2 \approx T$$

Suppose stock returns are generated by the standard lognormal model and consider what happens if you compute the quadratic variation of the log stock price. In order to do this, you would need to observe the stock price at a high frequency (i.e., with h being small). For example, you would like to observe multiple prices over the course of the trading day. The quadratic variation of the log of the stock price is

$$\sum_{i=1}^{n} \left(\ln[S(t+h)] - \ln[S(t)]\right)^2 = \sum_{i=1}^{n} \left(\ln[S(t+h)/S(t)]\right)^2$$
$$= \sum_{i=1}^{n} \left[(\alpha - \delta - 0.5\sigma^2)h + \sigma[Z(t+h) - Z(t)]\right]^2 \approx \sigma^2 T$$

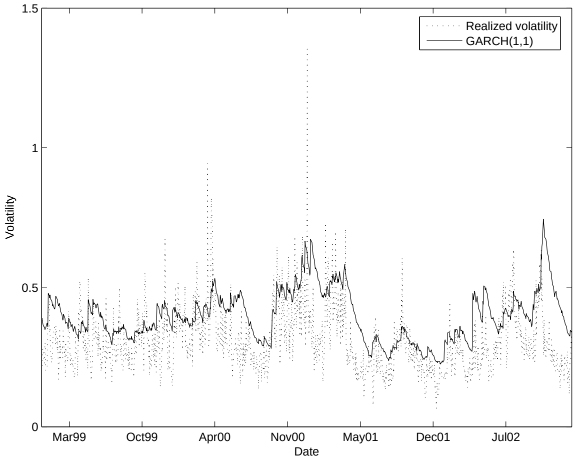
The left hand-side of this equation is the sum of squared, continuously compounded returns. The right-hand side is an estimate of variance. You can verify that as h gets smaller, the squared change in $Z(t)$ dominates the summation. (This is the same argument we used in Section 20.3.4 when discussing the multiplication rules for Itô processes.)

The sum of squared, continuously compounded returns, measured at high frequency, is the **realized quadratic variation** of the asset.[26] One well-known difficulty with using high-frequency data is that some observed price movements occur simply because transactions alternate between customer purchases (made at the dealer's offer price) and customer sales (made at the dealer's bid price). The resulting up and down movement in the price is called "bid-ask bounce".

Andersen *et al.* (2004) compare realized quadratic variation for currency prices with other variance estimates (such as GARCH(1,1)) and in the process

---

[26] Unfortunately, there does not appear to be a standard terminology for the realized quadratic variation of an asset price. It is common to call σ the "volatility". It would seem consistent to refer to the sum of squared returns as the "realized variance" and the "realized volatility" would then be the square root of the realized variance. In practice, however, the sum of squared returns is sometimes called the "realized volatility". See, for example, Andersen *et al.* (2004). Moreover, the realized quadratic variation only measures the variance of the stock price diffusion, σ^2, under certain regularity conditions, for example that there are no jumps. Thus, for clarity, we will use the term "realized quadratic variation," which is unambiguous, albeit clumsy.



25.6 Daily realized volatility for IBM, January, 1999 to December 2002, plotted against GARCH(1,1) volatility estimate. Source: Andersen *et al.* (2005)

they illustrate one way to measure realized volatility. At 30-minute intervals, they observe the bid and ask prices immediately preceding and following the 30-minute mark. They interpolate the averaged bid and ask prices to impute a price at 30 minute intervals. They then use this imputed price to measure the 30-minute continuously compounded return, from which they construct realized quadratic variation. In comparing forecasts based on realized quadratic variation with other methods of forecasting volatility, both in- and out-of-sample over one and ten-day horizons, they find that realized quadratic variation is generally at least as good as other estimates.

Figure 25.6 plots daily realized volatility for IBM against the GARCH(1,1) volatility estimate using the parameters computed in Example 25.2. As you might expect, the realized volatility series is considerably less smooth than the

GARCH estimate, with considerable outliers on specific days. In interepreting the realized volatility plot, it is important to recognize that if the price bounces around a great deal during the day, then realized volatility will be high even if the return for that day is not high. The empirical question is whether this intra-day bouncing around of the price predicts high volatility on subsequent days. The evidence in Andersen *et al.* (2004) is that it does.

25.3 HEDGING AND PRICING VOLATILITY

In this section we discuss derivative claims that have volatility as an underlying asset. We begin by discussing volatility and variance swaps (including one contract based on the VIX). We then look at an example of pricing a variance swap. Finally we discuss the construction of the history of the VIX volatility index reported by the Chicago Board Options Exchange (CBOE).

Throughout the section we will assume that the stock price follows the process

$$dS/S = \alpha dt + \sigma(S, X, t)dZ \tag{25.82}$$

This familiar looking equation is subtly different than in earlier chapters because we allow volatility to depend on the level of the stock price S, other variables, X, and time. Thus, volatility in equation (25.82) is stochastic. However, there are no jumps in equation (25.82), and we assume that the volatility function, $\sigma(S, X, t)$ does not move discontinuously. In this section we will let V denote measured volatility and V^2 measured variance.

Variance and Volatility Swaps

A **variance swap** is a forward contract that pays the difference between a forward price, $F_{0,T}(V^2)$, and some measure of the realized stock price variance over a period of time, times a notional amount. The payoff to a variance swap is

$$[\hat{V}^2 - F_{0,T}(V^2)] \times N$$

where N is the notional amount of the contract. There are numerous measurement details that we have to specify in order to write the contract for a variance swap:

- how frequently the return is measured
- whether returns are continuously compounded or arithmetic
- whether the variance is measured by subtracting the mean or by simply squaring the returns
- the period of time over which variance is measured
- how to handle days in which, unexpectedly, trading does not occur

Most of these design issues are straightforward, but the last deserves some comment. Most futures contracts settle based upon a final observable price. A variance contract, by contrast, settles based upon a *series* of prices. Therefore, failing to observe a price (for example, because the market is unexpectedly closed) creates a problem for measuring the realized variance. If the market is unexpectedly closed on day t, then the next measured return will be a *two-day* return, which will have a greater expected variance than a one day return. The following example shows how one contract deals with this issue.

Example 25.3 Three-month S&P 500 variance futures traded on the Chicago Futures exchange are an example of a variance swap. The payoff is based on squared, annualized, continuously compounded daily returns over a three-month period, $\hat{V}^2$. The measured price is quoted as $\hat{V}^2 \times 10,000$, and by definition a 0.01 change in this number is worth \$50.

