Solutions to selected problems from Brooks (2019)

Chapter 6 (p. 381–386): Autocorrelation, stationarity, ARMA models.

6.3 (a) The first two models are examples of AR(1) models, with AR parameter ϕ_1 equal to 1 (random walk) and 0.5 (stationary AR(1) process), respectively. The third model is an MA(1).

(b) For $|\phi_1| < 1$ the acf of the AR(1) model decays geometrically (it is equal to $\tau_k = \phi_1^k$). In the case $\phi_1 = 1$ the theoretical acf does not even exist (in that case the time series process is a random walk, which is nonstationary). However, if you would nevertheless estimate the acf, you would find that the estimator of τ_k decays only very slowly with k, and seemingly linearly. For an MA(1), the ACF drops to zero after the first lag.

(c) The random walk is most reasonable as a model for stock prices or stock indices. For $\phi < 1$ price changes would be predictable. However, traders exploiting this would drive the price up to the point where the price change would be unpredictable again. Hence in practice only very little (or no) predictability should be expected.

6.6 Note that Brooks uses a slightly different formula for the criteria: he deletes the $1 + \log(2\pi)$ term, and doesn't count the variance as a parameter, i.e., he uses $k = p + q + 1$ instead of the correct $k = p + q + 2$. Using his definition, we end up with the following values for each criterion and for each model order, with an asterisk denoting the smallest value of the information criterion in each case. SBIC stands for Schwarz Bayesian Criterion, also known as SIC or BIC.

ARMA order (p, q)	$k = p + q + 1$	$\log \hat{\sigma}^2$	AIC	SBIC
$(0, 0)$	1	0.932	0.942	0.944
$(1, 0)$	2	0.864	0.884	0.887
$(0, 1)$	2	0.902	0.922	0.925
$(1, 1)$	3	0.836	0.866	0.870
$(2, 1)$	4	0.801	0.841	0.847
$(1, 2)$	4	0.821	0.861	0.867
$(2, 2)$	5	0.789	0.839	0.846
$(3, 2)$	6	0.773	0.833^*	0.842^*
$(2, 3)$	6	0.782	0.842	0.851
$(3, 3)$	7	0.764	0.834	0.844

The result is pretty clear: both SBIC and AIC say that the appropriate model is an ARMA$(3, 2)$.

In general, the AIC and SBIC need not agree on the optimal model. Depending on the aim of the modeller, either the AIC or SBIC can be used. For constructing a model for prediction (where the aim is a small forecast error) the AIC is more suitable. The SBIC should be used when you want to obtain the true ARMA(p, q) orders of the process that generated the data. The assumption that the data are from a finite order ARMA(p, q) model is a rather strong one to start with, so in most applications recovering the true model order is not the primary goal of the modelling exercise.

6.7 We could perform the Ljung-Box test on the residuals of the estimated models to see if there was any linear dependence left unaccounted for by our postulated models.

Another test of the models' adequacy that we could use is to leave out some of the observations at the identification and estimation stage, and attempt to construct out of sample forecasts for these. For example, if we have 2000 observations, we may use only 1800 of them to identify and estimate the models, and leave the remaining 200 for construction of forecasts. We would then prefer the model that gave the most accurate forecasts.

6.8 This is not true in general. Yes, we do want to form a model which "fit" the data as well as possible. But in most financial series, there is a substantial amount of "noise". This can be interpreted as a number of random events that are unlikely to be repeated in any forecastable way. We want to fit a model to the data which will be able to "generalise". In other words, we want a model which fits to features of the data which will be replicated in future; we do not want to fit to sample-specific noise.

This is why we need the concept of "parsimony" — fitting the smallest possible model to the data. Otherwise we may get a great fit to the data in sample, but any use of the model for forecasts could yield terrible results.

Another important point is that the larger the number of estimated parameters (i.e. the more variables we have), then the smaller will be the number of degrees of freedom, and this will imply that coefficient standard errors will be larger than they would otherwise have been. This could lead to a loss of power in hypothesis tests, and variables that would otherwise have been significant are now insignificant.

