Stationary Time Series Notes

Technicalities

Concept	**Series** $\{Y_t\}$	**Random Variable** Y_t
Stationarity	Stationarity is a property of series.	Stationarity is not a property of random variables. If Y_t, Y_s belong to a stationary series $\{Y_t\}$, then the two are in some sense similarly distributed.
Autoregressive Model $AR(p)$	A series is said to follow an autoregressive model $AR(p)$ iff the random variables of that series are distributed according to $AR(p)$. Then, it is also said that the series is a $AR(p)$ process, and that it is generated by an $AR(p)$ model.	An autoregressive model $AR(p)$ describes the distribution of random variables.
Granger Causality	Granger causality is a relation that holds between two (or more) series.	
Unit Root	If the random variables that a series consists in have a unit root (and are distributed according to an $AR(p)$ model) then the series is a unit root autoregressive process, and follows a unit root autoregressive model.	Having a unit root is a property of random variables (that are distributed according to an $AR(p)$ model).
Order of Integration	Order of integration is a property of series.	Number of roots is a property of random variables (and is numerically equivalent to the order of integration of the series that such random variables constitute). $Y_t \sim I(d)$ denotes that Y_t has number of roots d.
Cointegration	Cointegration is not a property of series.	Cointegration is a property of random variables.

Time Series Data

Characterisation of Time Series Data

- Time series data arise from repeated observations of one entity. Cross-sectional data arise from observations of the same variable across different entities at a single point in time.

Comparison of Time Series Data and Cross-Sectional Data

- Time series data is **temporally ordered**, so the ordering of observations is meaningful. For example, it makes sense to think of Y_{t+1} as dependent on Y_t but not the reverse, and plots of Y_t against t are often highly informative. In contrast, the ordering (index) of cross-sectional data is not meaningful.
- Time series data exhibits **serial dependence**, i.e. Y_t is generally correlated with Y_{t-1} and Y_{t-h} potentially for large h. In contrast, that cross-sectional data is generated by random sampling implies that for all $i \neq j$, Y_i and Y_j are independent.
- Observations in time series data potentially have **non-identical and time-dependent distributions.** For example, the distribution of Y_t potentially varies with t across epochs.

Mathematical Model for Time Series Data

- A time series is some series $\{Y_1, Y_2, ..., Y_T\}$ of random variables. This is also written as $\{Y_t\}_{t=1}^{T}$ or $\{Y_t\}$, which leaves the range of the series implicit.
    - Both time series data and cross-sectional data are modelled mathematically as random variables. The randomness of cross-sectional data is a product of random sampling. In contrast, the randomness of time series data is a product of "real" randomness in the observed entity. It is supposed that the observed entity has a range of counterfactual histories, and the history observed is randomly realised.

Stationarity

- Description of a time series by some mathematical model requires that the time series is in some stable. The notion of stability is captured by the technical concept of stationarity.
- Time series $\{Y_t\}$ is weakly stationary iff the mean (of each observation Y_t) $\mathbb{E}Y_t$, variance $var(Y_t)$, and autocovariances $cov(Y_t, Y_{t-h})$ for all $h \in \mathbb{Z}$ are time-invariant, i.e. independent of t.
    - If a time series is weakly stationary, we apply the following notation: $\mu := \mathbb{E}Y_t$, $\sigma^2 := var(Y_t)$, $\gamma_h := cov(Y_t, Y_{t-h})$.
    - These describe the random distribution that each Y_t is drawn from, and are distinct from the sample mean, variance, and autocovariance. Roughly speaking, these are population parameters.
    - Two time series $\{Y_t\}, \{X_t\}$ are jointly weakly stationary iff their means, variances, autocovariances, and cross covariances $cov(Y_t, X_{t-h})$ and $cov(X_t, Y_{t-h})$ are time-invariant.
- Time series $\{Y_t\}$ is strictly stationary iff for all $k \geq 0$ and $s, t \geq 1$, $(Y_t, Y_{t+1}, \ldots, Y_{t+k})$ has the same distribution as $(Y_s, Y_{s+1}, \ldots, Y_{s+k})$.
    - It can be verified that strict stationarity implies weak stationarity.
    - Two time series $\{Y_t\}, \{X_t\}$ are jointly strictly stationary iff for all $k \geq 0$ and $s, t \geq 1$, $(Y_t, X_t, Y_{t+1}, Y_{t+1}, \ldots, Y_{t+k}, X_{t+k})$ has the same distribution as $(Y_s, X_s, Y_{s+1}, Y_{s+1}, \ldots, Y_{s+k}, X_{s+k})$.
- In what follows, "stationary" means weakly stationary.