For simplicity, we treat the payoff as if it were a forward contract, settling on one day. Let ϵ_i be the continuously compounded return on day i. The payoff at expiration is

$$\$50 \times \left[10,000 \times 252 \times \sum_{i=1}^{n_a-1} \frac{\epsilon_i^2}{n_e - 1} - F_{0,T}(V^2)\right]$$

In this formula, n_a is the actual number of S&P prices used in constructing V^2 (hence there are $n_a - 1$ returns), and n_e is the number of expected trading days at the outset of the contract. Thus, in the event of an unexpected trading halt, the sum of squared returns will be divided by a number larger than the number of squared returns. ■

A **volatility swap**, analogously to a variance swap, pays based on volatility differences. The payoff is

$$(\hat{V} - F_{0,T}(V)) \times N$$

where $F_{0,T}(V)$ is the forward price for volatility.

Example 25.4 The Chicago Board Options Exchange volatility index, the VIX, is the basis for a volatility futures contract that trades on the Chicago Futures Exchange. Unlike the variance futures contract, the volatility futures contract settles based upon the VIX index. The payoff is

$$1000 \times [\text{VIX}_T - F_{0,T}(V)]$$

■

In comparing the volatility futures contract to the variance futures contract, note that the two contracts are based on volatility measured over different periods of time. The variance contract settles based on realized quadratic variation over the period from 0 to T, and thus the futures price reflects volatility expectations from time 0 to time T. The VIX contract, since it is based on the VIX index, measures volatility expectations from time T to time $T + 30$ days. Thus, the volatility contract measures volatility going forward from the settlement date, while the variance contract looks backward from the settlement date.

There are at least two reasons that the variance contract is in some sense more "natural" than a volatility contract. First, we will see below that it is possible to price and hedge a variance forward contract (given some assumptions) using option prices. The pricing of volatility that a volatility forward contract is more complicated due to Jensen's inequality. Because the volatility is square root of the variance, Jensen's inequality implies that the volatility forward price will be less than the square root of the variance forward price of the variance.

Second, variance swaps arise naturally from dealers hedging their option positions. Recall from Chapter 13, in particular equation (13.11), that the profit of a delta-hedging dealer depends on the squared stock price change. Dealers can hedge this risk in realized variance by using variance swaps. For example a dealer with a negative gamma position could enter a swap that pays the dealer when the stock has a large price change. profit makes or loses money depending on the error in the dealer's hedging is related to the squared stock price change.

Pricing Volatility

We will see in this section one way to determine the fair price for a forward contract on variance.

Consider a variance contract that pays the sum of squared price changes from time 0 to time T. If price changes are measured over an interval of length

$h = T/n$, the contract would have the payoff

$$\text{Payoff} = \sum_{1}^{n} \left(\frac{S_{(i+1)h} - S_{ih}}{S_{ih}} \right)^2$$

Note that, since arithmetic and continuously compounded returns are close over small intervals, this is the same as the realized quadratic variation measure discussed in the previous section. As h gets small, equation (25.83) becomes

$$\text{Payoff} = \int_0^T \sigma(S_t, X_t, t)^2 dt \qquad (25.83)$$

where $\sigma(S, X, t)$ is the diffusion coefficient in equation (25.82). We want to answer two closely related questions. First, how is it possible to replicate the payoff to such a contract? Second, how should the contract be priced? As you might suspect, replicating the contract yields a way to price it.

In principle, the price of a forward contract on variance will be the expectation of equation (25.83) under the risk-neutral measure. In practice, how do we compute such an expectation? This section will provide one answer.

The Log Contract

Neuberger (1994) pointed out that a forward contract that pays

$$\ln(S_T/S_0) \qquad (25.84)$$

could be used to hedge and speculate on variance. A claim with the payoff in equation (25.84) is a **log contract**. As of early 2005, there is no exchange-traded log contract in existence, but for the moment suppose such a contract does exist.

Assuming the stock price follows equation (25.82), we can use Itô's Lemma to characterize the process followed by the log of the stock price. In using Itô's Lemma, we will implicitly assume that the stock price does not jump. Equation (25.82) permits a wide range of processes for the volatility, but the prospective volatility over the next instant is known. Applying Itô's Lemma, we have

$$d[\ln(S_t)] = \frac{1}{S}dS - 0.5\frac{1}{S^2}dS^2$$

Thus,

$$0.5\sigma^2 dt = \frac{1}{S}dS - d[\ln(S_t)]$$

Integrating this equation, we have

$$0.5\int_0^T \sigma^2(t)dt = \int_0^T \frac{1}{S}dS - \int_0^T d[\ln(S_t)]dt$$

The integral on the left-hand side is *realized quadratic variation* over 0 to T. Let $\hat{\sigma}^2$ denote annualized realized volatility from time 0 to time T. We can then rewrite this as

$$0.5T\hat{\sigma}^2 = \int_0^T \frac{1}{S}dS - \ln(S_T/S_0) \tag{25.85}$$

The right-hand side of equation (25.85) is the cumulative return to an investment in $1/S$ shares, less the return on a contract paying the realized continuously compounded return on the stock price from time 0 to time T. The left-hand side is annualized realized quadratic variation, which from equation (25.83) is also the payoff on a variance contract.

Take expectations of both sides of equation (25.85) with respect to the risk-neutral stock price distribution. The expectation of dS/S under the risk-neutral distribution is rdt. Hence we obtain

$$\begin{aligned} 0.5TE^{*}\left[\hat{\sigma}^{2}\right] &= E^{*}\left[\int_{0}^{T}\frac{1}{S}dS - \ln(S_T/S_0)\right] \\ &= rT - E^{*}\left[\ln(S_T/S_0)\right] \qquad (25.86) \end{aligned}$$

This expression seems to be of little help in pricing volatility. There is a trick, however, for pricing the log contract using other instruments.