6.9 (a) We class an autocorrelation coefficient or partial autocorrelation coefficient as significant if it exceeds $\pm 1.96\frac{1}{\sqrt{T}} = 0.196$ (Bartlett bands). Under this rule, the sample autocorrelation functions (sacfs) at lag 1 and 4 are significant, and the spacfs at lag 1, 2, 3, 4 and 5 are all significant. This clearly looks like the data are consistent with a first order moving average process since all but the first acfs are not significant (the significant lag 4 acf is a typical wrinkle that one might expect with real data and should probably be ignored), and the pacf has a slowly declining structure.

(b)

$$Q^* = T(T+2)\sum_{k=1}^{3}\frac{\hat{\tau}_k^2}{T-k} = 100 * 102 * \left(\frac{(0.42)^2}{99} + \frac{(0.104)^2}{98} + \frac{(0.032)^2}{97}\right) = 19.408$$

Under H_0 (no autocorrelation) $Q^* \sim \chi_3^2$. the 5% critical value is 7.815. The observed value exceeds the critical value, so we reject the null hypothesis of no serial correlation.

6.10 The forecasts are the conditional expectations given the information available at time $t-1$, $\Omega_{t-1} = (y_{t-1}, y_{t-2}, \ldots; u_{t-1}, u_{t-2}, \ldots)$.

(a)

$$\begin{aligned} f_{t-1,1} = E(y_t|\Omega_{t-1}) &= 0.036 + 0.69E(y_{t-1}|\Omega_{t-1}) + 0.42 \times E(u_{t-1}|\Omega_{t-1}) + E(u_t|\Omega_{t-1}) \\ &= 0.036 + 0.69 \times y_{t-1} + 0.42 \times u_{t-1} + 0 \\ &= 0.036 + 0.69 \times 3.4 + 0.42 \times -1.3 \\ &= 1.836 \end{aligned}$$

$$\begin{aligned} f_{t-1,2} = E(y_{t+1}|\Omega_{t-1}) &= 0.036 + 0.69E(y_t|\Omega_{t-1}) + 0.42 \times E(u_t|\Omega_{t-1}) + E(u_{t+1}|\Omega_{t-1}) \\ &= 0.036 + 0.69 \times E(y_t|\Omega_{t-1}) + 0 \\ &= 0.036 + 0.69 \times 1.836 \\ &\approx 1.3028 \end{aligned}$$

$$\begin{aligned} f_{t-1,3} = E(y_{t+2}|\Omega_{t-1}) &= 0.036 + 0.69E(y_{t+1}|\Omega_{t-1}) \\ &\quad + 0.42 \times E(u_{t+1}|\Omega_{t-1}) + E(u_{t+2}|\Omega_{t-1}) \\ &= 0.036 + 0.69 \times E(y_{t+1}|\Omega_{t-1}) + 0 \\ &\approx 0.036 + 0.69 \times 1.3028 \\ &\approx 0.935. \end{aligned}$$

(b) Given the forecasts and the actual value, it is very easy to calculate the MSE by plugging the numbers in to the relevant formula, which in this case is

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(x_{t-1+n} - f_{t-1,n})^2$$

if we are making N forecasts. Then the MSE is given by

$$\begin{aligned} MSE &= (1.836 + 0.032)^2 + (1.302 - 0.961)^2 + (0.935 - 0.203)^2 \\ &= 3.489 + 0.116 + 0.536 = 4.141. \end{aligned}$$

Notice also that 84% of the total sum of squared errors is coming from the error in the first forecast. Thus error measures can be driven by one or two times when the model fits very badly. For example, if the forecast period includes a stock market crash, this can lead the mean squared error to be much bigger than it would have been if the crash observations were not included. This point needs to be considered whenever forecasting models are evaluated. An idea of whether this is a problem in a given situation can be gained by plotting the forecast errors over time.

