14 Time Series

14.1 Stationarity

A (multivariate) time series y_t is an $m \times 1$ vector observed over time $t = 1, ..., n$. We think of the sample as a "window" out of an infinite past and infinite future.

A series y_t is strictly stationary if the joint distribution $(y_t, ..., y_{t+h})$ is constant across t for all h. An implication of stationary is that any finite moment is time-invariant.

A linear measure of dependence is autocovariance

$$\begin{aligned} \gamma(k) &= cov\left(y_t, y_{t-k}\right) \\ &= E\left(\left(y_t - Ey_t\right)\left(y_{t-k} - Ey_{t-k}\right)'\right) \end{aligned}$$

Stationarity implies that $\gamma(k)$ is constant over time t.

A loose definition of ergodicity is that $\gamma(k) \to 0$ as $k \to \infty$. A rigorous definition requires a measure-theoretic treatment, but intuitively that history is independent of the infinite past.

Simple time series models are built using fundamental "shocks" e_t, typically normalized so that $Ee_t = 0$.

(1) The most basic shock e_t is iid.

(2) Martingale Difference Sequence (MDS). e_t is a MDS if $E\left(e_t \mid e_{t-1}, e_{t-2}, ...\right) = 0$

(3) White noise. If $Ee_t e_{t-h} = 0$ for all $h \geq 1$.

An iid shock is a MDS, and a MDS is white noise, but the reverse is not true. An iid shock is unpredictable. A MDS is unpredictable in the mean. A white noise shock is linearly unpredictable.

A process is m-dependent if y_t and y_{t+k} are independent for $k > m$. In this case, $\gamma(k) = 0$ for $k > m$.

A simple model of time dependence is a moving average. A MA(1) is

$$y_t = e_t + \theta e_{t-1}$$

with e_t white noise. You can calculate that $\gamma(1) = \theta\sigma_e^2$. An MA(q) is q-dependent.

Another simple model is an autoregression. An AR(1) is

$$y_t = \alpha y_{t-1} + e_t$$

where e_t is iid. The series is stationary if $|\alpha| < 1$. You can calculate that $\sigma_y^2 = \gamma(1) = \sigma_e^2/(1 - \alpha^2)$ and $\gamma(k) = \alpha^k \sigma_y^2$. An AR process is not m-depenent.

An example of a MDS which is not iid is an ARCH process

$$\begin{aligned} e_t &= \sigma_t z_t \\ \sigma_t^2 &= \omega + \alpha e_{t-1}^2 \end{aligned}$$

or a GARCH(1,1)

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha e_{t-1}^2$$

where z_t is an iid shock. This series is predictable in the square (in the variance) but not in the mean.

An example of a white noise process which is not an MDS is the nonlinear MA

$$y_t = e_t + e_{t-1}e_{t-2}$$

You can calculate that this is white noise. Yet $E\left(e_t \mid e_{t-1}, e_{t-2}, ...\right) = e_{t-1}e_{t-2} \neq 0$. This is a nonlinear process.

14.2 Time Series Averages

Let $\theta = Ey_t$ be estimated by

$$\hat{\theta} = \frac{1}{n}\sum_{t=1}^{n} y_t$$

If y_t is stationary,

$$E\hat{\theta} = Ey_t = \theta$$

so the estimator is unbiased.

The variance of the standardized estimator is

$$\begin{aligned} var(\sqrt{n}\left(\hat{\theta} - \theta\right)) &= \frac{1}{n}E\left(\sum_{t=1}^{n}\left(y_t - \theta\right)\right)\left(\sum_{t=1}^{n}\left(y_t - \theta\right)\right)' \\ &= \frac{1}{n}E\left(\sum_{t=1}^{n}\sum_{j=1}^{n}\left(y_t - \theta\right)\left(y_j - \theta\right)'\right) \\ &= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{n}\gamma\left(t - j\right) \end{aligned}$$

This can be simplified to equal

$$\sum_{k=-(n-1)}^{n-1}\left(\frac{n-k}{n}\right)\gamma(j)$$

Notice that it is bounded by the inequality

$$var(\sqrt{n}\left(\hat{\theta} - \theta\right)) \leq \sum_{k=-\infty}^{\infty}|\gamma(j)|$$

Hence, if $\sum_{k=0}^{\infty}|\gamma(j)| < \infty$, then $var(\sqrt{n}\left(\hat{\theta} - \theta\right))$ is bounded. By Markov's inequality, it follows that $\hat{\theta} \to_p \theta$. This is a rather simple proof of consistency for time-series averages. This proof uses

stronger conditions than necessary.

