Forecasting Time Series Data
Notes from Rebecca Sela
Stern Business School
Spring, 2004

1 Time Series Concepts and Challenges

The linear regression model (and most other models) assume that observations are independent and identically distributed. This may not be true in time series, because current values often depend on what happened in previous periods. This may cause series to seem to have trends ("autocorrelation masquerading as trend") and will make standard errors wrong.

A time series is a stochastic process (a sequence of random variables). This allows us to define the following values:

- Mean: $E(X_t) = \mu_{X_t}$
- Variance: $Var(X_t) = E((X_t - \mu_{X_t})^2)$
- Standard Deviation: $StdDev(X_t) = \sqrt{Var(X_t)}$
- Covariance: $Cov(X_t, Y_s) = E((X_t - \mu_{X_t})(Y_s - \mu_{Y_s}))$
- Autocovariance: The covariance between X_t and X_{t-k}
- Correlation: $Corr(X_t, Y_s) = \frac{Cov(X_t, Y_s)}{\sqrt{Var(X_t)Var(Y_s)}}$, and $-1 \leq Corr(X_t, Y_s) \leq 1$
- Autocorrelation: The correlation between X_t and X_{t-k}. We call this number ρ_k.

If the correlation between two variables is zero, then the value of one has no effect on a linear prediction of the other. Autocorrelation means that past values of X_t can be used to predict future values. Even if the autocorrelation is zero, they may be related non-linearly, and thus are not necessarily independent.

Definition 1 *A time series is weak (covariance) stationary if the mean and variance are constant over time, and the autocovariance, $Cov(X_t, X_{t-k})$, depends only on the lag, k. A time series is wideset stationary if the distribution of $(X_t, X_{t-k_1}, X_{t-k_2}, ..., X_{t-k_n})$ does not depend on anything except $(k_1, ..., k_n)$.*

Definition 2 *A time series, ε_t, is white noise if $E(\varepsilon_t) = 0$ for all t and $Corr(\varepsilon_t, \varepsilon_s) = 0$ for all $t \neq s$. (It need not be normally distributed.) These are sometimes called the shocks or innovations in a time series, and they might not be observed directly.*

1.1 The basics of forecasting

Let h be the lead time, which is how far ahead the forecast is for. Let n be the current period, so that all values from period 0 to n are observed. That is, we know $X_0, ..., X_n$; this is called the information set, Ψ_n. A trace forecast is a forecast for leads $h = 1, ...H$. A point forecast is a single number for each X_{n+h}. An interval forecast provides a likely range for each X_{n+h}. Event forecasting estimates the probability of certain future events. A conditional forecasting approach takes the past as fixed and considers the future as random. Using the known information from the past helps to predict the future better.

Let $f_{n,h}$ be the forecast of X_{n+h} based on the information set, Ψ_n. Then $e_{n,h} = X_{n+h} - f_{n,h}$ is the forecast error. (This is a random variable.)

A cost function gives the cost of a wrong forecast, which may vary depending on how wrong the forecast is or whether it is above or below the realization. The cost function may affect the choice of intervals and forecasts. The simplest (and usually implicit) cost function is the squared error cost function, where the cost is the square of the distance between the forecast and the realization. This means that we want to minimize $E(e_{n,h}^2)$ with $E(e_{n,h}) = 0$.

1.2 Transformations of Time Series

Let X_t be a time series. Then, $Y_t = \Delta X_t = X_t - X_{t-1}$ is the first difference of the time series. This removes linear trends and, sometimes, non-stationarity. We recover a time series from its difference by:

$$\begin{aligned} X_t &= Z_t - Z_{t-1} \\ Z_t &= X_t + Z_{t-1} \\ &= X_t + X_{t-1} + ...X_{t-k-1} + Z_{t-k} \\ &= \sum_{k=0}^{\infty} X_{t-k} \end{aligned}$$

This is integration. Note that two time series that differ by a constant can have the same first difference.

In general, the d^{th} difference is of the form $X_t = Z_t - dZ_{t-1} + ... + (-1)^k \binom{d}{k} Z_{t-k} + ... + (-1)^d Z_{t-d} = \sum_{k=0}^{d} (-1)^k \binom{d}{k} Z_{t-k}$.

If the volatility of X_t seems to depend on the level of X_t, then it can be helpful to take the natural log of X_t before further analysis. This is particularly necessary for price series. The first difference of the natural log of a time series is approximately the return: $\frac{X_t - X_{t-1}}{X_{t-1}} \approx \ln(X_t) - \ln(X_{t-1})$.

2 ARIMA

2.1 Moving Average (MA)

2.1.1 The $MA(1)$ Model

This model is given by

$$X_t = \varepsilon_t + \beta\varepsilon_{t-1}$$

That is, the observed value is the weighted average of the last two shocks. With this model, we find:

$$\begin{aligned}
E(X_t) &= E(\varepsilon_t + \beta\varepsilon_{t-1}) = 0 + \beta * 0 = 0 \\
Var(X_t) &= Var(\varepsilon_t + \beta\varepsilon_{t-1}) = Var(\varepsilon_t) + \beta^2 Var(\varepsilon_{t-1}) + Cov(\varepsilon_t, \varepsilon_{t-1}) \\
&= (1+\beta^2)Var(\varepsilon_t) \\
Cov(X_t, X_{t-1}) &= Cov(\varepsilon_t + \beta\varepsilon_{t-1}, \varepsilon_{t-1} + \beta\varepsilon_{t-2}) \\
&= Cov(\varepsilon_t, \varepsilon_{t-1}) + \beta Cov(\varepsilon_{t-1}, \varepsilon_{t-1}) + \beta Cov(\varepsilon_t, \varepsilon_{t-2}) + \beta^2 Cov(\varepsilon_{t-1}, \varepsilon_{t-2}) \\
&= \beta Cov(\varepsilon_{t-1}, \varepsilon_{t-1}) = \beta Var(\varepsilon_t) \\
Corr(X_t, X_{t-1}) &= \frac{Cov(X_t, X_{t-1})}{\sqrt{Var(X_t)Var(X_{t-1})}} = \frac{\beta Var(\varepsilon_t)}{(1+\beta^2)Var(\varepsilon_t)} = \frac{\beta}{1+\beta^2} \\
Cov(X_t, X_{t-k}) &= Corr(X_t, X_{t-k}) = 0 \text{ if } k > 1
\end{aligned}$$

(This last result occurs because they have no shocks in common.) Thus the process has a very short memory.