6.11 (a) The shapes of the acf and pacf are perhaps best summarised in a table:

Process	acf	pacf
White noise	No significant coefficients	No significant coefficients
AR(2)	Geometrically declining	First 2 pacf coefficients significant, others insignificant
MA(1)	First acf coefficient significant, all others insignificant	Geometrically declining
ARMA(2,1)	Geometrically declining	Geometrically declining

(c) Since no values for the series y or the lagged residuals are given, the answers should be stated in terms of y and of u. Assuming that information is available up to and including time t, the 1-step ahead forecast would be for time $t+1$, the 2-step ahead for time $t+2$ and so on. A useful first step would be to write the model out for y at times $t+1$, $t+2$, $t+3$, $t+4$:

$$y_{t+1} = 0.21 + 1.32y_t + 0.58u_t + u_{t+1}$$

$$y_{t+2} = 0.21 + 1.32y_{t+1} + 0.58u_{t+1} + u_{t+2}$$

$$y_{t+3} = 0.21 + 1.32y_{t+2} + 0.58u_{t+2} + u_{t+3}$$

$$y_{t+4} = 0.21 + 1.32y_{t+3} + 0.58u_{t+3} + u_{t+4}$$

The 1-step ahead forecast would simply be the conditional expectation of y for time $t+1$ made at time t. Denoting the 1-step ahead forecast made at time t as $f_{t,1}$, the 2-step ahead

forecast made at time t as $f_{t,2}$ and so on. Denoting conditional expectation based on the information, Ω_t, available at time t by E_t, we have

$$\begin{aligned} E(y_{t+1}|y_t, y_{t-1}, \ldots) = f_{t,1} = E_t[y_{t+1}] &= E_t[0.021 + 1.32y_t + 0.58u_t + u_{t+1}] \\ &= = 0.21 + 1.32y_t + 0.58u_t \end{aligned}$$

since $E_t[u_{t+1}] = 0$. The 2-step ahead forecast would be given by

$$\begin{aligned} E(y_{t+2}|y_t, y_{t-1}, \ldots) = f_{t,2} &= E_t[y_{t+2}] = E_t[0.21 + 1.32y_{t+1} + 0.58u_{t+1} + u_{t+2}] \\ &= 0.21 + 1.32f_{t,1} \end{aligned}$$

since $E_t[u_{t+1}] = 0$ and $E_t[u_{t+2}] = 0$. Thus, beyond 1-step ahead, the MA(1) part of the model disappears from the forecast and only the autoregressive part remains. Although we do not know y_{t+1}, its expected value is the 1-step ahead forecast that was made at the first stage, $f_{t,1}$. The 3-step ahead forecast would be given by

$$\begin{aligned} E(y_{t+3}|y_t, y_{t-1}, \ldots) = f_{t,3} = E_t[y_{t+3}] &= E_t[0.21 + 1.32y_{t+2} + 0.42u_{t+2} + u_{t+3}] \\ &= 0.21 + 1.32f_{t,2} \end{aligned}$$

and the 4-step ahead by

$$\begin{aligned} E(y_{t+4}|y_t, y_{t-1}, \ldots) = f_{t,4} = E_t[y_{t+4}] &= E_t[0.21 + 1.32y_{t+3} + 0.42u_{t+3} + u_{t+4}] \\ &= 0.21 + 1.32f_{t,3}. \end{aligned}$$

(e) Moving average and ARMA models cannot be estimated using OLS – they are usually estimated by maximum likelihood. Autoregressive models can be estimated using OLS or maximum likelihood. Pure autoregressive models contain only lagged values of observed quantities on the RHS, and therefore, the lags of the dependent variable can be used just like any other regressors. However, in the context of MA and mixed models, the lagged values of the error term that occur on the RHS are not known a priori. Hence, these quantities are replaced by the residuals, which are not available until after the model has been estimated. But equally, these residuals are required in order to be able to estimate the model parameters. Maximum likelihood essentially works around this by calculating the values of the coefficients and the residuals at the same time. Maximum likelihood involves selecting the most likely values of the parameters given the actual data sample, and given an assumed statistical distribution for the errors.