The Ergodic Theorem states that if y_t is strictly stationary and ergodic, and $E|y_t| < \infty$, then $\hat{\theta} \to \theta$ a.s.

For a central limit theorem, we need a stronger summability condition on the covariances. As we showed above,

$$\begin{aligned} var(\sqrt{n}\left(\hat{\theta} - \theta\right)) &= \sum_{k=-(n-1)}^{n-1} \left(\frac{n-k}{n}\right)\gamma(j) \\ &\to \sum_{k=-\infty}^{\infty} \gamma(j) \\ &\equiv \Omega \end{aligned}$$

as $n \to \infty$. Thus the asymptotic variance of $\hat{\theta}$ is Ω, not $var(y_t)$. Under regularity conditions the estimator is asymptotically normal

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \to_d N(0, \Omega)$$

The variance Ω is sometimes called the "long-run covariance matrix" and is also a scale of the spectral density of y_t at frequency zero.

The fact that Ω involves an infinite sum of covariances means that standard error calculations for time-series averages needs to take this into account. Estimation of Ω is called "HAC" estimation in econometrics (for heteroskedasticity and autocorrelation consistent covariance matrix estimation), and is often called the "Newey-West estimator" due to an early influential paper (Econometrica, 1987) by Whitney Newey and Ken West.

14.3 GMM

This carries over to GMM estimation. If θ is the solution to

$$Em(y_t, \theta) = 0$$

where m is a known function, the GMM estimator minimizes a quadratic funtion in the sample moment of $m(y_t, \theta)$ to find $\hat{\theta}$. The asymptotic distribution of $\hat{\theta}$ is determined by the sample average of $m(y_t, \theta_0)$ at the true value θ_0, so if these are autocorrelated, then the asymptotic distribution of $\hat{\theta}$ will involve the long-run covariance matrix.

Theorem 1 *Under general regularity conditions the GMM estimator for stationary time-series data satisfies*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\right)$$

where

$$\boldsymbol{G} = \mathrm{E}\frac{\partial}{\partial \boldsymbol{\theta}'} m(y_t, \theta_0)$$

and

$$\boldsymbol{\Omega} = \sum_{k=-\infty}^{\infty} \mathrm{E}\left(\boldsymbol{m}_t \boldsymbol{m}_{t-k}'\right).$$

A important simplication occurs when $m(y_t, \theta_0)$ is serially uncorrelated. This occurs in dynamically well-specified models or correctly-specified MLE. In this case, $m(y_t, \theta_0)$ will be a MDS, serially uncorrelated, and thus

$$\boldsymbol{\Omega} = \mathrm{E}\left(\boldsymbol{m}_t \boldsymbol{m}_t'\right).$$

14.4 Linear Regression

Consider linear regression

$$y_t = \theta' x_{t-1} + e_t$$

This includes autoregressions, as x_{t-1} can include $y_{t-1}, y_{t-2}, ...$

In this model, the LS estimator of θ is GMM, using the moment condition

$$m(y_t, x_{t-1}, \theta) = x_{t-1}\left(y_t - \theta' x_{t-1}\right)$$

Note that at the true θ_0,

$$m(y_t, x_{t-1}, \theta_0) = x_{t-1} e_t$$

The properties of m_t depend on the properties of e_t.