Valuing the Log Contract

Demeterfi *et al.* (1999) and Carr and Madan (2002) independently showed that it is possible to use a portfolio of options to replicate the payoff on the log contract. Note first that

$$\int_a^b \frac{1}{K^2}(K - S_T)dK = \left[\ln(K) + \frac{S_T}{K}\right]_a^b = \ln(b) - \ln(a) + \frac{S_T}{b} - \frac{S_T}{a}$$

Use this to obtain the following identity, for any S_T (see Demeterfi *et al.* 1999):[27]

$$-\ln\left(\frac{S_T}{S_*}\right) = -\frac{S_T - S_*}{S_*} + \int_0^{S_*}\frac{1}{K^2}\max(K - S_T, 0)dK + \int_{S_*}^{\infty}\frac{1}{K^2}\max(S_T - K, 0)dK$$

---

[27] To interpret the expression on the right-hand side, notice that for any given S_T, the value of the first integral is zero for K below S_T, and the value of the second integral is zero for K above S_T. Thus, the effective integration bounds are not 0 and ∞, but rather depend upon S_T. For example, if $S_T = \bar{S} < K, S_*$, the equation becomes

$$-\ln\left(\frac{\bar{S}}{S_*}\right) = -\frac{\bar{S} - S_*}{S_*} + \int_{\bar{S}}^{S_*}\frac{1}{K^2}\max(K - \bar{S}, 0)dK + 0$$

Notice that if we take expectations of both sides with respect to the risk-neutral distribution for S_T, the integrals on the right-hand side become undiscounted option prices, and the expected stock price is the forward price. Thus, we obtain

$$-E^*\left[\ln\left(\frac{S_T}{S_*}\right)\right] = -\left[\frac{F_{0,T}-S_*}{S_*}\right] + e^{rT}\left[\int_0^{S_*}\frac{1}{K^2}P(K)dK + \int_{S_*}^{\infty}\frac{1}{K^2}C(K)dK\right]$$

Substitute this expression for $E^*(\ln(S_T)$ in equation (25.86). The result is

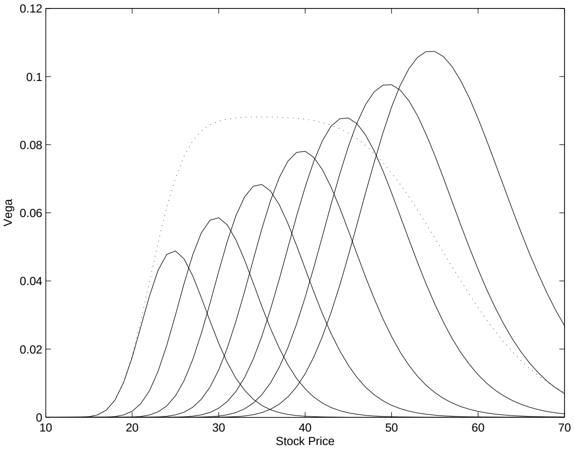
$$\begin{aligned}\hat{\sigma}^2 = \frac{2}{T}\Bigg[rT - \ln\left(\frac{S_*}{S_0}\right) - \frac{F_{0,T}-S_*}{S_*} \\ + e^{rT}\left(\int_0^{S_*}\frac{1}{K^2}P(K)dK + \int_{S_*}^{\infty}\frac{1}{K^2}C(K)dK\right)\Bigg] \qquad (25.87)\end{aligned}$$

Finally, note that if we set $S_* = F_{0,T}$, the first three terms on the right-hand side of equation (25.87) vanish, and we have:

$$\hat{\sigma}^2 = \frac{2e^{rT}}{T}\left[\int_0^{F_{0,T}}\frac{1}{K^2}P(K)dK + \int_{F_{0,T}}^{\infty}\frac{1}{K^2}C(K)dK\right] \qquad (25.88)$$

Remarkably, this formula gives us an estimate of expected realized variance that we compute using the observed prices of out of the money puts and out of the money calls! ("Out of the money" here is with respect to the forward price rather than the current stock price). It is important to note that we have *not* assumed that options are priced using the Black-Scholes formula or any other specific model.

One important characteristic of equation (25.88) is that the variance estimate can be replicated by trading options. It is possible to buy the strip of out-of-the-money puts and calls, weighted by the inverse squared strike price, to



25.7 Solid lines depict vegas of options with (from left to right) strikes of 25, 30, 35, 40, 45, 50, and 55. The dashed line is the weighted sum of the vegas, with each divided by the squared strike price, times 600. The calculations assume $\sigma = 0.30$, $r = 0.08$, $T = 0.25$, and $\delta = 0$.

create a portfolio that has the value of $\hat{\sigma}^2$.

To get a sense of why equation (25.88) works, we can examine the vega of a portfolio of options held in proportion to $1/K^2$. Figure 25.7 graphs vegas for a set of options and also displays the vega of a portfolio where the option holdings are weighted by the inverse squared strike price. The resulting portfolio has a vega that is not zero and is constant over a wide range of stock prices. If you hold such a portfolio, you make or lose money depending on volatility changes, but not depending on stock price changes. Note that while this intuitive explanation uses the Black-Scholes formula, the derivation of equation (25.87) simply assumes that there are no jumps and that $\sigma(S, X, t)$ is "well-behaved".

Computing the VIX

We can now explain the formula used to compute the CBOE's new volatility index. The calculation is based on equation (25.88). In practice option strike prices are discrete and there may be no option for which the strike price equals the index forward price. The actual formula used by the CBOE to compute is a discrete approximation to equation (25.88):

$$\hat{\sigma}^2 = \frac{2}{T} \sum_{K_i \le K_0} \frac{\Delta K_i}{K_i^2} e^{rT} \text{Put}(K_i) + \frac{2}{T} \sum_{K_i > K_0} \frac{\Delta K_i}{K_i^2} e^{rT} \text{Call}(K_i) - \frac{1}{T} \left[\frac{F_{0,T}}{K_0} - 1 \right]^2 \tag{25.89}$$

where K_0 is the first strike below the forward price for the index. The last term is a correction for the fact that there may be no option with a strike equal to the forward price.

25.4 EXTENDING THE BLACK-SCHOLES MODEL

In this section we examine three pricing models that are capable of generating volatility skew patterns resembling those observed in option markets. The goal is both to understand how the Black-Scholes model can be extended, and also to gain a sense of how these extensions help to better understand the data. The three models we consider are the Merton jump diffusion model, which relaxes the assumption that stock price moves are continuous; the constant-elasticity of variance model, primarily due to Cox, which relaxes the assumption that volatility is constant; and the Heston stochastic volatility model, which allows volatility to follow an Itô process that is correlated with the stock price process. These models have all been significantly generalized, but we can use them as touchstones for better understanding the economics of departures from the Black-Scholes lognormality assumption.