The series is positively correlated if $Corr(X_t, X_{t-1}) > 0$. Negatively correlated series are rare. If your series is negatively correlated, you may have overdifferenced. For example, if X_t is white noise, then $\Delta X_t = \varepsilon_t - \varepsilon_{t-1}$ is $MA(1)$ with $\beta = -1$.

Note that the theoretical forecastability, $R^2 = \frac{\beta^2}{(1+\beta^2)^2}$, is at most 0.5. Thus, $MA(1)$ series are not very forecastable.

To forecast in $MA(1)$ series, one step ahead, if β is known:

1. Note that $x_{n+1} = \varepsilon_{n+1} + \beta\varepsilon_n$.
2. The best forecast of ε_{n+1} is its expected value, 0.
3. Thus,we forecast: $f_{n,1} = \beta\varepsilon_n$.

However, we generally do not observe ε_n directly. Instead, we may estimate it from previous values:

1. Forecast $f_{0,1} = 0$.
2. Set $\widehat{\varepsilon_0} = x_1 - f_{0,1} = x_1$.
3. In general, set $\widehat{\varepsilon_k} = x_k - f_{k-1,1}$, and then use this to forecast $f_{k,1} = \beta\widehat{\varepsilon_k}$.
4. Continue until you have an estimate, $\widehat{\varepsilon_n}$.

2.1.2 $MA(q)$ Models

The $MA(q)$ model is given by $X_t = \varepsilon_t + \beta_1\varepsilon_{t-1} + \beta_2\varepsilon_{t-2} + ... + \beta_q\varepsilon_{t-q}$. This model has autocorrelation from lags 1 to q, and no autocorrelation after that. This model will be stationary for any set of β_i's.

The forecast for this model is: $f_{n,k} = \beta_k\varepsilon_n + ... + \beta_q\varepsilon_{n+k-q}$, since we include all the known shocks in the forecast and set all future shocks to 0. If the lead time, k, is greater than q, then the forecast is 0. The forecast error is $e_{n,k} = X_{n+k} - f_{n,k} = \varepsilon_{n+k} + \beta_1\varepsilon_{n+k-1} + ... + \beta_{k-1}\varepsilon_{n+1}$. The variance of this forecast error is $Var(e_{n,k}) = Var(\varepsilon_{n+k} + \beta_1\varepsilon_{n+k-1} + ... + \beta_{k-1}\varepsilon_{n+1}) = Var(\varepsilon_t)(1 + \beta_1^2 + ... + \beta_{k-1}^2)$. Note that the forecast error variance increases with the lead time. After q periods, the forecast error is simply the variance of X_t.

An $MA(q)$ model is invertible if we can write ε_t in terms of $X_t, X_{t-1}, ...$ In the case of the $MA(1)$, this is simply:

$$\begin{aligned} \varepsilon_t &= X_t - \beta\varepsilon_{t-1} \\ &= X_t - \beta(X_{t-1} - \beta\varepsilon_{t-1}) \\ &= ... \\ &= \sum_{k=0}^{\infty}(-\beta)^k X_{t-k} \end{aligned}$$

This will converge only if $|\beta| < 1$. More generally, an $MA(q)$ model, $X_t = \varepsilon_t + \beta_1\varepsilon_{t-1} + ... + \beta_q\varepsilon_{t-q}$ is invertible if the polynomial $z^q + \beta_1 z^{q-1} + ... + \beta_{q-1}z + \beta_q = 0$ has all of its roots inside the unit circle. (Generally, a model is not invertible if it has been over-differenced.)

2.2 Autoregressive Models

2.2.1 $AR(1)$ Model

This model is given by

$$X_t = \alpha X_{t-1} + \varepsilon_t$$

Note that, if the white noise is independent (not just uncorrelated), then this is a Markov process, and $E(X_t|X_{t-1}, ..., X_{t-k}) = E(X_t|X_{t-1})$.

We may write X_t in terms of past shocks, as an $MA(\infty)$ model, and use this

to compute the variance:

$$\begin{aligned} X_t &= \varepsilon_t + \alpha(\varepsilon_{t-1} + X_{t-2}) \\ &= ... \\ &= \sum_{j=0}^{N-1} \alpha^j \varepsilon_{t-j} + \alpha^N X_{t-N} \\ &= \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j} \\ Var(X_t) &= Var(\sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}) \\ &= \sum_{j=0}^{\infty} \alpha^{2j} Var(\varepsilon_t) \\ &= \frac{Var(\varepsilon_t)}{1-\alpha^{2j}} \end{aligned}$$

If $|\alpha| = 1$, then this model is a random walk. Its variance is infinite. If $|\alpha| > 1$, then the model is explosive. In most cases that we consider, $|\alpha| < 1$, and the process is zero-mean-reverting. That is, $E(X_{n+1}|X_n) = \alpha X_n$, which is closer to 0. The closer $|\alpha|$ is to 0, the faster the mean reversion happens. If $|\alpha| < 1$, then the autocorrelation is $Corr(X_t, X_{t-k}) = \alpha^k$. Thus, the current observation provides some information about every future value.