If the model is a true regression, then $\theta' x_{t-1}$ is the conditional mean, and the error is conditionally mean zero and thus a MDS:

$$E\left(e_t \mid I_{t-1}\right) = 0$$

where I_{t-1} contains all lagged information. In this case m_t is a MDS as well:

$$E\left(m_t \mid I_{t-1}\right) = E\left(x_{t-1} e_t \mid I_{t-1}\right) = x_{t-1} E\left(e_t \mid I_{t-1}\right) = 0$$

Thus the LS estimator is asymptotically normal, with a conventional covariance matrix.

On the other hand, if the model is an approximation, a linear projection, then $\theta' x_{t-1}$ is not necessarily the conditional mean, so e_t is not necessarily a MDS. Then m_t will not necessarily be serially uncorrelated, and Ω will contain the autocovariances of m_t

14.5 Density Estimation

Suppose y_t is univariate and strictly stationary with marginal distribution $F(y)$ with density $f(y)$. The kernel density estimator of $f(y)$ is

$$\hat{f}(y) = \frac{1}{nh} \sum_{t=1}^{n} k\left(\frac{y - y_t}{h}\right)$$

where $k(u)$ is a kernel function and h is a bandwidth.

As the function is linear, the expectation is not affected by time-series dependence. That is,

$$E\hat{f}(y) = E\frac{1}{h}k\left(\frac{y-y_t}{h}\right)$$

and this is the same as in the cross-section case. Hence the bias of $\hat{f}(y)$ is unchanged by dependence.

To calculate the variance, for simplicity assume that y_t is m-depenent. The variance is

$$\begin{aligned}
var(\hat{f}(y) &= E\left(\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{h}k\left(\frac{y-y_t}{h}\right) - E\frac{1}{h}k\left(\frac{y-y_t}{h}\right)\right)\right)^2 \\
&= \frac{1}{n^2}\sum_{t=1}^{n}\sum_{j=1}^{n} cov\left(\frac{1}{h}k\left(\frac{y-y_t}{h}\right), \frac{1}{h}k\left(\frac{y-y_j}{h}\right)\right) \\
&= \frac{1}{n}\sum_{k=-(n-1)}^{n-1}\left(\frac{n-k}{n}\right)\frac{1}{h^2}cov\left(k\left(\frac{y-y_t}{h}\right), k\left(\frac{y-y_{t-k}}{h}\right)\right) \\
&= \frac{1}{n}\sum_{k=-m}^{m}\left(\frac{n-k}{n}\right)\frac{1}{h^2}cov\left(k\left(\frac{y-y_t}{h}\right), k\left(\frac{y-y_{t-k}}{h}\right)\right) \\
&\simeq \frac{1}{nh^2}Ek\left(\frac{y-y_t}{h}\right)^2 + \frac{2}{n}\sum_{k=1}^{m}\frac{1}{h^2}Ek\left(\frac{y-y_t}{h}\right)k\left(\frac{y-y_{t-k}}{h}\right)
\end{aligned}$$

The second-to-last step uses m-dependence. The first part in the final line is the same as in the cross section case, and is asymptotically $\dfrac{R(k)f(y)}{nh}$.

Now take the second part in the final line, which is the sum of the m components. Let $f(u_0, u_k)$ be the joint density of (y_t, y_{t-k}) for $k > 0$. Then

$$\begin{aligned}
\frac{1}{h^2}Ek\left(\frac{y-y_t}{h}\right)k\left(\frac{y-y_{t-k}}{h}\right) &= \int\int \frac{1}{h^2}k\left(\frac{y-u_0}{h}\right)k\left(\frac{y-u_k}{h}\right)f(u_0,u_k)\,du_0du_k \\
&= \int\int k(v_0)\,k(v_k)\,f(y-hv_0, y-hv_k)\,dv_0dv_k
\end{aligned}$$

where I made two change-of-variables, $u_0 = y - hv_0$ and $u_k = y - hv_k$, which has Jacobian h^2. Expanding the joint density and integrating, this equals

$$\int k(v_0)\,dv_0 \int k(v_k)\,dv_k f(y,y) + o(1) = f(y,y)$$

the joint density, evaluated at (y, y). We have found that

$$\begin{aligned}
var(\hat{f}(y) &= \frac{R(k)f(y)}{nh} + \frac{2m}{n}f(y,y) + o\left(\frac{1}{n}\right) \\
&\simeq \frac{R(k)f(y)}{nh}
\end{aligned}$$

which is dominated by the same term as in the cross-section case. Time-series dependence has no effect on the asymptotic bias and variance of the kernel estimator! This is in strong contrast to the parametric case, where correlation affects the asymptotic variance.