At the outset, note that the Black-Scholes model easily accommodates time-varying volatility if the volatility pattern is deterministic. Specifically, Merton (1973) showed that if volatility is a deterministic function of time, then it is possible to price a European option with $T - t$ periods to maturity by substituting $\int_t^T \sigma^2(s)ds$ for $\sigma^2(T - t)$ in the Black-Scholes formula. We can think about this result in the context of a delta-hedging market-maker. As long as the market-maker knows the volatility at each point in time, the delta-hedge will work the same as if volatility were constant. What creates a problem is a *random* change in volatility.

Jump Risk and Implied Volatility

In chapter 21 we presented Merton's valuation formula for an option when the underlying asset can jump. In this section we consider the special case of that model in which the stock price can jump only to zero (see equation 21.61); we show that the jump model generates a volatility skew. The intuition in this case is particularly clear.

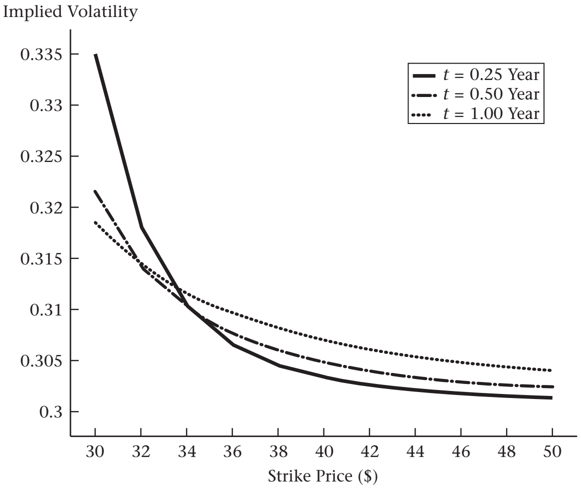
Suppose that the stock can jump to zero with 0.5% probability per year. If we let $S = \$40$, $K = \$40$, $\sigma = 30\%$, $r = 8\%$, $T - t = 0.25$, and $\delta = 0$, then

Option prices when the stock can and cannot jump. Assumes $S = \$40$, $\sigma = 30\%$, $r = 8\%$, $\delta = 0$, $\lambda = 0.5\%$, and $T - t = 0.25$.

Strike ($)	**Call Price** **Jump**	 **No Jump**	**Difference ($)**	**Call Vega**	**Jump-Implied σ**
40	2.8104	2.7847	0.0257	0.0781	0.303
35	6.1704	6.1348	0.0356	0.0436	0.308
30	10.6679	10.6320	0.0359	0.0083	0.334

we have call and put prices of \$2.81 and \$2.02, compared to the no-jump prices of \$2.78 and \$1.99. Now we do the following experiment: Generate "correct" option prices, i.e., prices properly accounting for the jump, for a variety of strikes and different times to maturity. We then ask what implied volatility we would compute for these options using the ordinary Black-Scholes formula. Table 25.2 shows the jump and no-jump prices for options at three different strike prices, along with the option vegas. The results are also graphed in Figure 25.8. Because of parity, puts and calls have the same implied volatility, so we need graph only one of them.

In every case, out-of-the money puts (in-the-money calls) have higher implied volatilities than at-the-money options. We can see why this is happening by examining the numbers more closely. The small possibility of a jump causes all the option prices to increase about 2.5–3.5 cents. The standard implied volatility calculation uses the Black-Scholes formula without a jump adjustment to compute the implied volatility. The no-jump prices all have implied volatilities of 30%. What volatility in the no-jump model is required



25.8 Implied volatilities computed using the Black-Scholes formula for three different times to maturity and different strike prices. The underlying option prices are generated when the stock can jump to zero with probability 0.5%/year. Assumes $S = \$40$, $\sigma = 30\%$, $r = 8\%$, $\delta = 0$.

to generate the prices in the jump column? For the 40-strike option, vega is 0.0781, so a change in volatility of approximately $0.0257/0.0781 = 0.323$ percentage points, or a volatility of 0.30323, will generate the higher price of \$2.81. For the 30-strike option, however, vega is only 0.0083. Thus, a change in volatility of approximately $0.0359/0.0083 = 4.3$ percentage points is required in order for the no-jump Black-Scholes model to explain the price of \$10.6679. The actual implied volatility in this case is 0.334. When vega is lower, a larger change in volatility is required to explain a given change in the option price.

This example is at most suggestive. In practice, jumps can be positive or negative and of uncertain magnitude. If jumps can occur in both directions, then we would expect to see higher implied volatilities for both in-the-money and out-of-the-money options. Furthermore, jump risk is unlikely to be purely diversifiable since there can be market-wide moves. The example does, however, illustrate important intuition for why jumps can generate volatility smiles.

Constant Elasticity of Variance

Cox (1975) proposed the constant elasticity of variance (CEV) model, in which volatility varies with the level of the stock price. Specifically, Cox proposed that the stock follow the process

$$dS = (\alpha - \delta)dt + \bar{\sigma} S^{\beta/2} dZ \tag{25.90}$$

Equation (25.90) describes the instantaneous dollar return on the stock. The instantaneous rate of return on the stock is

$$\frac{dS}{S} = (\alpha - \delta)dt + \bar{\sigma} S^{(\beta-2)/2} dZ \tag{25.91}$$

The instantaneous standard deviation of the stock return is therefore

$$\sigma(S) = \bar{\sigma} S^{(\beta-2)/2} \tag{25.92}$$

When $\beta > 2$, the CEV model implies that volatility is increasing with the stock price. Volatility decreases with the stock price when $\beta < 2$. When $\beta = 2$, the CEV model yields the standard lognormal process.

It is important to be clear that $\bar{\sigma}$ is a parameter that determines volatility, but the instantaneous rate of return volatility is $\bar{\sigma}S^{(\beta-2)/2}$. Thus, if we want the stock to have a volatility of σ_0 at the current stock price, S_0, then we must set $\bar{\sigma}$ so that $\sigma_0 = \bar{\sigma}S^{(\beta-2)/2}$, or

$$\bar{\sigma} = \sigma_0 S^{(2-\beta)/2}$$

From equation (25.90), the elasticity of the instantaneous stock price variance with respect to the stock price is a constant, β:

$$\frac{\partial(\bar{\sigma}^2 S^\beta)}{\partial S} \times \frac{S}{\bar{\sigma}^2 S^\beta} = \beta$$

This is where the name "constant elasticity of variance" comes from.