To forecast an $AR(1)$ model, we set all future errors to 0, and use the equations to find:

$$\begin{aligned} f_{n,1} &= \alpha X_n + E(\varepsilon_{n+1}) = \alpha X_n \\ f_{n,2} &= \alpha E(X_{n+1}) + E(\varepsilon_{n+2}) = \alpha^2 X_n \\ f_{n,h} &= \alpha^h X_n \end{aligned}$$

As $h \to 0$, $f_{n,h} \to 0 = E(X_{n+h})$.

2.2.2 $AR(p)$ Model

The $AR(p)$ model is given by

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \varepsilon_t$$

The autocorrelations are more complex, but decrease exponentially fast to 0 if the process is stationary.

An $AR(p)$ model is stationary if and only if the largest (in modulus) root, θ, of $z^p = \alpha_1 z^{p-1} + ... + \alpha_{p-1} z + \alpha_p$, has modulus less than one. (That is, all roots have modulus less than one – all roots lie within the unit circle.) If any root lies outside the unit circle, then the process is explosive. If at least one root is on the unit circle and all other roots are inside, then the process has a unit

root; that is, the differences (first, second, or more, depending on the number of unit roots) are stationary but the process itself is not mean-reverting. In the $AR(1)$ case, this simply tests whether $|\alpha|$ is greater than, less than, or equal to one.

2.3 The Box-Jenkins $ARMA(p, q)$ model

The $ARMA(p, q)$ model is given by:

$$X_t = \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + ... + \beta_q \varepsilon_{t-q}$$

Note that an $AR(p)$ model is an $ARMA(p, 0)$ model and an $MA(q)$ model is an $ARMA(0, q)$ model. Note that some non-trivial-looking ARMA models are really white noise: For example, $X_t = -0.5X_{t-1} + \varepsilon_t - 0.5\varepsilon_{t-1}$ reduces to $X_t = \varepsilon_t$.

To forecast using these models:

1. Forecast all future errors as 0.
2. Estimate past ε_t by $\widehat{\varepsilon_t} = X_t - f_{t-1,1}$.
3. Plug these values into the equation along with the observed X_t's.

The sum of two time series that are ARMA's is also an ARMA. In particular, the sum of an $ARMA(p_1, q_1)$ and an $ARMA(p_2, q_2)$ is $ARMA(p_1+p_2, \max\{p_1+q_2, p_2+q_1\})$.

2.4 The full Box-Jenkins $ARIMA(p, d, q)$ model

An $ARIMA(p, d, q)$ model is a model where the d^{th} difference is a stationary, invertible $ARMA(p, q)$ model. In this case, we say that the time series is integrated of order d. An $ARMA(p, q)$ model is an $ARIMA(p, 0, q)$ model.

Some special ARIMA models include:

- White noise: $ARIMA(0, 0, 0)$
- Random walk: $ARIMA(0, 1, 0)$

Any ARIMA model can also include a constant term. If an $ARIMA(p, 0, q)$ model has a constant term, then the series no longer necessarily has a zero mean. The constant is not the new mean, however. For example, in an $AR(1)$ model with a constant:

$$\begin{aligned} E(X_t) &= \alpha E(X_{t-1}) + c = \alpha E(X_t) + c \\ E(X_t) &= \frac{c}{1-\alpha} \end{aligned}$$

If an $ARIMA(p, 1, q)$ model has a constant term, then it has a linear trend.

2.4.1 Integrated Moving Averages and Exponential Smoothing

The exponential smoothing (exponentially weighted moving average) model assumes that each period has a local mean, $\overline{X_{t-1}}$. The next observation follows:

$$X_t = \overline{X_{t-1}} + \varepsilon_t$$

Since ε_t is white noise, we forecast $f_{n,h} = \overline{X_n}$, for any lead h. The local mean is updated according to:

$$\begin{aligned} \overline{X_t} &= \alpha X_t + (1-\alpha)\overline{X_{t-1}} \\ &= ... \\ &= \sum_{k=0}^{\infty} \alpha(1-\alpha)^k X_{t-k} \end{aligned}$$

This shows that the local mean is a moving weighted average of previous observations, and that the weights decrease exponentially. This forecasting method is equivalent to the forecasting method for the $ARIMA(0,1,1)$, with

$$X_t - X_{t-1} = \varepsilon_t - (1-\alpha)\varepsilon_{t-1}$$

Since the difference must be invertible, $-1 \leq (1-\alpha) \leq 1$, and $0 \leq \alpha \leq 2$.

Double exponential smoothing and Winter's method are more complex forms of this, allowing for separate parameters for the trend, level, and seasonality.

2.5 Model Identification

Model identification consists of choosing p, d, and q. First, choose d to make the time series stationary. Then, look at autocorrelations, partial autocorrelations, and other tests to choose p and q. In general, the most parsimonious model that seems to fit the data is the best option; more parameters lead to overfitting and bad forecasts, as well as unstable parameter estimates and multicollinearity.

2.5.1 Autocorrelation functions and partial autocorrelation functions

Recall that the autocorrelation function (ACF) for a stationary process is given by $\rho_k = \frac{Cov(X_t, X_{t-k})}{Var(X_t)}$. The autocorrelation function, ρ_k, for an $MA(q)$ process is zero when $k > q$. The autocorrelation function for an $AR(p)$ process dies down exponentially fast, but is technically non-zero for all k. Thus, if a graph of the autocorrelation function cuts off after a certain lag, q, the model is $MA(q)$.

The partial autocorrelation at lag k describes the relationship between X_t and X_{t-k} once the effects of $X_{t-1}, ..., X_{t-k+1}$ are removed. (This is done by regressing X_t on $X_{t-1}, ..., X_{t-k+1}$and looking at the correlation between the residuals and X_{t-k}. Note that the first partial autocorrelation equals the first autocorrelation.) The partial autocorrelation function (PACF) for an $AR(p)$ model cuts off after lag p, but the PACF for an $MA(q)$ model dies down

exponentially. Thus, if a graph of the partial autocorrelation function cuts off after a certain lag, p, the model is $AR(p)$.