6.12 (a) Some of the stylised differences between the typical characteristics of macroeconomic and financial data are as follows. One important difference is the frequency with which financial asset return time series and other quantities in finance can be recorded. This is of particular relevance for regressions with time series, since it is usually a requirement that all of the time-series data series used in estimating a given model must be of the same frequency. Thus, if, for example, we wanted to build a model for forecasting hourly changes in exchange rates, it would be difficult to set up a structural model containing macroeconomic explanatory variables since the macroeconomic variables are likely to be measured on a quarterly or at best monthly basis. This gives a motivation for using pure time-series approaches (e.g. ARMA models), rather than structural formulations with separate explanatory variables. It is also often of particular interest to produce forecasts of financial variables in real time. Producing forecasts from pure time-series models is usually simply an exercise in iterating with conditional expectations. But producing forecasts from structural models is considerably more difficult, and would usually require the production of forecasts for the structural variables as well.

(b) A simple"rule of thumb" for determining whether autocorrelation coefficients and partial autocorrelation coefficients are statistically significant at the the 5% level is to compare their absolute value to $1.96/\sqrt{T}$ In this case, T = 500, so a particular coefficient would be deemed significant if it is larger than 0.088 or smaller than –0.088. On this basis, the autocorrelation coefficients at lags 1 and 5 and the partial autocorrelation coefficients at lags 1, 2, and 3 would be classed as significant. The formulae for the Box-Pierce and the Ljung-Box test statistics are respectively

$$Q = T\sum_{k=1}^{m}\hat{\tau}_k^2$$

and

$$Q^* = T(T+2)\sum_{k=1}^{m}\frac{\hat{\tau}_k^2}{T-k}.$$

Note that we have only discussed the latter in class, and called it Q.

In this instance, the statistics would be calculated respectively as

$$Q = 500\left[0.307^2 + 0.013^2 + 0.086^2 + 0.0312^2 + 0.197^2\right] = 70.79$$

and

$$Q^* = 500 \times 502 \times \left[\frac{0.307^2}{500-1} + \frac{0.013^2}{500-2} + \frac{0.086^2}{500-3} + \frac{0.0312^2}{500-4} + \frac{0.197^2}{500-5}\right] = 71.39$$

The test statistics will both follow a χ^2 distribution with 5 degrees of freedom (the number of autocorrelation coefficients being used in the test). The critical values are 11.07 and 15.09 at 5% and 1% respectively. Clearly, the null hypothesis that the first 5 autocorrelation coefficients are jointly zero is resoundingly rejected

(c) Bartlett significance bands: $\pm 1.96/\sqrt{500} = \pm 0.088$. Setting aside the lag 5 autocorrelation coefficient, the pattern in the table is for the autocorrelation coefficient to only be significant at lag 1 and then to fall rapidly to values close to zero, while the partial autocorrelation coefficients appear to fall much more slowly as the lag length increases. These characteristics would lead us to think that an appropriate model for this series is an MA(1). Of course, the autocorrelation coefficient at lag 5 is an anomaly that does not fit in with the pattern of the rest of the coefficients. But such a result would be typical of a real data series (as opposed to a simulated data series that would have a much cleaner structure). This serves to illustrate that when econometrics is used for the analysis of real data, the data generating process was almost certainly not any of the models in the ARMA family. So all we are trying to do is to find a model that best describes the features of the data to hand. As one econometrician put it, all models are wrong, but some are useful!

(d) Strategy: replace future unknown values of u_t with 0, and future values of y_t with earlier forecasts. Hence, for Model A:

$$\begin{aligned}
\hat{x}_{z+1} &= 0.38 + 0.1u_z = 0.38 + 0.1 \cdot (-0.02) = 0.378 \\
\hat{x}_{z+2} &= 0.38 + 0.1\hat{u}_{z+1} = 0.38 + 0.1 \cdot 0 = 0.38 \\
\hat{x}_{z+3} &= 0.38 + 0.1\hat{u}_{z+2} = 0.38 + 0.1 \cdot 0 = 0.38 \\
\hat{x}_{z+4} &= 0.38 + 0.1\hat{u}_{z+3} = 0.38 + 0.1 \cdot 0 = 0.38
\end{aligned}$$

