8 Univariate Time Series Analysis

The previous chapters dealt with cross-sectional data, i.e. we had a sample $\{w_i\}_{i=1:n}$, with $w_i = (y_i, x_i')$ for example. With time series data, we have $\{w_t\}_{t=1:T}$. This chapter deals with univariate processes, i.e. w_t is a scalar, usually written as y_t (a single variable over time). This is more than just a change in the indexing letter; t denotes time, which implies that there is a natural ordering among the observations, unlike in the cross-sectional case. It also implies that we (usually) cannot think of the observations as being independent. For example, the unemplyoment rate in the second quarter of 2023 is informative of its value in the following quarter. The dependence of observations over time requires us to define a few new concepts to be able to analyze time series data, but it also allows us to look at new, dynamic properties of our data.

This chapter introduces the basics of time series econometrics by discussing univariate processes. First, Section 8.1 introduces a few foundational concepts, before Section 8.2 presents the General Linear Process (GLP) as well as ARMA(p, q) models – a popular class of models to approximate the GLP. Section 8.3 discusses how to deal with trends and seasonalities in raw data. Finally, Section 8.4 highlights the peculiarities that arise when estimating regressions using time series as opposed to cross-sectional data.

8.1 Time Series Econometrics: Fundamentals

Under cross-sectional data, we think of each y_i (and x_i) as a RV. The i.i.d. assumption tells us that our sample consists of n independent realizations of the same RV. As a result, when we take expectations in the cross-sectional context, we "average over cross-sectional units i"; $\mathbb{E}[y_i] = \int y_i f_Y(y_i) dy_i$ is not a function of i because $f_{Y_i} = f_Y$ is the same for all i. Under time series data, we also think of each y_t as a RV. However, without the i.i.d. assumption, we generally have T realizations of different and mutually dependent RVs. When we take the expectation of y_t, we average over all possible observations that we could have obtained at this particular period t; $\mathbb{E}[y_t] = \int y_t f_{Y_t}(y_t) dy_t = \mu_t$ can be a function of t because f_{Y_t} can be different across t. In a similar way, we obtain the variance $\gamma_{0,t}$ and the autocovariance $\gamma_{h,t}$, both of which can be different across time:

$$\gamma_{0,t} = \mathbb{E}\left[(Y_t - \mu_t)^2\right] = \int (y_t - \mu_t)^2 f_{Y_t}(y_t) dy_t \ ,$$

$$\gamma_{h,t} = \mathbb{E}\left[(y_t - \mu_t)(y_{t-h} - \mu_{t-h})\right] = \int\int (y_t - \mu_t)(y_{t-h} - \mu_{t-h}) f_{Y_t,Y_{t-h}}(y_t, y_{t-h}) dy_t dy_{t-h} \ .$$

In order to be able to analyze the process y_t,[1] we have to ensure that the probabilistic structure governing the dynamics is stable over time. All work in applied time series assumes either stationarity or some specific, predefined departure from it (see examples in Section 9.3.1).

Definition 39. *The (univariate) process y_t is weakly stationary (WS) if neither its mean nor its autocovariances depend on the time period t:*

1. $\mu_t = \mu \quad \forall\, t$, *and*
2. $\gamma_{h,t} = \gamma_h \quad \forall\, t$.[2]

The second condition means that the autocovariance $\gamma_{h,t}$ only depends on the number of periods h separating the two observations y_t and y_{t-h} (called "displacement") and is the same regardless of which period t we are looking at. Note that then $\gamma_h = \gamma_{-h}$.[3] We call $\{\gamma_h\}_{h=0,1,2,...}$ the autocovariance function (ACF) and $\{\rho_h\}_{h=0,1,2,...}$ for $\rho_h = \gamma_h/\gamma_0$ the autocorrelation function.

Intuitively, weak stationarity replaces the i.i.d. assumption made under cross-sectional data. Regarding the i.d.-part, it ensures that (at least) the first and second moments are the same for all RVs $\{Y_t\}_{t=1:T}$. It also puts some structure on the departure from independence by ensuring that the (linear) dependence between y_t and y_{t-h} is the same as that between y_τ and $y_{\tau-h}$ (for all τ). A stronger concept is strict stationarity.

[1] As in the previous chapters, I write y_t to denote the RV at the particular time t as well as its realization. Capital letters are reserved to emphasize the distinction between scalars, vectors and matrices rather than RVs and their realizations. In addition, as is common practice, if no confusion arises, I write y_t as shorthand for $\{..., y_{-1}, y_0, y_1, ..., y_T, y_{T+1}, ...\}$ to denote the whole process, i.e. set of RVs. Sometimes it is useful to think of the process as starting in the infinite past and extending to the infinite future, rather than being limited to the sample periods $t = 1 : T$.

[2] Sometimes this is called covariance stationarity. Also, some definitions add the requirement that the variance is finite: $0 < \gamma(0) < \infty$. By the Cauchy-Schwarz Inequality (see Appendix to Chapter 1), all γ_h are then bounded as well, because it holds that: $|\gamma_h| \leqslant \gamma_0 \,\forall\, h$.

[3] $\gamma_h = \mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)] = \mathbb{E}[(y_{t+h} - \mu)(y_{t-h+h} - \mu)] = \mathbb{E}[(y_t - \mu)(y_{t+h} - \mu)] = \gamma_{-h}$.

Definition 40. *A process is strictly stationary (SS) if for any values $h_1, \dots, h_k$, the joint distribution of $(y_t, y_{t-h_1}, \dots, y_{t-h_k})$ depends only on the intervals separating the dates (displacements) $h_1, \dots, h_k$ and not on the date t itself, i.e.*

$$f_{Y_t, Y_{t-h_1}, \dots, Y_{t-h_k}} = f_{Y_\tau, Y_{\tau-h_1}, \dots, Y_{\tau-h_k}} \quad \forall\, \tau \ .$$

For example, strict stationarity implies that the joint distribution of (y_{13}, y_9, y_1) is the same as that of $(y_{113}, y_{109}, y_{101})$. In contrast, weak stationarity only ensures that the first two moments of the two distributions are the same over time;[4] higher-order moments, such as $\mathbb{E}[y_t^3]$ or $\mathbb{E}[y_t y_{t-4}^2]$, could still depend on time t. Clearly, strict stationarity implies weak stationarity. If the process y_t is Gaussian – i.e. if the distribution $f_{Y_t, Y_{t-h_1}, \dots, Y_{t-h_k}}$ is multivariate Normal for all t and for all displacements $h_1, \dots, h_k$ –, then the reverse holds true as well; a weakly stationary and Gaussian process is strictly stationary, because the Normal distribution is fully specified by the first two moments.

Another useful concept is ergodicity. Intuitively, it ensures that observations become independent when we consider large enough displacements.