The technical requirement is that the joint density of the observations (y_t, y_{t-k}) is smooth at (y, y). In the cross-section case we only required smoothness of the marginal density. In the time series case we need smoothness of the joint densities, even though we are just estimating a marginal density.

The intuition is that kernel (nonparametric) estimation is averaging the data locally in the $y-$dimension, where there is no time-series dependence, not in the time-dimension. That is, $\hat{f}(y)$ is a average of observations y_t that are close to y. This subset of observations are not necessarily close to each other in time. Thus they have low joint dependence, and the contribution of joint dependence to the asymptotic variance is of small order relative to the nonparametric smoothing.

The theoretical implication of this result is that the theory of nonparametric kernel density estimation carries over from the iid case to the time-series case without essential modification. (However, proofs of the theorems requires attention to dependence and stronger smoothness conditions.)

14.6 One-Step-Ahead Point Forecasting

Given information up to n, we want a forecast f_{n+1} of y_{n+1}. What does this mean? We want f_{n+1} to be "close" to the realized y_{n+1}. Let $L(f, y)$ denote the loss associated with a forecast f of realized y. The risk of the estimator is the expected loss $R(f \mid I_n) = E\left(L(f, y) \mid I_n\right)$. A convenient loss function is quadratic $L(f, y, I_n) = (f - y)^2$ in which case the best forecast is the conditional mean $f = E\left(y_{n+1} \mid I_n\right)$. In this sense it is common for a point forecast for y_{n+1} to be an estimate of the conditional mean.

Let $x_{t-1} = (y_{t-1}, y_{t-2}, ..., y_{t-q})$, and assume that the conditional mean of y_t is (approximately) a function only of x_{t-1}. Then

$$\begin{aligned} y_t &= g\left(x_{t-1}\right) + e_t \\ E\left(e_t \mid I_{t-1}\right) &= 0 \end{aligned}$$

(One issue in nonparametrics is letting $q \to \infty$ as $n \to \infty$ so to allow for infinite dependence.) Then the optimal point forecast (given quadratic loss) for y_{n+1} is $g(x_n)$. A feasible point forecast is $\hat{g}(x_n)$ where $\hat{g}(x)$ is an estimator of g.

The conventional linear approach is to set $\hat{g}(x) = x'\hat{\theta}$ where $\hat{\theta}$ is LS of y on X. This imposes a linear model.

Any other nonparametric estimator can be used. In particular, a local linear estimator nests the linear (AR) model as a special case, but allows for nonlinearity.

14.7 One-Step-Ahead Interval Forecasting

An interval forecast is $\hat{C} = [f_1, f_2]$. The goal is select $\hat{C}$ so that $P(y_{n+1} \in \hat{C}) = .9$ (or some other pre-specified coverage probability) and so that $\hat{C}$ is as short as possible. Often people don't mention the goal of a short $\hat{C}$. But this is critical, otherwise we can design silly $\hat{C}$ with the desired coverage. For example, let $\hat{C} = \mathbb{R}$ with probability .9, and $\hat{C} = \hat{g}(x_n)$ with probability .1. This has exact coverage .9 but is clearly not using sample information intelligently.

A desired forecast inteval sets f_1 and f_2 equal to the .05 and .95 quantiles of the conditional distribution of y given x.