The CEV Pricing Formula

There is a relatively simple pricing formula for a European call when the stock price follows the CEV process.[28] Following Schroder (1989), define

$$\kappa = \frac{2(r-\delta)}{\bar{\sigma}^2(2-\beta)\left(e^{(r-\delta)(2-\beta)T}-1\right)}$$

$$x = \kappa S^{2-\beta} e^{(r-\delta)(2-\beta)*T}$$

$$y = \kappa K^{2-\beta}$$

---

[28] Cox (1996) originally derived a pricing formula in terms of infinite series for the case $\beta < 2$. Emanuel and MacBeth (1982) generalized Cox's analysis to the case where $\beta > 2$. Schroder (1989) showed that both cases could be expressed more compactly in terms of the non-central chi-squared cumulative distribution function. Davydov and Linetsky (2001) derive pricing formulas barrier and lookback options under a CEV process.

The CEV pricing formula for a European call is different for the cases $\beta < 2$ and $\beta > 2$. Let $Q(a, b, c)$ denote the non-central chi-squared distribution function with b degrees of freedom and non-centrality parameter c, evaluted at a.[29] The CEV call price is given by

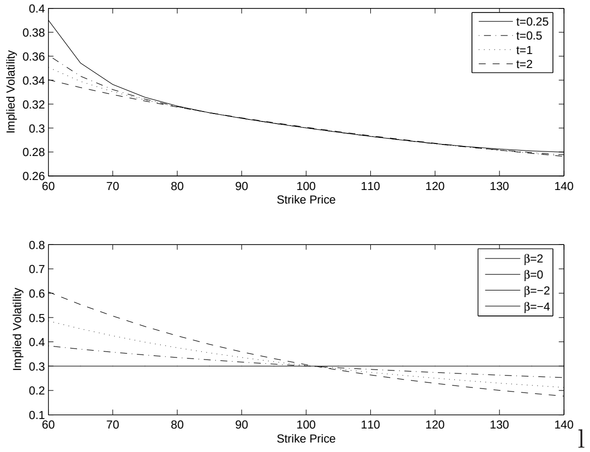
$$\begin{array}{lr} Se^{-\delta T}\left[1 - Q(2y, 2 + 2/(2-\beta), 2x)\right] - Ke^{-rT}Q(2x, 2/(2-\beta), 2y) & \beta < 2 \\ Se^{-\delta T}\left[1 - Q(2x, 2/(\beta-2), 2y)\right] - Ke^{-rT}Q(2y, 2 + 2/(\beta-2), 2x) & \beta > 2 \end{array} \tag{25.93}$$

Implied Volatility in the CEV Model

When $\beta < 2$, the CEV model generates a Black-Scholes implied volatility skew curve resembling that in Figure 25.1: Implied volatility decreases with the option strike price. To understand why the CEV model generates this volatility skew, note from equation (25.91) that when $\beta < 2$ and the stock price falls, volatility increases. Thus compared to the case of a constant volatility, an out-of-the-money put option has a greater chance of exercise and is likely to be deeper in the money when it is exercised. The only way for the Black-Scholes model to account for this higher price is with a higher volatility. As the strike price increases, less of the option value is due to the stock price behavior at low prices, and volatility therefore need not increase as much.

---

[29] The pricing formula is sometimes written in terms of the *complementary* non-central chi-squared distribution, which is $1 - Q(a, b, c)$. The non-central chi-squared distribution, unlike the chi-squared distribution, is not typically a standard function built into spreadsheets. However, it is available in programs such as Matlab and Mathematica.



25.9 Implied volatility in the CEV model. Both panels assume that $S = \$100$, $\sigma_0 = 0.30$, $r = 0.08$, $T = 1$. In the top panel, $\beta = 1$, while in the bottom panel, $T = 0.5$.

Figure 25.9 plots implied volatility curves generated by using the Black-Scholes formula to compute implied volatility for prices generated by the CEV model. The top panel shows the effect of time to maturity on the implied volatility curve. For a given strike price, volatility skew is less pronounced for longer times to maturity.

The Heston Model

In the CEV model, the instantaneous volatility of the stock evolves stochastically with the stock price, but volatility is a non-stochastic function of the stock price. Since in the CEV model there is only one stochastic process affecting the option price, market-makers can hedge option positions using only the stock.

A more general approach is to permit volatility to follow a stochastic process. The Heston model (see Heston 1993) allows volatility to vary stochastically, but still to be correlated with the stock.[30] Let $v(t)$ be the instantaneous stock return variance; hence, $\sqrt{v(t)}$ is the volatility. Suppose that the stock follows the process

$$\frac{dS}{S} = (\alpha - \delta)dt + \sqrt{v(t)}dZ_1 \tag{25.94}$$

Assume that the variance, $v(t)$, follows the mean-reverting process

$$dv(t) = \kappa[\bar{v} - v(t)]dt + \sigma_v\sqrt{v(t)}dZ_2 \tag{25.95}$$

We assume that $E(dZ_1 dZ_2) = \rho dt$.

The interpretation of equations (25.94) and (25.95) is familiar. Equation (25.94) for the stock is the same as equation (21.5) except that the volatility, $\sqrt{v(t)}$, is random. The equation for volatility, equation (25.95), has two noteworthy characteristics. First, the instantaneous variance, $v(t)$, is mean-reverting, tending toward the value $\bar{v}$, with a speed of adjustment given by κ. Second, the volatility of variance, $\sigma_v\sqrt{v(t)}$, depends on the square root of $v(t)$, and variance is therefore said to follow a *square root process.*

Suppose that the risk premium for the risk $\sigma_v\sqrt{v(t)}dZ_2$ can be written as $v(t)\beta_v$, where we assume β_v is constant. This assumption that the risk premium is proportional to the level of the variance is analytically convenient. Given this

---

[30] Earlier papers that modelled volatility as following a stochastic process included Hull and White (1987), Scott (1987), and Wiggins (1987). The Heston model has been generalized significantly by Duffie *et al.* (2000), who allow both jumps in the asset price and jumps in volatility.

assumption about the risk premium, the risk-neutral volatility process is

$$\begin{aligned} dv(t) &= \{\kappa[\bar{v} - v(t)] - v(t)\beta_v\}\, dt + \sigma_v \sqrt{v(t)} dZ_2^* \\ &= \kappa^* \left[\bar{v}^* - v(t)\right] + \sigma_v \sqrt{v(t)} dZ_2^* \end{aligned} \tag{25.96}$$

where $\kappa^* = \kappa + \beta_v$ and $\bar{v}^* = \bar{v}\kappa/(\kappa + \beta_v)$. This model of stochastic volatility is called the **Heston model** (Heston, 1993).