The ACF and PACF for an $ARMA(p,q)$ model (with $p > 0$ and $q > 0$) both die down exponentially. Thus, the ACF and PACF only identify pure AR and MA models.

Since we only know the realizations of the time series, we can only plot the sample ACF and PACF. Because these are only sample estimates, they will not be exactly zero, even after they theoretically should cut off. Instead, one should compare the sample ACF and PACF to the standard errors to determine if they are really zero. The simplest hypothesis test is whether the autocorrelation or partial autocorrelation is greater (in absolute value) than $\frac{2}{\sqrt{n}}$, where n is the sample size. A more correct way adjusts the standard errors for the k^{th} element to assume that the model is $MA(k-1)$ or $AR(k-1)$; this increases the standard error as k increases.

Note that both the ACF and the PACF may seem to cut off. That is an artifact of the data. Just because the ACF stops being significant at lag q and the PACF stops being significant at lag p does NOT mean that this is an $ARMA(p,q)$ series!

2.5.2 The (Corrected) Akaike Information Criterion

The Akaike Information Criterion (AIC) measures the "distance" of an estimated model from the "truth" (assuming that the truth is not in the set of possible models). This criterion decreases as the fit improves and increases with the number of parameters, which rewards parsimony. The standard version of the AIC is:

$$AIC(p,q) = -2n\log(\frac{SS}{n}) + 2(p+q)$$

where n is the number of observations (if the observations have been differenced, it is the sample size after differencing, that is, $n-d$) and SS is the residual sum of squares from the model. If a constant term is also being estimated, then $p+q$ should be replaced by $p+q+1$. However, the AIC may sometimes choose models with very large numbers of parameters. The corrected AIC (which fixes this) is:

$$AIC_c(p,q) = n\log(\frac{SS}{n}) + 2(p+q+1)\frac{n}{n-p-q-2}$$

Again, if a constant term is also being estimated, then $p+q$ should be replaced by $p+q+1$. In the case of a random walk, we many compute

$$SS = \sum(X_t - X_{t-1})^2$$

if the model has no constant, and

$$SS = \sum(X_t - X_{t-1} - \overline{(X_t - X_{t-1})})^2$$

Another test statistic (which assumes that the truth is one of the possible models) is the Bayesian Information Criterion (BIC):

$$BIC(p,q) = n\log(\frac{SS}{n}) + (p+q)\log n$$

Note that all of these test statistics assume that d is determined beforehand; they only help determine p and q.

2.6 Estimating Parameters

The most common and general approach to estimating parameters is optimization using a computer. Programs generally choose parameters to minimize squared forecast errors or use maximum likelihood estimates. Certain cases can be estimated by hand. (Since this is a different method, these estimates may not agree with the parameters found by optimization.)

2.6.1 Yule-Walker System of Equations (AR models)

In the $AR(1)$ model, the first autocorrelation is equal to the parameter, α. This is a simple estimate of the parameter. More generally, the Yule-Walker equations use the relationships among the autocorrelations to estimate the parameters:

$$\rho_k = \alpha_1\rho_{k-1} + \alpha_2\rho_{k-2} + ... + \alpha_p\rho_{k-p}$$

Substituting r_j for ρ_j, we find the Yule-Walker equations:

$$r_k = \widehat{\alpha_1}r_{k-1} + \widehat{\alpha_2}r_{k-2} + ... + \widehat{\alpha_p}r_{k-p}$$

and then solve for the $\widehat{\alpha_i}$.

2.6.2 $MA(1)$ Estimation

Recall that $\rho_1 = \frac{\beta}{1+\beta^2}$. We may use this fact to solve for β in terms of ρ_1. Solving and subsituting the sample autocorrelation r_1, we find the estimates:

$$\widehat{\beta} = \frac{1 \pm \sqrt{1-4r_1^2}}{2r_1}$$

In theory, since $|\rho_1| \leq 0.5$, this always has real solutions. However, sample autocorrelations might be greater than 0.5, which can be a problem. Assuming that the solutions are real, only one will lead to an invertible model; that is the solution with absolute value less than (or equal to) one.

2.7 Checking the model

Once the model has been estimated, there should be no structure in the residuals and the model should reflect known properties of the data.

2.7.1 Ljung-Box-Pierce Test

This test sums the squares of the autocorrelations, to test the null hypothesis that all of them are zero. This can be done over any number of autocorrelations; testing only the first few makes finding structure in those more likely, but may miss structure in residuals further out. If the Ljung-Box-Pierce test rejects the null hypothesis, then the model might require more lags, require that the data be used in logs, suggest seasonality, or reflect other factors.

2.7.2 Implied Forecasts

Plot both the forecasts and the forecast errors from the model, either at the end of the data or beginning in the middle of the series. If the forecast seems unreasonable or if the forecast intervals seem too wide or narrow, the model may need to be changed. In particular, the forecast intervals for any stationary time series will approach a constant, since the variance is constant into the future. If this seems unlikely, the model should not be stationary.

2.8 Box-Jenkins Model-Building

The following method is used to build an ARIMA model:

1. Transformation and pre-processing: Detrend, seasonally adjust, or take the logarithm of the series, if necessary. These parts of the series should be added back in after forecasting with ARMA to predict the level of the series better.
2. Identification: Choose p, d, and q.
3. Estimation: Estimate $\alpha_1, ..., \alpha_p, \beta_1, ..., \beta_q$.
4. Diagnostic checking: Ensure that the residuals are white noise. Make sure that the model is parsimonious. If this is not true, return to the identification step.

Note that ARMA models are only an approximation of the truth. The fact that they are not true implies an unmeasurable forecast error.