We discussed this problem earlier. Given an estimate $\hat{g}(x)$ for the conditional, set $\hat{e}_t = y_t - \hat{g}(x_{t-1})$ and estimate the CDF $\hat{F}(e \mid x)$, conditional distribution of $\hat{e}_t$ given x_{t-1}. (As a special case, you can use the unconditional DF $\hat{F}(e)$.) Let $\hat{q}_\alpha(x)$ be the α quantile of this conditional distribution. Then setting $\hat{f}_1 = \hat{g}(x_n) + \hat{q}_{.05}(x_n)$ and $\hat{f}_2 = \hat{g}(x_n) + \hat{q}_{.95}(x_n)$ the forecast interval is $\hat{C} = [\hat{f}_1, \hat{f}_2]$.

An alternative method to display forecast uncertainty is to use a density forecast. Let $f(y \mid x)$ denote the conditional density of y_t given x_{t-1}. An estimate of this conditional density is

$$\hat{f}(y \mid x) = \hat{f}_e(y - \hat{g}(x) \mid x)$$

where $\hat{f}_e$ is a (kernel) estimate of the conditional density of $\hat{e}_t$ given x_t. (As a special case you can use the unconditional density estimate, which is approximating the error e_t as being independent of x_{t-1}). The forecast density is then

$$\hat{f}(y \mid x_n) = \hat{f}_e(y - \hat{g}(x_n) \mid x_n)$$

14.8 Multi-Step-Ahead Forecasting

Practical forecasts are typically for multiple periods out of sample. Let $h \geq 1$ be the forecast horizon (a positive integer). It is convenient to switch notation and write the problem as forecasting y_{t+h} given x_t.

Given squared error loss, the optimal forecast is the conditional mean $f = E(y_{n+h} \mid I_n)$.

There are two main approaches to multi-step forecasting.

One method, the direct approach, is to specify a model for the h-step conditional mean, e.g.

$$\begin{aligned} y_{t+h} &= g(x_t) + e_{t+h} \\ E(e_{t+h} \mid I_t) &= 0 \end{aligned}$$

This is estimated by a (parametric or nonparametric) regression of y_{t+h} on x_t. The forecast is then $f = \hat{g}(x_n)$.

This requires a different "model" for each forecast horizon h, and could even imply different selected x_t for different horizons.

The strengths of this approach is that it directly models the object of interest. It is believed to

be relatively robust to misspecification. One weakness is that the error e_{t+h} cannot be uncorrelated whe $h > 1$. It is necessarily a MA$(h-1)$ process. This might invalidate conventional order selection methods (this is an open question)

The other method, called the "iterated" or "plug-in" approach, is to estimate a one-step-ahead forecast, and iterate. When $g(x) = x'\theta$ is linear and the goal is point forecasting, this is relatively simple, as iterated mean forecasts are functions only of θ. But for forecast intervals or when $g(x)$ is nonlinear, θ is insufficient. The only method is to estimate the one-step-ahead distribution

$$\hat{F}(y \mid x) = \hat{F}_e\left(y - \hat{g}(x) \mid x\right)$$

(or density), and iterate the entire one-step-ahead distribution. As this involves h-fold integration, it is easiest accomplished through simulation.

If $\hat{F}_e$ is estimated by NW, it can be written as

$$\hat{F}_e\left(e \mid x\right) = \sum_{t=1}^{n} p_t(x) 1\left(\hat{e}_t \leq e\right)$$

where

$$p_t(x) = \frac{K\left(H^{-1}\left(X_t - x\right)\right)}{\sum_{j=1}^{n} K\left(H^{-1}\left(X_j - x\right)\right)}$$

This is a discrete distribution, with probability mass $p_t(x)$ at $\hat{e}_t$. Thus to draw from $\hat{F}_e\left(e \mid x\right)$ is simply to draw from the empirical distribution of the residuals $\hat{e}_t$, with the weighted probabilities $p_t(x)$. (The simpliest case treats the error as independent of x_t, in which case $p_t(x) = n^{-1}$ for all t.)

[If a smoothed distribution estimator is used, then a simulation draw from the kernel is added.]

To make a draw from $\hat{F}(y \mid x)$, draw e^* from $\hat{F}_e(e \mid x)$ as described above, and then set $y^* = \hat{g}(x) + e^*$. Then y^* has the conditional distribution $\hat{F}(y \mid x)$.