Definition 41. *A SS process y_t is said to be ergodic if for any two bounded and measurable functions $f : \mathbb{R}^k \to \mathbb{R}$ and $g : \mathbb{R}^l \to \mathbb{R}$, we have*

$$\lim_{n\to\infty} \left| \mathbb{E}\left[f(y_{t_1}, ..., y_{t_k}) g(y_{t_1+n}, ..., y_{t_l+n})\right] \right| - \left| \mathbb{E}\left[f(y_{t_1}, ..., y_{t_k})\right] \right| \left| \mathbb{E}\left[g(y_{t_1+n}, ..., y_{t_l+n})\right] \right| = 0 \ ,$$

for all $t_1, ..., t_k$ (taking $k > l$, w.l.o.g.).

As a result, if we have a large enough sample, it is as if we had several realizations for the RVs that characterize our process, i.e. as if we had observed several samples instead of just one. This allows us to apply a LLN and a CLT.

Proposition 31 (LLN for SS & Ergodic Processes)**.**
If the process y_t is SS and ergodic with $\mathbb{E}[y_t] = \mu < \infty$, then $\frac{1}{T}\sum_{t=1}^T y_t \xrightarrow{p} \mu$.

[4] Note that the variance (covariance matrix) of the random vector (y_t, y_{t-4}, y_{t-12}) is given by $\begin{bmatrix} \gamma_{0,t} & \gamma_{4,t} & \gamma_{12,t} \\ \gamma_{-4,t} & \gamma_{0,t} & \gamma_{8,t} \\ \gamma_{-12,t} & \gamma_{-8,t} & \gamma_{0,t} \end{bmatrix}$. Under stationarity, we can drop the t-subscript, and we know $\gamma_h = \gamma_{-h}$.

Proposition 32 (CLT for SS & Ergodic Processes).
If the process y_t is SS and ergodic with $\mathbb{E}[y_t] = \mu < \infty$, $\mathbb{V}[y_t] = \sigma^2 < \infty$, and $\bar{\sigma}_T^2 = \mathbb{V}\left[\frac{1}{\sqrt{T}} \sum_{t=1}^T y_t\right] \xrightarrow{p} \bar{\sigma}^2 < \infty$, then

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_t \xrightarrow{d} N(\mu, \bar{\sigma}^2) \ .$$

Ergodicity is a rather philosophical concept and one cannot verify whether it holds in a given sample.[5] Counterexamples are rather exotic: e.g. something stochastic that is determined ahead of time and influences all observations in our particular sample (collection of realizations) so as to make it different from the underlying process (collection of RVs) that we want to learn about.[6]

Proposition 33 (Conditions for SS & Ergodicity).
If the process y_t consists of i.i.d. RVs, then y_t is SS and ergodic.
Moreover, suppose y_t is SS and ergodic and we define the process $z_t = h(..., y_{t-1}, y_t, y_{t+1}, ...)$ with a measurable function h. Then z_t is SS and ergodic as well.

Without ergodicity and strict stationarity, we can sometimes rely on the LLN and CLT for Marginale Difference Sequences (MDS).

Definition 42. *The process y_t is a martingale difference sequence (MDS)(w.r.t. the information sets $\{\mathcal{F}_t\}$) if $\mathbb{E}[y_t|\mathcal{F}_{t-1}] = 0 \; \forall \; t$.*

The information set $\mathcal{F}_t$ contains everything that is known to the researcher at time t. Usually, these are past values of the process(es) being analyzed; e.g. $\mathcal{F}_t = \{y_1, y_2, ..., y_t\}$.[7] We usually write $\mathbb{E}[y_t|\mathcal{F}_{t-1}]$ in shorthand as $\mathbb{E}_{t-1}[y_t]$.

[5] Even stationarity cannot be checked formally (without assuming a particular model that generated the data). However, the data can be inspected for obvious departures from stationarity such as persistently higher or more volatile realizations in some periods (indicative of instability of the mean and variance, respectively).

[6] For example, if $y_t = m + u_t$, $u_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and if we think of m as a RV in the process y_t, e.g. with $m \sim N(\mu, 1)$, while all our observations are characterized by a particular realization of m, then the process y_t is not ergodic. We are forever stuck with the particular m drawn ahead of time and we can never learn $\mu = \mathbb{E}[y_t]$ based on $\frac{1}{T}\sum_{t=1}^T y_t \xrightarrow{p} m$.

[7] The name MDS originates from the fact that an MDS y_t could be constructed as $y_t = x_t - x_{t-1}$ where x_t is a martingale, i.e. it holds that $\mathbb{E}[x_{t+s}|\mathcal{F}_t] = x_t$ for all s, t, whereby $\mathcal{F}_t = \{x_1, x_2, ..., x_t\}$.

Proposition 34 (LLN and CLT for MDS).
Let $\{y_t, \mathcal{F}_t\}$ be an MDS such that $\mathbb{E}[|y_t|^{2r}] < c < \infty$ for some $r > 1$ and all t. Then

1. $\frac{1}{T}\sum_{t=1}^{T} y_t \to 0$.
2. *If in addition* $\bar{\sigma}_T^2 = \mathbb{V}[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} y_t] \to \bar{\sigma}^2 > 0$, *then* $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} y_t \xrightarrow{d} N(0, \bar{\sigma}^2)$.

8.2 The General Linear Process and its Approximation

An important process is the White Noise (WN) process. It consists of uncorrelated – i.e. linearly independent – RVs and (usually) has zero mean.

Definition 43. *The process u_t is called white noise (WN), written as $u_t \sim WN(\mu, \sigma^2)$, if*

$$Cov(u_t, u_{t-h}) = \begin{cases} \sigma^2 & if\, h = 0 \\ 0 & otherwise \end{cases} .$$

We call u_t strong WN if $u_t \perp\!\!\!\perp u_{t-h}\ \forall\ t, h$ (i.e. u_t and u_{t-h} are independent).

Usually, we deal with mean-zero WN: $u_t \sim WN(0, \sigma^2)$. The distinction between (weak) WN and strong WN is only important for models where higher moments of u_t matter, e.g. conditional heteroskedasticity models (see Section 9.3.1). If the WN process u_t is Normal, then it is strong WN, as linear independence implies independence for Normal RVs. We can then simply write $u_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Note that the Normal WN process is SS and ergodic as it consists of i.i.d. RVs at each t (see Proposition 33 above).

Proposition 35 (Wold Decomposition).
Any mean-zero WS process y_t can be uniquely represented as

$$y_t = \sum_{l=0}^{\infty} b_l u_{t-l} \ , \quad where \quad u_t \sim WN(0, \sigma^2) \ , \quad b_0 = 1 \ \ , \ \sum_{l=0}^{\infty} b_l^2 < \infty \ .$$

The Wold decomposition tells us that any WS process with a mean of zero can be written as a linear process that consists of (infinitely many) WN components. The square summability condition $\sum_{l=0}^{\infty} b_l^2 < \infty$ ensures that the coefficients approach zero sufficiently quickly so that the impact of any u_{t-l} on y_t vanishes as $l \to \infty$. We refer to u_t as the innovation to the process y_t at time t. Note that the Wold decomposition does not restrict the distributional family of u_t, nor does it exclude higher-order dependencies among u_t. However, if u_t is i.i.d. (i.e. strict WN and i.d.), then by Proposition 33 y_t is SS and ergodic.[8] This happens, for example, if $u_t \sim N(0, \sigma^2)$.