2.9 Prediction Intervals

To estimate the variance of forecasts, we use the $MA(\infty)$ representation of a time series:

$$X_t = \varepsilon_t + c_1\varepsilon_{t-1} + c_2\varepsilon_{t-2} + ...$$

This representation exists for any time series. However, the sequence $\{c_k\}$ might not converge to zero for non-stationary series. We may then write future values as:

$$\begin{aligned} X_{n+h} &= (\varepsilon_{n+h} + c_1\varepsilon_{n+h-1} + ... + c_{h-1}\varepsilon_{n+1}) + (c_h\varepsilon_n + c_{h+1}\varepsilon_{n-1} + ...) \\ &= FutureShocks + PastShocks \end{aligned}$$

Since the best forecast uses the actual values of all past and present shocks and sets all future shocks to zero (their expected value), we have:

$$\begin{aligned} f_{n,h} &= c_h\varepsilon_n + c_{h+1}\varepsilon_{n-1} + ... \\ e_{n,h} &= \varepsilon_{n+h} + c_1\varepsilon_{n+h-1} + ... + c_{h-1}\varepsilon_{n+1} \\ Var(e_{n,h}) &= (1 + c_1^2 + ... + c_{h-1}^2)Var(\varepsilon_t) \end{aligned}$$

If we assume that the shocks are independent and normally distributed, then our 95% prediction interval is:

$$f_{n,h} \pm 1.96\sqrt{Var(e_{n,h})}$$

This shows us that the width of the prediction interval depends on the volatility of the shocks, the lead time, the model and the parameters, and the error rate (95%) in this case. This is a prediction interval conditional on all of the shocks observed through this period.

All of this assumes that the model and its parameters are known, not estimated, and that errors are normally distributed. If any of these assumptions are not true, then the prediction interval may be incorrect (and should probably be wider).

3 Testing for Time Series Properties

3.1 Testing for Autocorrelation: Durbin-Watson

Recall that in any time series regression, autocorrelation in the errors will mean that the true standard errors are larger than the estimated standard errors. The Durbin-Watson test checks for this autocorrelation.

First, fit the model using ordinary least squares regression and get the residuals, $\{e_t\}$. To test for autocorrelation, we have the Durbin-Watson test statistic:

$$\begin{aligned} d &= \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \\ &= \frac{\sum_{t=2}^{n}(e_t^2 + e_{t-1}^2 - 2e_te_{t-1})}{\sum_{t=1}^{n} e_t^2} \\ &\approx 1 + 1 - 2\frac{\sum_{t=2}^{n} e_te_{t-1}}{\sum e_t^2} \\ &\approx 2 - 2r_1 = 2(1 - r_1) \end{aligned}$$

where r_1 is the sample first autocorrelation. Under the null hypothesis. $\rho_1 = 0$ and $d = 2$; if $\rho_1 > 0$ then $d < 2$. Informally, values of d much smaller than 2 (0 is a lower bound) indicate autocorrelation. However, cutoffs and p-values are hard to construct. In addition, this test will not catch if higher autocorrelations are non-zero if the first autocorrelation is 0. Finally, the test can be inconclusive, leaving one to have to decide whether to correct for autocorrelation.

3.2 Testing for Random Walks and Unit Roots

Given an $AR(p)$ model, $x_t = \alpha_1 x_{t-1} + ... + \alpha_p x_{t-p} + \varepsilon_t$, $\{x_t\}$ has a unit root if the polynomial $z^p = \alpha_1 z^{p-1} + ... + \alpha_{p-1} z + \alpha_p$ has a root on the unit circle (and none outside it). In particular, any $ARIMA(p, 1, q)$ process has a unit root and therefore is $I(1)$, and any $ARIMA(p, 2, q)$ process has two unit roots and is $I(2)$. We test for unit roots under the null hypothesis of a unit root, using the Dickey-Fuller unit root test.

The null hypothesis is that $x_t = x_{t-1} + \varepsilon_t$ and the alternative hypothesis is that $x_t = c + \rho x_{t-1} + \varepsilon_t$, with $|\rho| < 1$. The test statistic is simply the t-test from the regression of x_t on a constant and x_{t-1}:

$$\widehat{\tau}_\mu = \frac{\widehat{\rho} - 1}{se(\widehat{\rho})}$$

However, under the null hypothesis, $\widehat{\tau}_\mu$ does not have a standard normal distribution. Instead, one must use Dickey-Fuller-specific cutoff values (for example, -2.86 instead of -1.645). An augmented Dickey-Fuller test is used to distinguish between $ARIMA(p, 1, 0)$ and $ARIMA(p+1, 0, 0)$. The same test statistic (from a regression involving more lagged values) but different cutoff values are used in this case.

To test for a random walk with drift, the null hypothesis is $x_t = x + x_{t-1} + \varepsilon_t$, and the alternative hypothesis is $x_t = \alpha_0 + \alpha_1 t + y_t$, where $y_t = \rho y_{t-1} + \varepsilon_t$ (a trend-stationary process). In this case, we regress x_t on a constant, a trend, and x_{t-1}, and test whether the coefficient on x_{t-1} is 1. That is, we have a test statistic $\widehat{\tau}_\tau = \frac{\widehat{\rho}-1}{se(\widehat{\rho})}$, which again has a non-normal distribution.

3.2.1 Testing for Cointegration

Definition 3 *Suppose that two series, $\{x_{1,t}\}$ and $\{x_{2,t}\}$ are $I(1)$, but the series $\{x_{1,t} - \beta x_{2,t}\}$ is $I(0)$. Then we say that $\{x_{1,t}\}$ and $\{x_{2,t}\}$ are cointegrated, or that there is a stationary equilibrium.*

If we know or hypothesize a specific value for β, then we may test for cointegration by testing that the two original series have unit roots but that $\{x_{1,t} - \beta x_{2,t}\}$ does not, using the Dickey-Fuller test.