Let $V[S(t), v(t), t]$ represent the price of a derivative on the stock when the stock price and volatility are given by equations (25.94) and (25.95). Suppose we proceed with the Black-Scholes derivation, in which we hold the option and try to hedge the resulting risk. We immediately encounter the problem that there are *two* sources of risk, dZ_1 and dZ_2. A position in the stock will hedge dZ_1, but what can we use to hedge risk resulting from stochastic volatility? Apart from other options, there will typically be no asset that is a perfect hedge for volatility.[31] In that case, we rely on the equilibrium approach to pricing the option. The PDE for the derivative $V[S(t), v(t), t]$ is then:

$$\begin{aligned} &\tfrac{1}{2}v(t)S^2V_{SS} + \tfrac{1}{2}\sigma_v^2 v(t)V_{vv} + \rho v(t)\sigma_v S V_{Sv} \\ &+ (r - \delta)SV_S + \{\kappa[\bar{v} - v(t)] - v(t)\beta_v\}\, V_v + V_t = rV \end{aligned} \tag{25.97}$$

This equation is the multivariate Black-Scholes equation, described in Appendix 21.A. The third term is due to the covariance between the stock return and volatility. Since there is no asset to hedge volatility, the coefficient on the V_v

---

[31] It might be possible to use other options on the same stock to hedge volatility, but the option would then be priced *relative* to the price of the option used as a hedge.

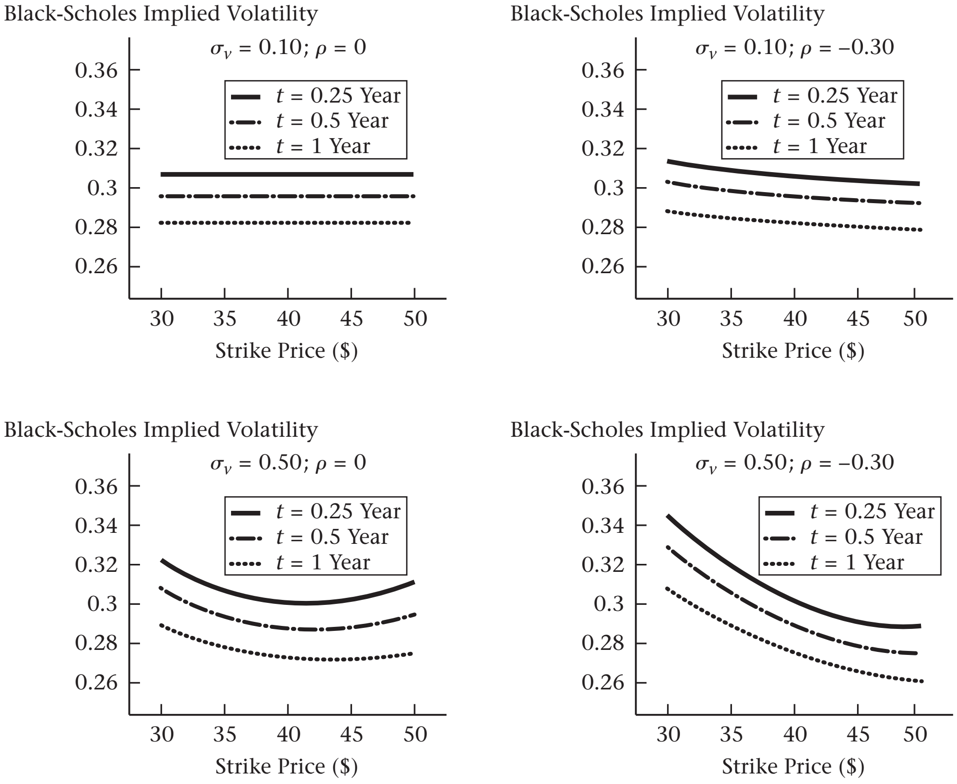
term has a correction for the risk premium associated with volatility.

Heston (1993) shows that equation (25.97) has an integral solution that can be evaluated numerically. Given this solution, we can see how implied volatility behaves when volatility is stochastic. Similar to the analysis of jumps in Section 21.2, we price options for different strikes and expirations under the stochastic volatility model, and then use Black-Scholes to compute implied volatilities. We assume that the stock price is \$40, and compute implied volatilities for options with strike prices ranging from \$30 to \$50, and with maturities from 3 months to 1 year.

Figure 25.10 shows the result of this experiment for two different values of σ_v and ρ. In the figure the long-run volatility, $\bar{v}^*$, is 25%, less than the current volatility, 32%. Because volatility reverts to the mean, implied volatility decreases with time to maturity in every case. In the panel where $\sigma_v = 0.10$ and $\rho = 0$, there is almost no skew, although the mean reversion in volatility is apparent. When $\sigma_v = 50\%$ and $\rho = 0$, the figure exhibits both symmetric skew and mean reversion. The asymmetric skew in both right-hand panels of Figure 25.10 arises from assuming a negative correlation between volatility and the stock price.

Evidence

The challenge for an option pricing model is to match the observed volatility



25.10 Implied volatilities computed using the Black-Scholes formula when prices are computed using the Heston model for three different times to maturity, different strike prices, and two different volatilities of volatility, σ_v. The top panel assumes that $\sigma_v = 0.10$, while the bottom panel assumes that $\sigma_v = 0.5$. In both panels, $\kappa^* = 2.0$, $v(t) = 0.32$, $\bar{v}^* = 0.25$, $r = 8\%$, and $\delta = 0$.

skew.[32] The literature investigating ways to do this is too large to adequately summarize here. Instead, we will sketch the nature of findings in the literature

---

[32] The true model should give equal implied volatilities for options at different strikes and maturities. For example, if the stochastic volatility model were true and option prices were consistent with equation (25.97), then Black-Scholes implied volatilities would exhibit skew, but if the Heston model were used to compute implied volatility, then the options in Figure 25.10 would all have implied volatilities of 32%.

and highlight issues that arise when trying to match models to data.