The process above is referred to as the General Linear Process (GLP). We write it in shorthand as $y_t = B(L)u_t = \sum_{l=0}^{\infty} b_l u_{t-l}$, where $B(L) = b_0 + b_1 L + b_2 L^2 + ...$ is a lag-polynomial and L is the lag-operator. The lag-operator has the property that $Lu_t = u_{t-1}$ and, more generally, $L^k u_t = u_{t-k}$. Also, commutativity and distributivity hold: $L(cu_t) = c(Lu_t) = cu_{t-1}$ and $L(u_t + v_t) = u_{t-1} + v_{t-1}$.[9] For the GLP, it is easy to show that

$$\mathbb{E}[y_t] = 0 \; , \quad \text{and } \gamma_h = \sigma^2 \sum_{k=0}^{\infty} b_k b_{k+h} \; .^{10}$$

We cannot use the GLP directly for modeling our data because it contains infinitely many coefficients. Instead, we approximate it with more parsimonious models, usually from the Autoregressive-Moving Average (ARMA) class of models.

MA(1)

$$y_t = c + u_t + \theta u_{t-1} \; , \quad u_t \sim WN(0, \sigma^2) \; .$$

The MA(1) process cuts off the GLP after the first lag of u_t. We obtain $\mathbb{E}[y_t] = c$ and

$$\gamma_{h,t} = \mathbb{E}[(u_t + \theta u_{t-1})(u_{t-h} + \theta u_{t-h-1})] = \begin{cases} (1+\theta^2)\sigma^2 & \text{if } h = 0 \\ \theta\sigma^2 & \text{if } h = 1 \\ 0 & \text{if } h > 1 \end{cases} .$$

Therefore, the MA(1) process is WS regardless of the value of θ.

[8] And so are the processes $\{y_t^2\}$ or $\{y_t y_{t-h}\}$, all of which means that $\frac{1}{T}\sum_{t=1}^T y_t \to \mathbb{E}[y_t]$, $\frac{1}{T}\sum_{t=1}^T y_t^2 \to \mathbb{E}[y_t^2]$ and $\frac{1}{T}\sum_{t=1}^T y_t y_{t-h} \to \mathbb{E}[y_t y_{t-h}]$.

[9] For example, we have $(a + bL)Lx_t = ax_{t-1} + bx_{t-2}$ and $(1 - aL)(1 - bL)x_t = x_t - (a + b)x_{t-1} + abx_{t-2}$.

[10] $\gamma_h = \mathbb{E}[y_t y_{t-h}] = \mathbb{E}[(\sum_{l=0}^{\infty} b_l u_{t-l})(\sum_{k=0}^{\infty} b_k u_{t-h-k})] = \sum_{l=0}^{\infty}\sum_{k=0}^{\infty} b_l b_k \mathbb{E}[u_{t-l}u_{t-h-k}] = \sum_{k=0}^{\infty} b_{k+h} b_k \sigma^2$ because $\mathbb{E}[u_t u_s] = 0 \; \forall \; s \neq t$.

MA(q)

$$y_t = c + u_t + \theta_1 u_{t-1} + \ldots + \theta_q u_{t-q} \ , \quad u_t \sim WN(0, \sigma^2) \ .$$

In short, we write $y_t = c + \theta(L)u_t$ with $\theta(L) = 1 + \theta_1 L + ... + \theta_q L^q$. The MA(q) process cuts off the GLP after the qth lag of u_t. Note that the GLP is a MA(∞) and the WN process is a MA(0). We obtain again $\mathbb{E}[y_t] = c$ and

$$\begin{aligned}\gamma_{h,t} &= \mathbb{E}\left[(u_t + \theta_1 u_{t-1} + \ldots + \theta_q u_{t-q})(u_{t-h} + \theta_1 u_{t-h-1} + \ldots + \theta_q u_{t-h-q})\right] \\ &= \mathbb{E}\left[\theta_h u_{t-h}^2 + \theta_{h+1}\theta_1 u_{t-h-1}^2 + \ldots + \theta_q \theta_{q-1} u_{t-q}^2\right] \\ &= \begin{cases} (\theta_h\theta_0 + \theta_{h+1}\theta_1 + \ldots + \theta_1\theta_{q-h})\sigma^2 & \text{if } h = 0, 1, ..., q \ , \quad \text{where } \theta_0 = 1 \ , \\ 0 & \text{if } h > q \end{cases} .\end{aligned}$$

Again the MA(q) is WS regardless of the values of $\theta_1, ..., \theta_q$.

AR(1)

$$y_t = c + \phi y_{t-1} + u_t \ , \quad u_t \sim WN(0, \sigma^2) \ .$$

The AR(1) posits a linear relationship between y_t and its value from last period, y_{t-1}. Repeatedly inserting this expression on the RHS for the lags y_{t-l}, $l = 1, 2, ...$, we get $y_t = c(1 + \phi + \phi^2 + ...) + \sum_{l=0}^{\infty} \phi^l u_{t-l}$. Only under $|\phi| < 1$ is this well-defined, and we get

$$y_t = \frac{c}{1-\phi} + \sum_{l=0}^{\infty} \phi^l u_{t-l} \ ,$$

which in turn leads to $\mathbb{E}[y_t] = \frac{c}{1-\phi}$ and $\gamma_h = \frac{\phi^h}{1-\phi^2}\sigma^2$, and it shows that for $|\phi| < 1$ the process is WS.[11] This calculation illustrates that the AR(1) process gives us a particular, restricted MA(∞) – and hence an approximation of the GLP – with just one parameter. This puts some structure on our model as it induces a monotonically decaying ACF (in absolute value).

Under $|\phi| < 1$, the AR(1) process is stable, since the effect of u_{t-h} on y_t dies out as $h \to \infty$. If instead $|\phi| > 1$, then this effect diverges to infinity and we have an exploding process. If $|\phi| = 1$, provided that $c = 0$, y_t is simply the sum of past innovations u_t: $y_t = \sum_{h=0}^{\infty} u_{t-h}$. In that case, we call y_t a unit-root or random walk process (with drift if $c \neq 0$).[12] Under $|\phi| \geq 1$, the process is not WS as its variance diverges to infinity.[13]

[11] $\gamma_h = \mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)] = \mathbb{E}\left[\left(\sum_{l=0}^{\infty} \phi^l u_{t-l}\right)\left(\sum_{k=0}^{\infty} \phi^k u_{t-h-k}\right)\right] = \sum_{k=0}^{\infty} \phi^{h+k}\phi_k\sigma^2 = \sigma^2\phi^h \sum_{k=0}^{\infty} \phi^{2k}$, and $\sum_{k=0}^{\infty} \phi^{2k} = \frac{1}{1-\phi^2}$.