4 Long Memory Models and ARFIMA

In ARIMA models, the d term is always an integer. In ARFIMA models, we allow it to be a fraction (usually between 0 and 1). This means that models will have more memory than stationary ARMA models but less memory than $ARIMA(p, 1, q)$ models.

Definition 4 *Let B be the backshift (lag) operator, so that $Bx_t = x_{t-1}$. Then we may represent the difference operator as $\Delta = 1 - B$. We define the d^{th} difference as $\Delta^d = (1 - B)^d = \sum_{k=0}^{\infty} (-1)^k \binom{d}{k} B^k$, where $\binom{d}{k} = \frac{d(d-1)...(d-k+1)}{k!}$. This holds for both integer and fractional d.*

An ARFIMA model is one in which the number of differences is fractional. For example, an $ARFIMA(0,d,0)$ has $\Delta^d x_t = \varepsilon_t$. This gives us an $MA(\infty)$ representation of:

$$\begin{aligned} x_t &= \Delta^{-d}(\Delta x_t) = \Delta^{-d}\varepsilon_t = (1-B)^{-d}\varepsilon_t \\ &= \varepsilon_t + (-1)\binom{-d}{1}\varepsilon_{t-1} + (-1)^2\binom{-d}{2}\varepsilon_{t-2} + ... \\ &= \varepsilon_t + d\varepsilon_{t-1} + \frac{d(d-1)}{2}\varepsilon_{t-2} + \frac{d(d-1)(d-2)}{6}\varepsilon_{t-2} + ... \end{aligned}$$

The coefficients on this MA model die down more slowly than the exponential rate we have previously seen. This model is stationary and invertible if $-\frac{1}{2} < d < \frac{1}{2}$. Similarly, the $AR(\infty)$ representation is:

$$\begin{aligned} \Delta^d x_t &= \varepsilon_t \\ x_t - \binom{d}{1}x_{t-1} + \binom{d}{2}x_{t-2} - \binom{d}{3}x_{t-3} + ... &= \varepsilon_t \end{aligned}$$

$$\begin{aligned} x_t &= \varepsilon_t + \binom{d}{1}x_{t-1} - \binom{d}{2}x_{t-2} + \binom{d}{3}x_{t-3} - ... \\ &= \varepsilon_t + dx_{t-1} - \frac{d(d-1)}{2}x_{t-2} + \frac{d(d-1)(d-2)}{6}x_{t-3} - ... \end{aligned}$$

Again, the coefficients decay more slowly than an exponential decay.

In an ARFIMA model with $|d| < 0.5$, the autocorrelations have a power law decay: $\rho_k \propto k^{2d-1}$. This is a slower decay than for ARMA models, since k^{2d-1} decays more slowly than α^k. Because the autocorrelations decay more slowly, the standard error of sample means (and other such statistics) decay more slowly as well. In particular, $Var(\overline{X}_n) \propto n^{2d-1}$ instead of $Var(\overline{X}_n) \propto \frac{1}{n}$ when $d = 0$. Thus, getting precise estimates is (even) harder.

5 Non-Linear Models

Definition 5 *A stationary time series is linear if it can be represented as an $MA(\infty)$ model with independent shocks, $\{e_t\}$. That is, $X_t = e_t + a_1 e_{t-1} + a_2 e_{t-2} + ...$*

Definition 6 *The best linear forecast (in terms of mean squared errror), $f_{n,h}^{Lin}$ is the forecast such that $f_{n,h}^{Lin}$ is a linear combination of $x_n, x_{n-1}, ...$ and the forecast error $x_{n+h} - f_{n,h}^{Lin}$ is uncorrelated with all linear combinations of $x_n, x_{n-1}, ...$*

Definition 7 *The optimal forecast (in terms of mean squared error) is simply the conditional expectation, $E(x_{n+h}|x_n, x_{n-1}, ...)$.*

In all of these cases, notice that conditioning on all previous values of $\{x_t\}$ is equivalent to conditioning on all previous values of $\{\varepsilon_t\}$, since one may be derived from the other in a completely specified model.

Non-linear models improve forecasts when uncorrelated errors are not independent. That is, errors that cannot be predicted linearly may be able to be predicted in other ways. If the shocks are independent (that is, strict white noise), then the linear model is the best model, and the optimal linear forecast is also the optimal forecast. Note that all multivariate normal variables that are uncorrelated are also independent. Thus, any model with multivariate normal shocks is linear and any non-linear model has shocks that are not multivariate normal.

Definition 8 *A white noise process $\{\varepsilon_t\}$ which satisfies $E(\varepsilon_{n+1}|\varepsilon_n, \varepsilon_{n-1}) = 0$ is called a martingale difference.*

Definition 9 *A process $\{x_t\}$ is a martingale if, for all n and h, $E(x_{n+h}|x_n, x_{n-1}, ...) = 0$. The differences in this process are martingale differences.*

The optimal linear forecast is the optimal forecast if and only if the shocks are a martigale difference.

5.1 Bilinear models

Bilinear models are time series in which:

$$X_t + a_1 X_{t-1} + ... + a_p X_{t-p} = e_t + b_1 e_{t-1} + ... + b_q e_{t-q} + \sum_{i=1}^{I} \sum_{j=1}^{J} c_{ij} e_{t-i} X_{t-j}$$

where $\{e_t\}$ is independent white noise. Note that this is an ARMA model with the additional terms in $\sum_{i=1}^{I} \sum_{j=1}^{J} c_{ij} e_{t-i} X_{t-j}$. To test whether a series is non-linear, one may estimate this model and then test whether all the c_{ij} are zero.

5.2 Threshold Autoregressive Models

A threshold autoregressive model is given by:

$$X_t = \begin{cases} a^{(1)} X_{t-1} + \varepsilon_t^{(1)} & X_{t-1} < d \\ a^{(2)} X_{t-1} + \varepsilon_t^{(2)} & X_{t-1} \geq d \end{cases}$$

where d is the threshold value. Such a model can have "limit-cycles" if the model cycles between the regimes over the forecast period.