The pricing models in this section illustrate ways in which modifying the Black-Scholes assumptions can enable a pricing model to better fit observed option prices. For example, all the pricing models we have discussed are capable of generating higher implied volatilities for in-the-money (low-strike) calls. The Merton jump model and the CEV model in the examples above both generate implied volatility curves that are flatter as time to maturity increases. Combinations of the models, such as a Heston model that also allows jumps, seem able to reproduce *qualitative* features of Figure 25.1. However, matching models to data is a more involved exercise than just a visual comparison of implied volatility curves.

To illustrate the issues, suppose you want to match the Heston stochastic volatility model to data. There are a number of ways you might proceed. First, on a given day, you could find a set of model inputs that best matches the volatility curves for that day. This entails finding a return variance ($v(t)$), a volatility of volatility (σ_v), a mean reversion rate (κ^*), a long-run risk-neutral variance ($\bar{v}^*$), and a correlation between volatility and the stock return (ρ), that match the data for a particular day.

Matching implied volatilities across a set of options on a given day is a *cross-sectional* test of the model. Once you admit multiple days of data, the model has *time-series* implications as well. Equations (25.95) and (25.96) imply that volatility evolves over time in a specific way. If you look at the evolution of volatility over time, does it match equation (25.95)? Are the parameters that enable the model to fit the cross-section consistent with those implied by the volatility time series? When there is a risk premium in the equilibrium pricing model (as in the Heston model), it is potentially easier to reconcile the behavior of the stock with option prices because there is an additional parameter. However, as Bates (2003) emphasizes, a risk premium must be plausible.

Bakshi *et al.* (1997) and Bates (2000) both asked whether option pricing models incorporating jumps and stochastic volatility can generate realistic volatility skew for options based on the S&P 500.[33] Both studies find greater volatility skew at short maturities than at long maturities. If you compare Figures 25.8 and 25.10, you can see that this pattern is generated by the jump model but not as obviously by the Heston model. This explains why, although Bakshi *et al.* (1997) found that the stochastic volatility model provided the best overall explanation of prices, they added jumps to account for skew at short maturities.[34] They also found that permitting stochastic interest rates (which can be added in the same fashion as stochastic volatility) helped explain prices at longer maturities.

Bates (2000) also found that jump models (as in Figure 25.8) fit near-term option prices better but found the jump parameters implausible: the stock price

---

[33] Bakshi *et al.* (1997) examined European options on the S&P 500 index while Bates (2000) examined options on the S&P 500 futures contract.

[34] Carr and Wu (2003) formalize the intuition that jumps matter more if there is a short time to expiration.

does not appear to jump as often as implied by the estimates necessary to explain implied volatility. Bates also concluded that in order for the stochastic volatility model to explain skew, the volatility of volatility had to be implausibly large. However, Duffie *et al.* (2000) developed a pricing procedure which permitted jumps in both the asset price and the volatility, and noted that allowing jumps in volatility potentially addressed the problem of an implausibly large volatility of volatility.

Broadie *et al.* (2004) conclude that "models with jumps in returns drastically improve overall pricing performance ... [and] jumps in volatility offer a significant pricing improvement in the cross-section unless a model with jumps in returns is allowed to have a volatility-of-jump risk premium." In other words, jumps in at least some dimensions are important, and risk premia can be important as well.[35]

To add one more layer of complication, casual observation suggests that in some cases volatility changes deterministically over time. When a firm announces earnings, for example, volatility will be higher than on ordinary days. You can show that this is true by comparing the volatility of returns on earnings announcement days against that on other days. Dubinsky and Johannes (2004)

---

[35] Another way to assess risk premia is to look at the returns on zero-delta positions. Coval and Shumway (2001) find negative returns from zero-delta written straddles on the S&P 100 and 500 index. Since the written straddle loses money when volatility increases, this finding may be at least in part attributable to a volatility risk premium. Bakshi and Kapadia (2003) find smaller risk premia associated with delta-hedged individual stocks than with index options.

show that this effect is also apparent in option prices, which imply a higher volatility before an earnings announcement than after.[36] This finding suggests that in addition to the use of increasingly sophisticated mathematical pricing models, careful option pricing will require data sets that identify *anticipated* days of unusual volatility.[37]

CHAPTER SUMMARY

For options on a given underlying asset on a given day, implied volatility generally varies across option strikes and across maturities. Implied volatility also varies over time. As a result there is great interest in measuring volatilities and in pricing options when volatility can vary.

Methods of measuring volatility using past data include historical volatility, exponentially-weighted moving average volatility, ARCH, GARCH, and realized quadratic variation. ARCH and GARCH estimates are based upon a formal statistical model in which volatility is random. Realized quadratic variation exploits high frequency data to obtain a reliable volatility estimate using data from a short time horizon.

---

[36] In fact, we saw the effect of earnings announcements in Example 25.2 when we estimated a GARCH(1,1) model for IBM and found the estimates sensitive to the inclusion of four earnings announcement days.

[37] This is not just an issue for individual firms. Governments make economic announcements on pre-specified days at set times, and these announcements sometimes generate large moves in prices. For example, Hanweck (1994) shows that implied volatility in Eurodollar futures options is greater on days when the government announces aggregate employment.

Both variance and volatility swaps permit hedging and speculation on volatility. The variance forward price can be obtained as a weighted sum of the prices of traded European options, a calculation which is the basis for the VIX measure of implied volatility.

The Black-Scholes model does not perfectly explain observed option prices; there is volatility skew, which means that implied volatility varies with the strike price and time to expiration. Two modifications to the model are to permit jumps in the stock price and to allow volatility to be stochastic. Both changes generate option prices that exhibit volatility skew and that better fit the data than the unmodified Black-Scholes model.

Attempts to explain prices of traded options suggest that it is important to account for jumps in both the asset price and volatility, and that risk premiums on one or both jumps may be important.

FURTHER READING

Early studies of stock returns (e.g., Fama 1965) found that continuously compounded returns exhibit too many large returns to be consistent with normality. In recent years, research has focused on specifying stock price processes that give theoretical option prices consistent with observed prices. Introductions to GARCH models include Royal Swedish Academy of Sciences (2003), and Bollerslev *et al.* (1994). Alexander (2001) is a readable text for less technical readers. Realized quadratic variation as a measure of volatility is presented and applied in Andersen *et al.* (2004).

Demeterfi *et al.* (1999) is a clear and well-written discussion of volatility hedging, and the paper also develops the volatility measure used now to construct the VIX. See also Chicago Board Options Exchange (2003).