[12] From a purely mathematical point of view, the equation for y_t is a first-order linear difference equation, and $|\phi| < 1$ is the condition for its stability.

[13] We get $\mathbb{V}[y_t] = \phi^2\mathbb{V}[y_{t-1}] + \mathbb{V}[u_t]$, since $\text{Cov}(y_{t-1}, u_t) = 0$ because u_t is uncorrelated with all past u_ts, and y_t is a linear function of all past u_ts. If $|\phi| \geq 1$, then $\mathbb{V}[y_t] > \mathbb{V}[y_{t-1}]$ (if $\mathbb{V}[u_t] > 0$). If we think of the

With higher order AR-models, we can approximate the GLP and its ACF more flexibly, while still retaining the parsimony of having only a few parameters. For their analysis, it is useful to note that we can write the AR(1) process as $(1-\phi L)y_t = c + u_t$, and the above calculation makes clear that $(1-\phi L)^{-1} = \lim_{h\to\infty}(1+\phi L+\phi^2 L^2+...)$, provided that $|\phi| < 1$.

AR(2)

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + u_t \; , \quad u_t \sim WN(0, \sigma^2) \; .$$

This process can also be written as $\phi(L)y_t = c + u_t$ for $\phi(L) = (1 - \phi_1 L - \phi_2 L^2)$. We can factorize this lag-polynomial as $(1-\phi_1 L - \phi_2 L^2) = (1-\lambda_1 L)(1-\lambda_2 L)$, and by the above we know that we can invert it if $|\lambda_1| < 1$ and $|\lambda_2| < 1$.[14] It turns out that this is the condition for WS of the AR(2) process.

The process y_t can be written in so-called "companion form" as a particular, restricted vector-autoregression (VAR) of order 1, VAR(1):

$$\underbrace{\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix}}_{\tilde{y}_t} = \underbrace{\begin{bmatrix} c \\ 0 \end{bmatrix}}_{\tilde{c}} + \underbrace{\begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix}}_{\Phi_1} \underbrace{\begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix}}_{\tilde{y}_{t-1}} + \underbrace{\begin{bmatrix} u_t \\ 0 \end{bmatrix}}_{\tilde{u}_t} \; .$$

The first equation in this system gives the above equation for the AR(2) process of y_t, while the second just gives the identity $y_{t-1} = y_{t-1}$. Intuitively, the process y_t is stable if and only if the vector-valued process $[y_t, y_{t-1}]'$ is stable. And just like stability of an AR(1) depends on the absolute value of the autoregressive coefficient ϕ, stability of a VAR(1) depends on the absolute values of the eigenvalues of the autoregressive matrix Φ_1. For our particular, restricted VAR(1) here, it turns out that finding the eigenvalues of the matrix Φ_1, i.e. the values λ_1, λ_2 that solve $\lambda^2 - \phi_1\lambda - \phi_2 = 0$ (the roots of this polynomial), is is the same calculation as finding the values λ_1, λ_2 that give $(1 - \phi_1 L - \phi_2 L^2) = (1 - \lambda_1 L)(1 - \lambda_2 L)$.[15] The process for y_t is WS iff these eigenvalues are both less than one in absolute value.[16]

Once we know that y_t is WS, we can find the mean by using the fact that $\mathbb{E}[y_t] = \mu \; \forall \; t$. We

process as initialized in the infinite past, then this implies $\mathbb{V}[y_t] = \infty$, which means that these comparisons are ill-defined. For this reason, some textbooks add the requirement of finite variances to the definition of stationarity.

[14] i.e. λ_1 and λ_2 lie "inside the unit circle".

[15] To find λ_1 and λ_2 s.t. $(1 - \phi_1 z - \phi_2 z^2) = (1 - \lambda_1 z)(1 - \lambda_2 z)$ (we replace the lag operator L with some generic variable z), there are two options to proceed. One is to find the two roots z_1 and z_2 that give $(1 - \phi_1 z - \phi_2 z^2) = 0$, which implies that $\lambda_1 = 1/z_1$ and $\lambda_2 = 1/z_2$. The other is to divide this polynomial by z^2, to define $\lambda = 1/z$ and to find the two roots λ_1 and λ_2 that solve $z^{-2} - \phi_1 z^{-1} - \phi_2 = \lambda^2 - \phi_1\lambda - \phi_2 = 0$.

[16] From a mathematical point of view, y_t follows a second-order difference equation and this is the condition for its stability.

get

$$\mu = c + \phi_1\mu + \phi_2\mu + 0 = \frac{c}{1-\phi_1-\phi_2} \; .$$

To find the autocovariances, we use $\gamma_{h,t} = \gamma_h \; \forall \; t$ and $\gamma_h = \gamma_{-h} \; \forall \; h$. Multiplying the expression $y_t - \mu = \phi_1(y_{t-1}-\mu) + \phi_2(y_{t-2}-\mu) + u_t$ by $(y_{t-h}-\mu)$ and taking expectations yields

$$\begin{aligned}\gamma_h &= \phi_1\gamma_{h-1} + \phi_2\gamma_{h-2} + \text{Cov}(u_t, y_{t-h}) \\ &= \phi_1\gamma_{h-1} + \phi_2\gamma_{h-2} + \sigma^2 \, \mathbf{1}\{h=0\} \; .\end{aligned}$$ [17]

This is the Yule-Walker (second-order) difference equation for the ACF of an AR(2) process. Writing this equation out for $h = 0, 1, 2$ gives a system of equations that can be solved for γ_0, γ_1 and γ_2:

$$\begin{aligned}\gamma_0 &= \phi_1\gamma_1 + \phi_2\gamma_2 + \sigma^2 \; , \\ \gamma_1 &= \phi_1\gamma_0 + \phi_2\gamma_1 \; , \\ \gamma_2 &= \phi_1\gamma_1 + \phi_2\gamma_0 \; .\end{aligned}$$

We can then solve for $\{\gamma_h\}_{h=3,4,...}$ sequentially using the difference equation above. The additional flexibility of the AR(2) relative to the AR(1) process allows for a hump-shaped or sinusoidal ACF. Note that – as for the AR(1) – the ACF is decaying gradually towards zero.[18]

AR(p)

$$y_t = c + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + u_t \; , \quad u_t \sim WN(0, \sigma^2) \; .$$

We can write this as $\phi(L)y_t = c + u_t$ with $\phi(L) = (1 - \phi_1 L - ... - \phi_p L^p)$. Again, we can write y_t in companion form and it is WS iff the eigenvalues of the matrix

$$\Phi_1 = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & & \phi_p \\ 1 & 0 & \cdots & & 0 \\ 0 & 1 & & & \vdots \\ \vdots & & \ddots & & \\ 0 & & & 1 & 0 \end{bmatrix} = \begin{bmatrix} (\phi_1, ..., \phi_p)_{1\times p} \\ I_{p-1} \quad 0_{(p-1)\times 1} \end{bmatrix}$$

[17] We know $\text{Cov}(u_t, y_{t-h}) = \mathbb{V}[u_t] = \sigma^2$ for $h = 0$ and zero otherwise.