This can be generalized to multiple states. In the limit, this can become:

$$X_t = \lambda(X_{t-1}) + \varepsilon_t$$

where λ is a non-linear function. λ can be estimated with smoothing technology.

5.3 Chaos Theory

Suppose that $X_t = f(X_{t-1})$, where f is a non-stochastic mapping. Then paths are given by $\{X_0, f(X_0), f(f(X_0)), ...\}$. These paths may be periodic, explosive, or chaotic. If $f(x) = x$, then x is a fixed point. In the chaotic case, these paths can appear to be a stochastic time series, even though the path is entirely predictable.

6 Volatility and (G)ARCH Models

In general, it is possible that the variance (volatility) of a series is not constant over time. It may vary because of some underlying process or because of the effect of previous observations (as in ARCH models). The simplest non-parametric measure of volatility is the sum of squared returns at a high frequency over a short period. This is called the realized volatility. Plotting this (or its log) can show whether volatility seems to be constant or changing over time.

6.1 $ARCH(q)$ Model

The Autoregressive Conditional Heteroskedasticity model assumes that the variance of the shocks changes over time, conditional on the previous shocks. In particular, we have shocks, $\{\varepsilon_t\}$, such that:

$$\begin{aligned} & \varepsilon_t|\Psi_{t-1} \tilde{} Normal(0, h_t) \\ h_t &= \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 \end{aligned}$$

For stationarity and no possibility of having the series become constant at zero, we assume that $\omega > 0$, $\alpha_i \geq 0$, and $\sum_{i=1}^{q} \alpha_i < 1$. Then, $h_t \geq \omega > 0$ for all t. The conditional mean and variance are $E(\varepsilon_t|\Psi_{t-1}) = 0$ and $Var(\varepsilon_t|\Psi_{t-1}) = h_t$. The shocks are zero mean white noise, because their unconditional mean and variance are constant at $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \frac{\omega}{1-(\sum_{i=1}^{q} \alpha_i)}$. In addition, this is a martingale difference, but the shocks are not independent, because we may forecast the volatility. In this model, volatility can be quite persistent; one large shock increases the probability of a large shock in the next period as well. Since ε_t is a mixture of normal distributions, it is not Gaussian and has very heavy tails (how heavy they are depends on the α_i's).

If we let $\eta_t = \varepsilon_t^2 - h_t$, then $\{\eta_t\}$ is white noise and $\varepsilon_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + ... + \alpha_q \varepsilon_{t-q}^2 + \eta_t$ is an $AR(q)$ process. Note that $E(\varepsilon_t^2|\Psi_{t-1}) = h_t$, and $E(h_t) = E(E(\varepsilon_t^2|\Psi_{t-1})) = E(\varepsilon_t^2) = Var(\varepsilon_t) = \frac{\omega}{1-(\sum_{i=1}^{q} \alpha_i)}$.

6.2 The $ARIMA(k,l)$ Model with $ARCH(q)$ Errors

Suppose that the process $\{x_t\}$ obeys the model:

$$\begin{aligned} x_t &= a_1 x_{t-1} + ... + a_k x_{t-k} + b_1 \varepsilon_{t-1} + ... + b_l \varepsilon_{t-l} + \varepsilon_t \\ & \quad \varepsilon_t | \Psi_{t-1} \tilde{} Normal(0, h_t) \\ h_t &= \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 \end{aligned}$$

Then $\{x_t\}$ is $ARMA(k,l)$ with $ARCH(q)$ errors. We compute point forecasts just as we did before (since the shocks are martingale differences, these are still the optimal forecasts). However, the width of our forecast intervals change. Since $\varepsilon_{n+1} | \Psi_n \tilde{} Normal(0, h_{n+1})$, the correct width is $z_{\frac{\alpha}{2}} \sqrt{h_n}$, where $z_{\frac{\alpha}{2}}$ is the correct critical value in the normal distribution. Thus, the width depends on the recent volatility of the shocks (by way of the current variance).

However, we cannot estimate this more than one step ahead, because the two forecast errors come from normal distributions with different variances (and the variance of the second distribution depends on the first error). Simulation may help solve this problem.

6.2.1 Choosing the order of the ARCH Model

Since $\{\varepsilon_t\}$ is white noise, the ACF and PACF will be identically zero. However, $\{\varepsilon_t^2\}$ is an $AR(q)$ process. Thus, the ACF of $\{\varepsilon_t^2\}$ should die down, while the PACF should cut off after the q^{th} lag. In addition, the AICC can be used to choose a model. The formula in this case is:

$$AIC_C = -2 \log likelihood + 2(q+1) \frac{N}{N-q-2}$$

for a model that is zero mean white noise. (Add one to q if a constant term is being estimated as well.)

6.2.2 Combining ARCH and ARIMA

The simplest way to estimate an ARIMA+ARCH model is to estimate the ARIMA model and then use the residuals from that in the ARCH estimation.

However, the ARIMA model assumes that the errors are unconditionally normally distributed, which is not true in the ARCH model. This means that the standard errors of the ACF and PACF are wrong. In particular, the standard errors for an $ARCH(1)$ for the k^{th} autocorrelation should be

$$se(r_k) = \sqrt{\frac{1}{n}(1 + \frac{2\alpha_1^k}{1 - 3\alpha_1})}$$

Notice that this is not defined for $\alpha_1 \geq \frac{1}{\sqrt{3}}$; in that case, the tails are too fat for the variance of r_k to exist. In addition, the AICC is not exactly correct (but

should be close). This is only a problem in choosing the ARMA model; the AICC, ACF and PACF are valid for the squared errors in the ARCH model.

Ideally, one would estimate both the ARIMA and the ARCH parameters jointly. However, that is more complicated.