For Model B, we find

$$\hat{x}_{z+1} = 0.63 + 0.17x_z - 0.09x_{z-1} = 0.63 + 0.17 \cdot 0.31 - 0.09 \cdot 0.02 = 0.6809$$
$$\hat{x}_{z+2} = 0.63 + 0.17\hat{x}_{z+1} - 0.09x_z = 0.63 + 0.17 \cdot 0.6809 - 0.09 \cdot 0.31 = 0.7179$$
$$\hat{x}_{z+3} = 0.63 + 0.17\hat{x}_{z+2} - 0.09\hat{x}_{z+1} = 0.63 + 0.17 \cdot 0.7179 - 0.09 \cdot 0.6809 = 0.6908$$
$$\hat{x}_{z+4} = 0.63 + 0.17\hat{x}_{z+3} - 0.09\hat{x}_{z+2} = 0.63 + 0.17 \cdot 0.6908 - 0.09 \cdot 0.7179 = 0.6828.$$

(e) We have only discussed one method: residual diagnostics. This would involve examining the acf and pacf of the residuals from the estimated model. If the residuals showed any "action", that is, if any of the acf or pacf coefficients showed statistical significance, this would suggest that the original model was inadequate. It is worth noting that this model evaluation procedure would only indicate a model that was too small. If the model were too large, i.e. it had superfluous terms, these procedures would deem the model adequate. This could be solved by testing the model parameters for significance, and removing insignificant ones, provided the residual correlogram still looks OK after doing that. In principle, this can lead to so-called subset ARMA models, where some individual lag orders are removed. This is not supported in `statsmodels`, however.

(f) We compute the MSE for both models. For Model A, we find

$$\begin{aligned} MSE &= \frac{1}{4}\sum_{i=1}^{4}(x_{z+i} - \hat{x}_{z+i})^2 \\ &= \frac{1}{4}\Big[(0.62 - 0.378)^2 + (0.19 - 0.38)^2 \\ &\qquad + (-0.32 - 0.38)^2 + (0.72 - 0.38)^2\Big] = 0.175. \end{aligned}$$

For Model B,

$$\begin{aligned} MSE &= \frac{1}{4}\sum_{i=1}^{4}(x_{z+i} - \hat{x}_{z+i})^2 \\ &= \frac{1}{4}\Big[(0.62 - 0.6809)^2 + (0.19 - 0.7179)^2 \\ &\qquad + (-0.32 - 0.6908)^2 + (0.72 - 0.6828)^2\Big] = 0.326. \end{aligned}$$

So Model A performs better with respect to MSE. Note that for this contrived example, the forecasts differ quite a lot. In practice, if you have two competing models that both pass the residual diagnostics, they will tend to produce similar forecasts.

Chapter 8 (p. 492–496): (Co-)Integration

8.2 (a) The null hypothesis is H_0: $\Psi = 0$ while the alternative hypothesis is H_a: $\Psi < 0$.

(b) The null hypothesis should be rejected if the ratio $\frac{\widehat{\Psi}}{SE(\widehat{\Psi})} \leq \tau_{\mu,0.05} = -2.86$. This is the critical value when there is a drift μ in the regression, but no deterministic trend of the form λt. (Note that Brooks has a typo in the critical value.) The observed value is $\frac{\widehat{\Psi}}{SE(\widehat{\Psi})} = -0.06$, which is too large to reject the null hypothesis.

(c) There is insufficient evidence for rejecting the null, suggesting that the data might be $I(1)$. To verify this, one could difference the data and apply the DF test once more. If the original series is indeed $I(1)$, this will most likely lead to a rejection of the null hypothesis after differencing.

(d) The usual critical values do not apply since, due to the non-stationarity of y_t, the ratio $\frac{\widehat{\Psi}}{SE(\widehat{\Psi})}$ does not follow a standard t-distribution under the null hypothesis.

8.3 (a) The DF test statistic in this case is $\frac{\widehat{\Psi}}{SE(\widehat{\Psi})} = \frac{-0.52}{0.16} = -3.25$ which is smaller than the critical value $\tau_{\mu,0.05} = -2.86$, hence there is evidence that the data are actually generated by a stationary process.

(b) One might include a deterministic trend and apply the DF test once more, since a mis-specification of this type (omitting it while it should be there) can lead to over-rejection of the test (rejecting at a rate larger than the nominal rejrection rate ($\alpha = 0.05$ in this case) under the null hypothesis).