[18] This is easiest to see using the expression $y_t = (1-\lambda_1 L)^{-1}(1-\lambda_2 L)^{-1}u_t = (1+\lambda_1 L+\lambda_1^2 L^2+...)(1+\lambda_2 L+\lambda_2^2 L^2+...)u_t$, where we know λ_1 and λ_2 are both less than one in absolute value. As a result, lags of u_t from the more distant past have ever smaller coefficients in front of them. When computing $\gamma_h = \text{Cov}(y_t, y_{t-h})$, all the u_{t-l} for $l = 0, 1, ..., h-1$ in the expression for y_t are irrelevant because the u_ts are linearly independent, while u_{t-l} for $l = h, h+1, ...$ are multiplied by small coefficients.

are all less than one in absolute value. Also analogously to before, we get $\mathbb{E}[y_t] = (1 - \phi_1 - ... - \phi_p)^{-1}c$ and can compute the ACF using the Yule-Walker difference equation

$$\gamma_h = \phi_1 \gamma_{h-1} + ... + \phi_p \gamma_{h-p} + \mathbf{1}\{h = 0\}\sigma^2 \ .$$

ARMA(p,q)

$$y_t = c + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + u_t + \theta_1 u_{t-1} + ... + \theta_q u_{t-q} \ , \quad u_t \sim WN(0, \sigma^2) \ .$$

This can be written compactly as $\phi(L)y_t = c + \theta(L)u_t$ with $\phi(L) = (1 - \phi_1 L - ... - \phi_p L^p)$ and $\theta(L) = 1 + \theta_1 L + ... + \theta_q L^q$. The condition for WS is equivalent to the condition for WS under the corresponding AR(p) model with the same coefficients $\phi_1, ..., \phi_p$, i.e. the MA-coefficients $\theta_1, ..., \theta_q$ can be ignored (set to zero) when analyzing stationarity.

The MA(q) process essentially uses q coefficients to freely determine the first q autocovariances, while all remaining $\{\gamma_h\}_{h=q+1,q+2,...}$ are set to zero. The AR(p) process uses p coefficients to induce a particular, restricted shape for the ACF at all displacements $h = 0, 1, 2, ...$. The ARMA(p,q) process combines these two different modeling philosophies and can accommodate a wide range of ACF patterns. When it comes to WS processes, the only real drawback of ARMA models is that they do not capture higher-order serial correlations such as $\mathbb{E}[y_t^2 y_{t-h}^2]$ which are important for modeling financial data. Models that do that as well as models that depart from WS are briefly discussed in the context of state space models in Section 9.3.1. Many of these models build upon ARMA models.

8.3 Trend and Seasonality

Raw data is seldom stationary. It often displays a trend as well as some regular, fluctuating pattern, referred to as "seasonality". We can write the raw series as

$$y_t^{raw} = \text{trend}_t + \text{seasonality}_t + y_t \ .$$

Once the trend and seasonality are subtracted, we would be left with y_t, which is typically WS and can be modeled as an ARMA process. (Alternatively, if it displays some instability that (potentially) prevents it from being WS, we can model its dynamics using more advanced models that display, for example, conditional heteroskedasticity, stochastic volatility or time-varying parameters (see Section 9.3.1).)

Often, one can download data which was seasonally adjusted by the statistical agency that provides it. If not, one can consult techniques that deal with seasonalities (while leaving the trend intact). The simplest way to do so is to run a regression with a dummy for each season (e.g. a dummy for each month of the year), subtract the fitted values from the original series and add back the average coefficient across seasons to preserve the original scaling of the variable.

One can also often download data which is already de-trended as well. However, modeling a trend is more consequential than modeling seasonalities. There are two possible ways to deal with trends. The trend-stationary specification assumes a certain functional form of the trend and assumes that the deviation of the raw but seasonally adjusted series $y_t^{s.a.}$ from this trend is stationary. For example, under a linear trend,

$$y_t^{s.a.} = c + g \cdot t + y_t \ , \quad y_t \text{ is WS with mean zero} \ .$$

This leads to $y_{t+h}^{s.a.} = c + g(t+h) + y_{t+h}$, and hence $y_{t+h}^{s.a.} - y_t^{s.a.} = gh + y_{t+h} - y_t$. We call this a deterministic trend, because y_{t+h} and y_t are disturbances that push the series $y_t^{s.a.}$ only temporarily away from its trend. Regardless of its past values, the series $y_t^{s.a.}$ reverts back to its original trend and continues to grow at rate g per period. This is seen best if y_t is independent, which yields $\mathbb{E}_t[y_{t+h}^{s.a.}] = c + g(t+h)$.

The difference-stationary specification assumes that the first difference of $y_t^{s.a.}$ is stationary:

$$\Delta y_t^{s.a.} = g + y_t \ , \quad y_t \text{ is WS with mean zero} \ .$$

This implies $y_t^{s.a.} = y_{t-1}^{s.a.} + g + y_t$. In other words, $y_t^{s.a.}$ follows a random walk with drift and (potentially serially correlated) disturbances y_t. This leads to $y_{t+h}^{s.a.} = g\ h + \sum_{l=0}^{h} y_{t-l} + y_t^{s.a.}$. We call this a stochastic trend because the past value of the series $y_t^{s.a.}$ has a permanent effect on all future values of the series; the process does not revert back to trend, but continues to grow starting from $y_t^{s.a.}$. Even under independence of y_t, we have $\mathbb{E}_t[y_{t+h}^{s.a.}] = g\ h + y_t^{s.a.}$. As before, the per-period growth rate is g.

In practice, researchers often choose the specification under which the remainder y_t comes closer to being stationary or is easier to model. Under the difference-stationary model, one simply takes first differences and continues to work with $\Delta y_t^{s.a.}$. Under the trend-stationary model, one has to estimate the trend. Usually, this approach leads to a two-step procedure: one first estimates the trend and then continues to work with the remainder. In asymptotic analyses, the fact that the trend is estimated is usually ignored, in particular if one uses a series that has been de-trended by the statistical agency providing the data rather than by oneself. Estimating the trend and the supposed model for the remainder in one step requires special care with the asymptotic analysis (see Section 8.4.3).

Finally, note that the difference in trend-modeling philosophies applies even if y_t is not stationary, but is modeled based on some, pre-defined departure from stationarity, e.g. as a stochastic volatility process. If indeed $\Delta y_t^{s.a.}$ is WS, then we say that $y_t^{s.a.}$ is integrated of order one, I(1). For example, GDP is I(1) as GDP growth is stationary. A WS series is I(0). The only example of an I(2) series in economics is the price level in some countries and some periods, as the inflation series behaves like a random walk, i.e. an I(1) process.