6.3 Extensions of the ARCH Model and other Conditional Heteroskedasticity Models

6.3.1 I-ARCH

In the I-ARCH model, we model:

$$h_t = \omega + \varepsilon_{t-1}^2$$

In this case, the unconditional variance is infinite and therefore the process is non-stationary. $\{\varepsilon_t^2\}$ is a random walk with drift, so that the forecasted volatility is not mean reverting and tends toward infinity.

6.3.2 ARCH-T

Sometimes, even with an ARCH correction, the (corrected) errors have tails that are too heavy. We try to correct this by assuming that the errors have a t-distribution instead of a normal distribution. (The number of degrees of freedom for the t-distribution must be specified, though.) In this case,

$$\frac{\varepsilon_t}{h_t} | \Psi_{t-1} \tilde{} t_k$$

instead of $\frac{\varepsilon_t}{h_t} | \Psi_{t-1} \tilde{} Normal(0,1)$.

6.3.3 GARCH Models

The GARCH model allows for more persistent volatility than a parsimonious ARCH model. The most common GARCH model is $GARCH(1,1)$, which is defined by:

$$\begin{aligned} & \varepsilon_t | \Psi_{t-1} \tilde{} Normal(0, h_t) \\ h_t &= \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1} \\ \alpha + \beta &< 1 \end{aligned}$$

In this case, ε_t^2 is $ARMA(1,1)$, with the AR term equal to $\alpha + \beta$ and the MA term equal to β. If $\alpha + \beta = 1$, then this is an I-GARCH model. Notice that $Var(\varepsilon_t) = \frac{\omega}{1-\alpha-\beta}$, and h_t is an exponentially decaying moving average of $\{\varepsilon_t^2\}$.

Some other models also relate predicted volatility ($\widehat{\sigma}_T^2$) to squared returns

(R_T^2):

$$\begin{aligned}
\widehat{\sigma}_T^2 &= (1-\theta)\sum_{i=1}^{\infty}\theta^{i-1}R_{T+1-i}^2 \\
\widehat{\sigma}_T^2 &= \alpha_0 + \sum_{i=1}^{p}\alpha_i R_{T+1-i}^2 \\
\widehat{\sigma}_T^2 &= \alpha_0 + \sum_{i=1}^{p}\alpha_i R_{T+1-i}^2 + \sum_{i=1}^{q}\beta_i\widehat{\sigma}_{T-i}^2 \\
&= \delta_0 + \sum_{i=1}^{\infty}\delta_i R_{T+1-i}^2
\end{aligned}$$

6.3.4 Stochastic Volatility Models

Suppose that e_t are independent shocks and $\{h_t\}$ is a latent, NOT observation driven process. Then we model returns as

$$r_t = \exp(h_t)e_t$$

However, using this assumes that the process $\{h_t\}$ is known. Since $\{h_t\}$ can be any process, it can be adapted to any purpose (such as long memory).

6.3.5 Long Memory Volatility (FIGARCH)

For longer memory of volatility, one can use fractionally integrated GARCH. In this case,

$$\begin{aligned}
\varepsilon_t|\Psi_{t-1} \tilde{} N(0,h_t) \\
h_t &= \omega + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + ...
\end{aligned}$$

where the α_i decay as the $AR(\infty)$ coefficients of an $ARFIMA(1,d,0)$.

6.4 Aggregating Risk

The total risk over a group of time periods differs from the risk over a single time period. If observations are independent and identically distributed, then the variance over k periods is k times the variance in one period, since $Var(r_1+...+r_k) = Var(r_1)+...+Var(r_k) = k*Var(r)$. However, under the GARCH model, the volatility is not constant and the volatility one period affects volatility the next period. If the returns, $\{\varepsilon_t\}$ are martingale differences, then $Var(\varepsilon_{n+1} + ... + \varepsilon_{n+h}|\Psi_n) = \sum_{i=1}^{h} Var(\varepsilon_{n+i}|\Psi_n) = E(\varepsilon_{n+1}^2|\Psi_n) + ... + E(\varepsilon_{n+h}^2|\Psi_n)$. As $h \to \infty$, $E(\varepsilon_{n+h}^2|\Psi_n)$ approaches the unconditional variance. Because of this tendency toward the unconditional variance, conditional heteroskedasticity can be harder to detect if measurements are taken farther apart.

7 Seasonality

A seasonal effect is one that occurs identically at certain intervals (such as months or days of the week). One can often detect seasonality in evenly-spaced spikes in the ACF or PACF of the level or the difference.

The simplest method to deal with seasonality is to subtract out monthly averages from each observation. However, this assumes that there is no change in seasonal patterns over time. The Census has procedures called X-11 and X-12 that are complex but built into SAS for seasonal adjustment.

Exponential smoothing can be used to account for slowly changing seasonal factors. In this, we define the current monthly effect as a weighted average of the previous calculated effect and the current observation. That is,

$$Effect_t = \alpha(Effect_{t-1}) + (1-\alpha)x_t$$

Alternatively, we may use SARIMA. We define seasonal differences as $y_t^{(M)} = x_t - x_{t-M}$, where M is the period (such as 12 for monthly data). This should remove any seasonal components that repeat exactly. However, this leads to a seasonal difference of the stationary parts, so autocorrelations as lag M might be created. For example, if $S(t)$ is a seasonal component and z_t is the non-seasonal component:

$$\begin{aligned} y_t^{(M)} &= (S(t) + z_t) - (S(t-M) + z_{t-M}) \\ &= z_t - z_{t-M} \end{aligned}$$

A more complex form is Seasonal Multiplicative ARIMA, described by $(p,d,q) \times (P,D,Q)_S$. In this model, the values for a fixed period (such as Janurary) follow an $ARIMA(P,D,Q)$ model; the model is the same for every different period. The shocks across months follow an $ARIMA(p,d,q)$ model. This is a more parsimonious model than just including M or $2M$ lags.