(c) A solution to serial dependence in the residuals is to perform an augmented Dickey Fuller test (including lagged values Δy_{t-k} in the regression, $k = 1, \ldots, p$). The order p can be selected automatically by minimizing some information criterion.

Chapter 9 (p. 569–572): GARCH Models

9.1 (a) Some stylized features of returns data are: i) little to no autocorrelation, ii) heavy tails, iii) volatility clustering, iv) leverage effects. Linear (ARIMA) models can only be used to model the autocorrelation (and the heavy tails, if the error distribution is non-normal).

(b) A GARCH(1,1) model (with normally distributed errors) can be used to model the volatility clustering. It also offers a partial explanation for the heavy tails, because it can be shown that volatility clustering leads to excess kurtosis in the unconditional distribution of returns.

(c) It can be shown that a GARCH(1,1) model corresponds to an ARCH(∞) model, with a specific structure for the ARCH coefficients. It allows the researcher to model an autocorrelation structure in the squared returns for which a high-order ARCH model might be required, while requiring the estimation of only 3 parameters. Despite its parsimony, the GARCH(1,1) structure often proves sufficient to model the volatility clustering.

(d) The simple GARCH(1,1) model cannot address the leverage effects found in financial returns. Extensions that allow for this include the GJR and EGARCH specifications. Furthermore, the GARCH model with normally distributed errors cannot account for all of the heavy-tailedness of the data; hence a non-normal error distribution may be warranted, such as the Student's t or GED.

(e) Daily returns typically have a mean close to zero, so $\mu \approx 0$. The estimates for the GARCH equation are often close to the stationarity boarder, e.g., $\alpha_1 = 0.05, \beta = 0.94$. The unconditional volatility is given by $\sigma^2 = \alpha_0/(1-\alpha_1-\beta)$. We expect an annualized volatility of perhaps 30%, corresponding to a daily volatility of about 1.875% $(30\%/\sqrt{256})$. Hence we expect that $\alpha_0 \approx 0.000003515625$, so that $\sigma = \sqrt{0.000003515625/(1-0.99)} = 0.01875$. If the returns are expressed in percent (i.e., they have been multiplied by 100), then α_0 will be 10.000 times as large.

(f) (has not been covered in class)

(g) First, observe that $\hat{\alpha}_1 + \hat{\beta} = 1.1062 > 1$, so this model is non-stationary. In view of the next question, this appears to be a typo, so let's assume that $\hat{\beta} = 0.811$, leading to a stationary model. Then

$$\begin{aligned} \text{var}_T(y_{T+1}) &= \sigma^2_{T+1} = \alpha_0 + \alpha_1 y_T^2 + \beta\sigma_T^2 \\ \text{var}_T(y_{T+k}) &= \sigma^2 + (\alpha_1+\beta)^{k-1}(\sigma^2_{T+1} - \sigma^2), \quad k \geq 2 \end{aligned} \tag{1}$$

where we have put T in place of t. Plugging in the estimates σ^2_{T+1}, $\hat{\alpha}_1$, $\hat{\beta}$, and $\hat{\sigma} = \sqrt{\hat{\alpha}_0/(1-\hat{\alpha}_1-\hat{\beta})}$, this can be used to forecast the conditional variance (and hence the volatility) k days ahead.

(h) The forecast equation in the previous question is valid only if the model is stationary. Here, we have $\hat{\alpha}_1 + \hat{\beta} = 1.1051 > 1$, so this model is nonstationary. A similar derivation shows that

$$\text{var}_T(y_{T+k}) = \alpha_0 \sum_{j=1}^{k-1} (\alpha_1+\beta)^{j-1} + (\alpha_1+\beta)^{k-1}\sigma^2_{T+1}, \quad k \geq 2.$$

If $(\alpha_1+\beta)$, the volatility forecast will explode exponentially.

9.3 (a) The conditional variance $\sigma^2_{t+k} \equiv \text{var}_t(y_{t+k}) \equiv \text{var}(y_{t+k}|\Omega_t)$ measures the variability of future returns, *conditional* on the information available at time t. The unconditional variance $\sigma^2 \equiv \mathbb{E}[\sigma^2_t]$, on the other hand, does not 'use' the information whether period t is in a low- or high volatility regime. It measures the *average* variance. A 1-day forecast relies strictly on the conditional variance. If the model is stationary, then according to (1), the multi period forecasts converge to the unconditional variance.