The first papers to suggest alternative assumptions about the stock price for option pricing were Cox and Ross (1976), Cox (1996), and Merton (1976). Merton noted in his paper that the jump model had the potential to explain volatility skew patterns noted by practitioners at the time. The first stochastic volatility models were proposed by Hull and White (1987), Scott (1987), and Wiggins (1987). The Heston (1993) model has been generalized by Duffie *et al.* (2000), who develop a pricing framework that can accomodate jumps in volatility as well as jumps in the stock price.

The empirical literature examining the ability of option pricing models to fit observed prices is rapidly evolving. Well-known papers include Bakshi *et al.* (1997), Bates (2000), and Pan (2002). Current research (which include citations to numerous other papers) include Andersen *et al.* (2005) and Broadie *et al.* (2004). Dubinsky and Johannes (2004) examine deterministic volatility changes, for example due to earnings announcements.

For the first eight problems you will need to use data on the CD accompanying the book.

1. Using weekly price data (constructed Wednesday to Wednesday), compute historical *annual* volatilities for IBM, Xerox, and the S&P 500 index for 1991 through 2004. Annualize your answer by multiplying by $\sqrt{52}$. Also compute volatility for each for the entre period.
2. Compute daily volatilities for 1991 through 2004 for IBM, Xerox, and the S&P 500 index. Annualize by multiplying by $\sqrt{252}$. How do your answers compare to those in Problem 1?
3. For the period 1999-2004, using daily data:
   - Compute an EWMA estimate, with $b = .95$, of IBM's volatility using all data.
   - Compute an EWMA estimate, with $b = .95$, of IBM's volatility, at each date using only the previous 60 days of data.

   Plot both estimates. How different are they?
4. Replicate the GARCH(1,1) estimation in Example 25.2, using daily returns from on IBM from January 1999 to December 2003. Compare your estimates with and without the four largest returns.
5. Estimate a GARCH(1,1) for the S&P 500 index, using data from January 1999 to December 2003.

6. Using the average of the bid and ask option prices for IBM on the CD, compute implied volatility curves for the nearest-term options, using the average of the bid and ask prices. Also compute implied volatility for options with 2 years to expiration. Compare the results. For which options are the implied Black-Scholes volatilities most different for the calls and puts? Why does this occur?

7. Using the average of the bid and ask option prices for the S&P 500 index on the CD, compute implied volatility curves for the nearest-term options, using the average of the bid and ask prices. Also compute implied volatility for the options with the longest term to expiration. Compare the results. Are implied Black-Scholes volatilities different for the calls and puts? What would you expect to see?

8. Using the IBM option prices, calculate implied volatilities separately for the bid and ask prices and for the calls and puts, for the nearest term options and the longest term options.

How sensitive are the implied volatilities to estimation using the bid or ask prices? For which options is the difference greatest? Why?

9. Suppose $S = \$100$, $r = 8\%$, $\sigma = 30\%$, $T = 1$, and $\delta = 0$. Use the Black-Scholes formula to generate call and put prices with the strikes ranging from \$40 to \$250, with increments of \$5. Compute the implied volatility from these prices by using the formula for the VIX (equation (25.89)). What happens to your estimate if you use strikes that differ by \$1 or \$10, or strikes that range only from \$60 to \$200?

10. Explain why the VIX formula in equation (25.89) overestimates implied volatility if options are American.

The following three problems use the Merton jump formula. As a base case, assume $S = \$100$, $r = 8\%$, $\sigma = 30\%$, $T = 1$, and $\delta = 0$. Also assume that $\lambda = 0.02$, $\alpha_J = -0.20$ and $\sigma_J = 0.30$.

11. Using the Merton jump formula, generate an implied volatility plot for $K = 50, 55, ...150$.

- How is the implied volatility plot affected by changing α_J to -0.40 or -0.10?
- How is the implied volatility plot affected by changing λ to 0.01 or 0.05?
- How is implied volatility plot affected by changing σ_J to 0.10 or 0.50?

12. Using the base case parameters, plot the implied volatility curve you obtain for the base case against that for the case where there is a jump to zero, with the same λ.

13. Repeat problem 11, except let $\alpha_J = 0.20$, and in part (b) consider expected alternate jump magnitudes of 0.10 and 0.50.

The following two problems both use the CEV option pricing formula. Assume in both that $S = \$100$, $r = 8\%$, $\sigma_0 = 30\%$, $T = 1$, and $\delta = 0$.

14. Using the CEV option pricing model, set $\beta = 1$ and generate option prices for strikes from 60 to 140, in increments of 5, for times to maturity of 0.25, 0.5, 1.0, and 2.0. Plot the resulting implied volatilties. (This should reproduce Figure 25.9.)

15. Using the CEV option pricing model, set $\beta = 3$ and generate option prices for strikes from 60 to 140, in increments of 5, for times to maturity of 0.25, 0.5, 1.0, and 2.0. Plot the resulting implied volatilties.

APPENDIX

Here is one way to set up a spreadsheet in order to estimate a GARCH model by maximum likelihood using Excel.

1. Enter daily prices in column B, beginning in B10.
2. Compute continuously compounded returns for a 5-year period in column C, beginning in row 11. Leave cells a1:c7 empty.
3. In column D compute the squared continuously-compounded return. This will be ϵ^2.
4. In cell E11, enter the variance of the continuously-compounded returns. This will be your starting value for q.
5. In cell E12, enter the formula "=$B$1+$B$2*D12+$B$3*E11". Be sure to pay attention to which cells are absolute and which are relative references.

Copy this formula down the length of your data.

6. In cell F13, enter the formula $= -\ln(E13) - D13/E13$. Copy this formula down. This is your log likelihood function for each observation.

7. Suppose that your last return is in row 1200. In cell B4 enter the formula = SUM(F13:F1200). This is the log likelihood function for your data.

8. In Solver set up the following constraints: $B1 \geq 0.0000001$, $B2 \geq 0.00000001$, $B3 > 0$, $B2 \leq 0.99999999 - B3$.

9. In cell B5 enter the formula B1/(1-B2-B3). This is your unconditional variance estimate.

10. In cell B6 enter the formula SQRT(B5*252). This is your unconditional annualized standard deviation.

11. Set up Solver to maximize cell B4 (the likelihood) by varying cells B1:B3 (the parameters).

12. Solve!

Your solution will likely be quite sensitive: to starting values, to the Solver options, and to unusually large squared returns. You should change the tolerance (in Solver options) to 1% or less. You should also experiment with different starting values for the parameters.