8.4 Inference for Univariate Time Series Models

This section deals with the estimation of regressions using time series data and highlights the peculiarities that arise relative to the case with cross-sectional data. First, Section 8.4.1 deals with the estimation of AR(p) models, in particular the AR(1) model. Estimation of other univariate time series models, including MA and ARMA models, is treated in Section 9.3.4 in the context of state space model estimation. Then, Section 8.4.2 and Section 8.4.3 discuss how standard errors need to be adjusted when errors are autocorrelated and how the asymptotic analysis requires special care when deterministic trends are included in the regression.

Four things change when going from cross-sectional to time series econometrics. First, observations are correlated over time. As a result, one has to apply particular LLNs and CLTs for ergodic and SS processes (or MDS). This is illustrated in Section 8.4.1 in the context of the estimation of WS AR(p) processes. Second, for AR(p) processes that are not WS but feature a unit-root, the asymptotic distribution of frequentist estimators becomes non-standard. In addition, even though the asymptotic distribution of processes that have roots close to but different than unity stay intact, they become a poor approximation of an estimator's finite sample distribution. In contrast, Bayesian inference is unaffected by the degree of persistence in the data, which is a primary reason for its popularity in time series applications. All of this is discussed in Section 8.4.1. Third, when regressing one series on another, one has to adjust standard errors for the autocorrelation in the error terms, as discussed in Section 8.4.2. Fourth, as shown in Section 8.4.3, the asymptotic analysis of models with deterministic trends becomes somewhat more involved. These last two points are further reasons for the popularity of Bayesian inference for time series models, because it conditions on the data and hence is not concerned with asymptotics.[19]

[19] In other words, deriving the posterior and conducting finite sample inference does not rely on asymptotics.

8.4.1 Estimation of AR(p) Models

8.4.1.1 Inference Under Weak Stationarity

We can write the AR(1) model in linear regression form as

$$y_t = x_t'\phi + u_t \ , \quad \text{where} \quad x_t = [1, y_{t-1}]' \ , \quad \phi = [\phi_0, \phi_1]' \ , \quad \text{and} \quad u_t \sim WN(0, \sigma^2) \ .$$

Because u_t is WN and y_{t-1} is only a function of $\{u_{t-l}\}_{l=1,2,...}$, u_t and x_t are uncorrelated: $\mathbb{E}[x_t u_t] = 0 \ \forall \ t$. If u_t is strict WN, then u_t and x_t are independent: $\mathbb{E}[u_t|x_t] = \mathbb{E}[u_t] = 0$. The analogous holds for an AR(p), whereby $x_t = [1, y_{t-1}, ..., y_{t-p}]'$ and $\phi = [\phi_0, \phi_1, ..., \phi_p]'$.

We can derive the joint likelihood of data $\{y_t\}_{t=1:T}$ by factorizing it as the product of conditionals:

$$\begin{aligned} p(Y_{1:T}|\theta, y_0) &= p(y_T|\theta, y_{T-1}, y_{T-2}, ..., y_0) p(y_{T-1}, y_{T-2}, ..., y_1|\theta, y_0) \\ &= ... \\ &= \prod_{t=1}^{T} p(y_t|\theta, y_{t-1}, ..., y_1, y_0) \ . \end{aligned}$$

Under an AR(1), only the first lag is important for the distribution of y_t: $p(y_t|\theta, y_{t-1}, ..., y_1, y_0) = p(y_t|\theta, y_{t-1})$. In addition, assuming $u_t \sim N(0, \sigma^2)$, we get $y_t|y_{t-1} \sim N(\phi y_{t-1}, \sigma^2)$, and thus

$$\begin{aligned} p(Y_{1:T}|\theta, y_0) &= \prod_{t=1}^{T} (2\pi\sigma^2)^{-1/2} exp \ \left\{ -\frac{1}{2\sigma^2}(y_t - \phi y_{t-1})^2 \right\} \\ &= (2\pi\sigma^2)^{-n/2} exp \ \left\{ -\frac{1}{2\sigma^2} \sum_{t=1}^{T} (y_t - \phi y_{t-1})^2 \right\} \\ &= (2\pi\sigma^2)^{-n/2} exp \ \left\{ -\frac{1}{2\sigma^2} (Y - X\phi)'(Y - X\phi) \right\} \ . \end{aligned}$$

We call $p(Y_{1:T}|\theta, y_0)$ the conditional likelihood because it conditions on the initial observation y_0. The unconditional likelihood can be constructed by using the marginal distribution of y_0: $p(Y_{0:T}|\theta) = p(Y_{1:T}|\theta, y_0)p(y_0|\theta)$. For an AR(p) model, the conditional likelihood is $p(Y_{1:T}|\theta, y_0, y_{-1}, ..., y_{-p+1})$- While we can find $p(y_0|\theta)$ for an AR(1), finding the marginal distribution of $(y_0, y_{-1}, ..., y_{-p+1})$ is non-trivial, which is why in practice typically the conditional likelihood is used. Even for an AR(1), the unconditional likelihood prevents analytical manipulations. Note that we need data on $\{y_t\}_{t=-p+1:T}$ in order to form the (conditional) likelihood for $\{y_t\}_{t=1:T}$, i.e. our sample size T equals the number of time periods we have in our dataset minus p, the number of observations lost to the initial conditions. For example, estimating an AR(2) with a sample of 100 observations, we have $T = 98$.

As before in the cross-sectional case, we get the MLE and OLS estimator $\hat{\phi} = (X'X)^{-1}X'Y$, which equals $\phi + (X'X)^{-1}X'U$ if the model is specified correctly.[20] Unlike before, its asymptotic analysis is complicated by the fact that the observations $\{y_t, x_t\}_{t=1:T}$ are not i.i.d. across t. Instead of relying on the i.i.d. assumption to apply the WLLN and CLT (together with the Slutsky theorems), for the following results we must assume that y_t is SS and ergodic in order to apply the LLN and CLT for SS and ergodic processes. Note that assuming that the WN process u_t is Normal implies that y_t is SS and ergodic.

By the LLN for ergodic and SS processes,

$$\frac{1}{T}\sum_{t=1}^{T} x_t x_t' \xrightarrow{p} \mathbb{E}[x_t x_t'] \quad \text{and} \quad \frac{1}{T}\sum_{t=1}^{T} x_t u_t \xrightarrow{p} \mathbb{E}[x_t u_t] = 0 \ .$$

Together with Slutsky's theorem, these two results imply that $\hat{\phi}$ is consistent. By the CLT for ergodic and SS processes,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t u_t \xrightarrow{d} N\left(0, \mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t u_t\right]\right) \quad ,$$

where

$$\begin{aligned}
\mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t u_t\right] &= \frac{1}{T}\mathbb{V}\left[\sum_{t=1}^{T} x_t u_t\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{V}[x_t u_t] + \frac{1}{T}\sum_{t=1}^{T}\sum_{\tau\neq t}\text{Cov}(x_t u_t, x_\tau u_\tau) \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[u_t^2 x_t x_t'] + \frac{1}{T}\sum_{t=1}^{T}\sum_{\tau\neq t}\mathbb{E}[x_t u_t x_\tau u_\tau] \\
&= \sigma^2\mathbb{E}[x_t x_t'] \ ,
\end{aligned}$$

using the fact that u_t is independent across t, as it is a Normal WN process.[21] Overall, we get the analogous asymptotic distribution to the cross-sectional case:

$$\sqrt{T}(\hat{\phi} - \phi) \xrightarrow{d} N\left(0, \sigma^2\mathbb{E}[x_t x_t']^{-1}\right) \ .$$

[20] As before, it is unbiased if $\mathbb{E}[u_t|x_t] = 0$.