To make multi-step-ahead draws:

1. Given x_n, draw e_{n+1} from $\hat{F}_e(e \mid x_n)$ and set $y_{n+1} = \hat{g}(x_n) + e_{n+1}$.
2. Define $x_{n+1} = (y_{n+1}, x_n)$.
3. Given x_{n+1}, draw e_{n+2} from $\hat{F}_e(e \mid x_{n+1})$ and set $y_{n+2} = \hat{g}(x_{n+1}) + e_{n+2}$.
4. Iterate until you attain y_{n+h}.

This creates one draw from the estimated h-step-ahead distribution, say $y_{b,n+h}$. Repeat for $b = 1, ..., B$, similar to bootstrapping.

The point forecast, the conditional mean, is the expected value of the conditional distribution, so is estimated by the average of the simulations

$$f_h = \frac{1}{B} \sum_{b=1}^{B} y_{b,n+h}.$$

A 90% forecast interval is constructed from the 5% and 95% quantiles of the $y_{b,n+h}$. A forecast density can be calculated by applying a kernel density estimator to $y_{b,n+h}$.

This procedure actually creates a joint distribution on the multi-step-ahead conditional distribution $(y_{n+1}, ..., y_{n+h})$. Perhaps this could be constructively used for some purpose. (e.g., what is the probability that unemployment rate will remain above 7% for all of the next 12 months?)

The acknowledged good feature of the iterative method (in parametric models) is that it is more accurate when the one-step-ahead distribution is correctly specified. The ackowledged downside is that it is is believed to be non-robust to misspecification. Errors in the one-step-ahead distribution can be magnifide when iterated multiple times. It is unclear how these statements apply to the nonparametric setting. The models are both correctly specified (as the bandwidth decreases the models become more accurate) yet are explcitly misspecified (in finite samples, any fitted nonparametric model is incomplete and biased).

14.9 Model Selection

Information criterion are widely used to select linear forecasting models.

Suppose the K'th model has K parameters and residual variance $\hat{\sigma}_K^2 = n^{-1}\sum \hat{e}_t^2$ where $\hat{e}_t = y_t - \hat{\theta}' x_{t-1}$

Popular methods in econometrics include AIC, BIC

$$\begin{aligned} AIC_K &= n\ln\left(\hat{\sigma}_k^2\right) + 2K \\ BIC_K &= n\ln\left(\hat{\sigma}_k^2\right) + \ln(n)K \end{aligned}$$

Less popular in econometrics, but widely used in time-series more generally, is Predictive Least Squares (PLS) introduced by Rissanen

$$\begin{aligned} PLS_K &= \sum_{t=P}^{n} \tilde{e}_t^2 \\ \tilde{e}_t &= y_t - x_{t-1}'\tilde{\theta}_{t-1} \\ \tilde{\theta}_{t-1} &= \left(\sum_{j=1}^{t-1} x_{j-1}x_{j-1}'\right)^{-1} \sum_{j=1}^{t-1} x_{j-1}y_j \end{aligned}$$

This is a time-series genearlization of CV. $\tilde{e}_t$ is a predictive residual. The sequential estimate $\tilde{\theta}_{t-1}$ uses observations up to $t-1$ for a one-step forecast. The PLS criterion is the sum of squared out-of-sample prediction errors. The PLS criterion needs a start-up sub-sample P before evaluating the first residual. Unfortunately the criterion can be sensitive to P, and there is no good guide for its selection.

While PLS is not typically used in econometrics for explicit model selection, it is very commonly used for model comparison. Models are frequently compared by so-called "out of sample performance". In practice, this involves comparing the PLS criterion across models. When this is

done, it is frequently described as if this is an "objective" comparison of performance. In fact, it is just a comparison of the PLS criterion. While a good criterion, it is not necessarily superior to other criterion, and is not at all infallible as a model selection criterion.

It is widely asserted that these methods can be applied to the direct multi-step-ahead context. I am skeptical, as the proofs are typically omitted, and the multi-step-ahead model has correlated errors. This may be worth investigating.