(b) The regression in question is a regression of y_t on just a constant. If heteroskedasticity is present but ignored, then μ will be estimated consistently, but its standard error will be off. Hence one should use heteroskedasticity consistent ('White') standard errors. Newey-West standard errors can also be used; they are consistent even if there is autocorrelation in addition to the heteroskedasticity. Hence their name, 'HAC'.

(c) Historical volatility is simple to compute, but reacts slowly to shocks and displays a 'ghosting' feature. GARCH models are harder to estimate, but extremely successful in forecasting conditional volatility. EWMA is the method used by RiskMetrics. It is a special case of the GARCH model; all parameters are fixed, so no estimation is necessary. The model is thus very simple, but less flexible than a GARCH model. In particular, it is only suitable for daily data. Finally, implied volatility has the advantage that it is a forward-looking measure. However, the empirical evidence on the accuracy of implied versus statistical forecasting models is mixed, and some research suggests that implied volatility systematically over-estimates the true volatility of the underlying asset returns. This may arise from the use of an incorrect option pricing formula to obtain the implied volatility. As a result, other features of the data (such as excess kurtosis) are proxied for by the (implied) volatility.

Chapter 11 (p. 654–656): Panel data

10.1 (a) A pooled regression refers to just stacking the time series for all individuals on top of one another and treating the data like one big sample, ignoring the panel character. The main advantage of the fixed effects estimator compared to a pooled regression is that the individual fixed effects allow us to account for the individual heterogeneity. As a consequence of this, the estimator is immune to omitted variable bias, as long as the omitted variables are time invariant, because their effect will be removed by the within transformation. Examples of such time invariant variables are company size, industry, CEO ability, etc. Note that the last one of these is unobservable, so we will never be able to obtain data on it; hence we cannot include it in a regression, and thus the fixed effects estimator is the only way to get rid of the omitted variable bias associated with it. A similar argument holds for the time effects; their inclusion rids us of all the effects of omitted variables that are constant across individuals, but vary over time (such as recessions).

(c) A balanced panel has the same number of observations for each individual, so that the total number of observations is NT. A typical example of an unbalanced panel is stock market data, because some companies may get delisted.

10.2 (a) The LSDV estimator can also be obtained by subtracting, from each outcome y_i of individual i, the quantity $\bar{y}_{i\cdot} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, i.e., the mean over all observations for that individual. The same is done with each x_{it}. This yields

$$\ddot{y}_{it} = y_{it} - \bar{y}_{i\cdot}, \quad \ddot{x}_{j,it} = x_{j,it} - \bar{x}_{j,i\cdot},$$

This is called the *within transformation*. The transformed model can be estimated by OLS:

$$\ddot{y}_{it} = \sum_{j=1}^{l} \ddot{x}_{j,it}\beta_j + \ddot{\nu}_{it}.$$

It can be shown that this is equivalent to including a dummy for each individual. The benefit is that it removes the hassle of having to estimate an extra N parameters. For constructing tests and confidence intervals, we must only take care to use the proper degrees of freedom (because we implicitly still estimate the fixed effects); software implementations usually do that automatically. If the models includes time fixed effects as well, then the within transformation becomes

$$\ddot{y}_{it} = y_{it} - \bar{y}_{i\cdot} - \bar{y}_{\cdot t}, \quad \ddot{x}_{j,it} = x_{j,it} - \bar{x}_{j,i\cdot} - \bar{x}_{j,\cdot t},$$

and we can again apply OLS to the transformed variables.

(b) The RE estimator assumes that the individual heterogeneity is random, and INDEPENDENT of the regressors. The latter is unlikely to be true.

(c) If the individual effects are indeed independent of the regressors, then RE is better (consistent and efficient), FE is only consistent. If independence does not hold, then FE is consistent and efficient, but RE is neither. So FE is the safer choice. We can choose between them by means of a Hausman-Wu test.