[21] By independence of u_t across t, we have $\mathbb{E}[u_t^2 x_t x_t'] = \mathbb{E}[u_t^2]\mathbb{E}[x_t x_t']$ and $\mathbb{E}[x_t u_t x_\tau u_\tau] = \mathbb{E}[x_t x_\tau]\mathbb{E}[u_t]\mathbb{E}[u_\tau] = 0$.

As in the cross-sectional case, we can use the analogy principle to estimate σ^2 and $\mathbb{E}[x_t x_t']^{-1}$ as $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^T \hat{u}_t^2$ and $\left[\frac{1}{T}\sum_{t=1}^T x_t x_t'\right]^{-1}$. However, unlike the cross-sectional case, x_t consists of lagged values of y_t, and we can also use the moments implied by our supposed model to estimate $\mathbb{E}[x_t x_t']^{-1}$. For example, under an AR(1), we have $x_t = [1, y_{t-1}]'$, and hence

$$\mathbb{E}[x_t x_t'] = \mathbb{E}\begin{bmatrix} 1 & y_{t-1} \\ y_{t-1} & y_{t-1}^2 \end{bmatrix} = \begin{bmatrix} 1 & \mathbb{E}[y_t] \\ \mathbb{E}[y_t] & \mathbb{E}[y_t^2] \end{bmatrix} ,$$

and we know that $\mathbb{E}[y_t] = \phi_0/(1-\phi_1)$ and $\mathbb{E}[y_t^2] = \mathbb{V}[y_t] + \mathbb{E}[y_t]^2 = \sigma^2/(1-\phi_1^2) + \phi_0^2/(1-\phi_1)^2$, which we can estimate by plugging in our estimates for σ^2, ϕ_0 and ϕ_1.[22]

Note that $x_t u_t$ is a MDS (w.r.t. the information set $\mathcal{F}_{t-1} = \{y_{t-1}, y_{t-2}, ...\}$).[23] This allows us to estimate the AR(p) model also with GMM, again adjusting the asymptotic analysis for dependence of observations over time. In sum, time series models can often be written in linear regression form and estimated using OLS or MLE, as shown above, or one can use the commonly encountered MDS property to estimate them using GMM.[24]

Once the model is written in linear regression form, Bayesian inference is conducted in the same way as before in Section 4.5. For example, under the prior $\phi \sim N(0, \lambda^{-1}I)$, we get the posterior $\phi|Y \sim N(\bar{\phi}, \bar{V})$ with $\bar{V} = [X'X + \lambda^{-1}I]^{-1}$ and $\bar{\phi} = \bar{V}X'Y$. As before, for regular priors,[25] the posterior of ϕ gets ever tighter concentrated around $\hat{\phi}$ so that, asymptotically, the posterior mean and mode equal the MLE, i.e. asymptotically, Bayesians and frequentists agree (see Section 4.5). This asymptotic agreement breaks down under unit-roots.

8.4.1.2 Inference Under Unit-Roots

Consider the AR(1) process under the presence of a unit-root:

$$y_t = \phi_0 + \phi_1 y_{t-1} + u_t , \quad \text{with} \quad \phi_1 = 1 .$$

For simplicity, assume $\phi_0 = 0$ and write $\phi_1 = \phi = 1$.[26] Bayesian inference remains unaffected: starting from some prior on the parameter ϕ, we use the likelihood to update our beliefs to the posterior distribution on ϕ. The domain of ϕ and, correspondingly, whether the process

[22] Note that we use stationarity to write $\mathbb{E}[y_{t-1}]$ as $\mathbb{E}[y_t]$ and $\mathbb{E}[y_{t-1}^2]$ as $\mathbb{E}[y_t^2]$.

[23] As a result, we get $\frac{1}{T}\sum_{t=1}^T x_t u_t \to 0$ and $\frac{1}{\sqrt{T}}\sum_{t=1}^T x_t u_t \xrightarrow{d} N\left(0, \mathbb{V}[\frac{1}{\sqrt{T}}\sum_{t=1}^T x_t u_t]\right)$ even without relying on ergodicity and SS. As long as $\frac{1}{T}\sum_{t=1}^T x_t x_t'$ is convergent, we have consistency.

[24] In fact, GMM came out in the 1980s as an estimation technique for rational expectations models in macroeconomics, as in them MDS arise quite naturally (just write an optimality condition like the Euler equation in the form $\mathbb{E}[f(x_t, u_t)|\mathcal{F}_{t-1}] = 0$).

[25] i.e. priors with positive probability mass on the whole domain of ϕ

[26] While $\phi = -1$ would also be a unit-root, this case is not encountered in economics.

y_t is WS is irrelevant for this analysis.

In contrast, frequentist inference requires substantial and non-trivial adjustments. The derivation of the OLS or MLE estimator $\hat{\phi} = (X'X)^{-1}X'Y$ is as before. However, for the asymptotic analysis we cannot rely neither on the LLN and CLT for ergodic and SS processes (as y_t is not stationary) nor on the LLN and CLT for MDS (as the variance of y_t is exploding). In fact, for the mean-zero AR(1), one can see that the asymptotic distribution derived under WS is not well-defined for $\phi = 1$: we get $\sqrt{T}(\hat{\phi} - \phi) \xrightarrow{d} N\left(0, \sigma^2/\mathbb{V}[y_{t-1}]\right) = N\left(0, 1 - \phi^2\right)$.

A formal treatment is beyond the scope of this exposition.[27] The bottom line is that $T(\hat{\phi} - 1)$ converges to a non-standard distribution called the Dickey-Fuller distribution. The convergence to this asymptotic distribution is faster than in the WS case; $\hat{\phi}$ converges at the rate T rather than $\sqrt{T}$, which is referred to as superconsistency. However, even for the simple case of an AR(1) it beocmes much more complicated to construct frequentist tests and confidence intervals, let alone for larger models, as commonly used in empirical analyses.

In addition, even for models with near unit-roots – i.e. ϕ close to, but below one –, the finite sample distribution of $\hat{\phi}$, as illustrated in simulations, is left-skewed and therefore features systematic deviations from its Normal asymptotic counterpart. These deviations are more pronounced the closer ϕ is to unity and the smaller the sample size. As a result, even for models without unit roots, standard frequentist inference can only be trusted if the degree of persistence (the value of ϕ in an AR(1) model) is sufficiently low and/or the sample size T is sufficiently large. It is fair to say that in many if not most time series models we know very little about the finite sample behavior of frequentist estimators. This leads many researchers to use Bayesian inference when dealing with time series data.

8.4.2 Estimating Regressions with Autocorrelated Errors

Suppose we are interested in estimating $y_t = x_t'\beta + u_t$ where x_t contains k variables measured over time (and potentially a constant). In such time series regressions, the error u_t is likely autocorrelated, i.e. $\mathbb{E}[u_t u_\tau] \neq 0$ for $t \neq \tau$. This leads to different standard errors for $\hat{\beta} = (X'X)^{-1}X'Y$ than under the i.i.d. (cross-sectional) case.

Assuming that $\mathbb{E}[x_t u_t] = 0$ and that both x_t and u_t are ergodic and SS processes, we can apply the LLN for such processes (together with Slutsky's theorem) to show consistency of

[27] For some intuition, note that for the mean-zero AR(1), we have $y_t = \sum_{\tau=1}^{t} u_\tau$. If u_t is i.i.d. across t, we have $\frac{1}{\sqrt{t}} y_t = \frac{1}{\sqrt{t}} \sum_{\tau=1}^{t} u_\tau \xrightarrow{d} N(0, \sigma^2)$ by CLT. As a result, the sample average $\frac{1}{T}\sum_{t=1}^{T} y_t = \frac{1}{T}\sum_{t=1}^{T}\sum_{\tau=1}^{t} u_\tau = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\sqrt{\frac{t}{T}}\frac{1}{\sqrt{t}}\sum_{\tau=1}^{t} u_\tau\right]$ converges (in distribution) to a RV rather than to some point, like the mean for a WS AR(1).

$\hat{\beta}$ in the same way as done for $\hat{\phi}$ under the WS AR(1) above. Also analogously to that case, we can use the LLN and CLT for ergodic and SS processes (together with Slutsky's theorem) to get the following asymptotic distribution:

$$\sqrt{T}(\hat{\beta}-\beta) \xrightarrow{d} N\left(0\ ,\ \mathbb{E}[x_t x_t']^{-1}\mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t u_t\right]\mathbb{E}[x_t x_t']^{-1}\right) .$$

Different assumptions on the properties of the process $u_t|\{x_t\}_{t=1:T}$ lead to different expressions for this asymptotic variance. The term in the middle can be written as

$$\begin{aligned}\mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t u_t\right] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{V}[x_t u_t] + \frac{1}{T}\sum_{t=1}^{T}\sum_{\tau\neq t}\text{Cov}(x_t u_t, x_\tau u_\tau) \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[x_t x_t' u_t^2] + \frac{1}{T}\sum_{t=1}^{T}\sum_{\tau\neq t}\mathbb{E}[x_t x_\tau' u_t u_\tau] \\ &= \mathbb{E}[x_t x_t' u_t^2] + \sum_{h=1}^{T-1}\mathbb{E}[u_t u_{t-h}(x_t x_{t-h}' + x_{t-h}x_t')] .\end{aligned}$$

If $x_t u_t$ is not autocorrelated, the second term is zero. If in addition $\mathbb{V}[u_t|x_t] = \gamma_0$ is not a function of $\{x_t\}$, the first term simplifies to $\gamma_0\mathbb{E}[x_t x_t']$ by LIE. For the asymptotic variance of $\hat{\beta}$, we end up with $\gamma_0\mathbb{E}[x_t x_t']^{-1}$, the same expression as obtained under a homoskedastic linear regression model for cross-sectional data. If $\mathbb{V}[u_t|x_t]$ is a function of x_t, the first term does not simplify, and we construct the heteroskedasticity-robust variance $\mathbb{E}[x_t x_t']^{-1}\mathbb{E}[x_t x_t' u_t^2]\mathbb{E}[x_t x_t']^{-1}$. We estimate these objects by replacing the expectation operators with time-averages and u_t with $\hat{u}_t$.

If $x_t u_t$ is autocorrelated, we can construct the heteroskedasticity and autocorrelation (HAC) robust variance by replacing $\mathbb{E}[u_t u_{t-h}(x_t x_{t-h}' + x_{t-h}x_t')]$ by $\frac{1}{T-h}\sum_{t=h+1}^{T}\hat{u}_t\hat{u}_{t-h}(x_t x_{t-h}' + x_{t-h}x_t')$. Because for high h, we have little observations based on which we can estimate this term, it is custom to multiply this expression by a Kernel (i.e. a weighting-term) w_h. For example, the Newey-West estimator sets $w_h = 1 - \frac{h}{H+1}$ for $h = 1 : H$ and $w_h = 0$ for $h > H$ (Newey and West, 1987).

8.4.3 Estimating Regressions with Deterministic Trends

Consider the deterministic trend model $y_t = \beta_0 + \beta_1 t + u_t = x_t'\beta + u_t$, where u_t is some mean-zero WS process and $x_t = [1, t]'$. The asymptotic analysis of such regressions with deterministic trends requires special care, as $\frac{1}{T}\sum_{t=1}^T x_t x_t'$ does not converge to a non-singular matrix. Hence, one cannot derive the asymptotic distribution of $\sqrt{T}(\hat{\beta} - \beta) = \left(\frac{1}{T}\sum_{t=1}^T x_t x_t'\right)^{-1} \frac{1}{\sqrt{T}}\sum_{t=1}^T x_t u_t$.

Instead, one can derive the asymptotic distribution of

$$\sqrt{T} G_T(\hat{\beta} - \beta) \; , \quad \text{where} \quad G_T = \begin{bmatrix} 1 & 0 \\ 0 & T \end{bmatrix} \; .$$

If u_t is WN, we get $\sqrt{T} G_T(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 \mathbb{E}[x_t x_t']^{-1})$. This adjustment reflects the fact that $\hat{\beta}_0$ converges at rate $T^{1/2}$, while $\hat{\beta}_1$ converges at rate $T^{3/2}$. Intuitively, this is because the variance of the regressor $x_{2,t} = t$ explodes as $T \to \infty$. We say $\hat{\beta}_1$ is superconsistent. Based on this result, we can approximate the finite sample distribution of $\hat{\beta}$ as $N(\beta, \frac{\sigma^2}{T} G_T^{-1} \mathbb{E}[x_t x_t']^{-1} G_T^{-1})$.

Generally, in models with different convergence speeds of estimators, one has to find some matrix G_T so that $\frac{1}{T}\sum_{t=1}^T G_T^{-1} x_t x_t' G_T^{-1\prime}$ converges to a non-singular matrix. In other words, for purposes of asymptotic analysis, one considers the rotated regressors $G_T^{-1} x_t$. In the above example, this is quite straightforward. In more general models, additional complications can arise. For example, in the model $y_t = \beta_0 + \beta_1 t + \beta_2 y_{t-1} + u_t$, we cannot just multiply both regressors t and y_{t-1} by T^{-1} because asymptotically, the two transformed regressors would be perfectly collinear.

Appendix