Descriptive Statistics

- Stationarity implies that sample estimators of means, variances, and autocovariances are consistent for their population counterparts.
- The sample mean of stationary time series $\{Y_t\}$ is $\overline{Y}_T := \frac{1}{T}\sum_{t=1}^T Y_t$. This converges in probability to the population mean $\mu := \mathbb{E}Y_t$ of each Y_t (which, given stationarity, is time-invariant).
    - If $\{Y_t\}$ is not stationary, then its sample mean $\overline{Y}_T := \frac{1}{T}\sum_{t=1}^T Y_t$ converges to the average population mean $\frac{1}{T}\sum_{t=1}^T \mathbb{E}Y_t$ in the range $\{1, \ldots, T\}$. Note that in this case, $\mathbb{E}Y_t$ is not time-invariant, hence in general, for $t, t' \in \{1, \ldots, T\}$, $\mathbb{E}Y_t \neq \mathbb{E}Y_{t'}$.
- The h^{th} sample autocovariance of stationary time series $\{Y_t\}$ is $\hat{\gamma}_h := c\hat{o}v(Y_t, Y_{t-h}) := \frac{1}{T}\sum_{t=h+1}^T (Y_t - \overline{Y}_T)(Y_{t-h} - \overline{Y}_T)$. This converges in probability to the h^{th} population autocovariance $\gamma_h := cov(Y_t, Y_{t-h})$ (which, given stationarity, is time-invariant).
    - Note that the sample autocovariance is computed from $T-h$ observations of (Y_t, Y_{t-h}), but that the denominator in the computation is T rather than $T-h$ (or $T-h-1$ which is entailed by an analogous application of Bessel's correction). This is motivated by the aim of reducing the likelihood of obtaining erroneously large estimates at long lags.
- The sample variance of stationary time series is $\{Y_t\}$ is $\hat{\gamma}_0 = \frac{1}{T}\sum_{t=1}^T (Y_t - \overline{Y}_T)^2 =: v\hat{a}r(Y_t)$. This converges in probability to the population variance $var(Y_t)$.
- The h^{th} sample autocorrelation of stationary time series $\{Y_t\}$ is $\hat{\rho}_h := \frac{c\hat{o}v(Y_t, Y_{t-h})}{v\hat{a}r(Y_t)} = \frac{\hat{\gamma}_h}{\hat{\gamma}_0}$. This converges in probability to the h^{th} population autocorrelation $\rho_h := \frac{cov(Y_t, Y_{t-h})}{var(Y_t)} = \frac{\gamma_h}{\gamma_0}$.
    - The autocorrelation function of time series $\{Y_t\}$ is the function from lag h to autocorrelation ρ_h. In practice, the autocorrelation ρ_h is generally not known, so the autocorrelation function is estimated by the sample autocorrelation function, which is the function from lag h to sample autocorrelation $\hat{\rho}_h$.
- **Note that the population counterparts $\mathbb{E}Y_t, var(Y_t), cov(Y_t, Y_{t-h})$ are properties of the distribution of the random variable Y_t (and Y_{t-h}), which are time-invariant if $\{Y_t\}$ is stationary. These quantities do not (as such) describe the distribution of observations in $\{Y_t\}$, but describe the random distribution Y_t due to the "real" randomness in the entity observed, that has a range of counterfactual histories.**
- The persistence of a time series $\{Y_t\}$, heuristically, is the speed with which $\{Y_t\}$ reverts to its mean, or equivalently, the extent of serial autocorrelation in $\{Y_t\}$.
    - More persistent time series have smoother trajectories and take lengthier excursions from their mean.
    - For a more persistent time series, autocorrelation decays to zero more slowly with increasing lags.

Autoregressive Models

- An autoregressive model is a model for generating a series of observations. It is the very rough analogue of a causal model in the cross-sectional setting.
- An autoregressive model $AR(p)$ of order p is some set of equations that describes the relationships between a time series variable Y_t and its lags $Y_{t-1}, \ldots, Y_{t-p}$, namely the equation $Y_t = \beta_0 + \sum_{i=1}^p \beta_i Y_{t-i} + u_t$, where $\mathbb{E}[u_t|\mathcal{Y}_{t-1}] = 0$, i.e. u_t is

mean-zero and is not forecastable on the basis of past realisations of Y, Y_0 is given, and u_t is some stationary shock.

- $\mathcal{Y}_{t-1} := Y_{t-1}, Y_{t-2}, \ldots$.
- If each variable Y_t in some time series $\{Y_t\}$ satisfies this equation, then $\{Y_t\}$ is said to be an $AR(p)$ process, and equivalently, it is said that $\{Y_t\}$ is generated by an $AR(p)$ model. It is also said that Y_t is $AR(p)$.
- $AR(p)$ is a property of each random variable $Y_t \in \{Y_t\}$. Being an $AR(p)$ process, or being generated by an $AR(p)$ model is a property of time series $\{Y_t\}$.

- Note that, under an $AR(p)$ model (i.e. given some $\{Y_t\}$ that satisfies the $AR(p)$ model), it is not necessarily the case that Y_t is uncorrelated with Y_{t-p-1}, it is only implied that the correlation of Y_t and Y_{t-p-1} is entirely a product of Y_t's dependence on $Y_{t-1}, \ldots, Y_{t-p}$ which are in turn dependent n Y_{t-p-1}.

Stationarity

- An $AR(1)$ process (i.e. a time series $\{Y_t\}$ that satisfies $Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$ where $\mathbb{E}[u_t|\mathcal{Y}_{t-1}] = 0$) is stationary iff for all t, $\mathbb{E}Y_t = \frac{\beta_0}{1-\beta_1}$ and $var(Y_t) = \frac{\sigma_u^2}{1-\beta_1^2}$ (where σ_u^2 denotes the common variance of each u_t).
  - Suppose that $\{Y_t\}$ is a stationary $AR(1)$ process.
  - $$\begin{aligned}\mathbb{E}Y_t &= \beta_0 + \beta_1 \mathbb{E}Y_{t-1} + \mathbb{E}u_t \\ &= \beta_0 + \beta_1 \mathbb{E}Y_t + \mathbb{E}[\mathbb{E}[u_t|\mathcal{Y}_{t-1}]], \\ &= \beta_0 + \beta_1 \mathbb{E}Y_t\end{aligned}$$
  - $\Leftrightarrow \mathbb{E}Y_t = \frac{\beta_0}{1-\beta_1}$.
  - $$\begin{aligned}var(Y_t) &= \beta_1^2 var(Y_{t-1}) + var(u_t) + 2\beta_1 cov(Y_{t-1}, u_t) \\ &= \beta_1^2 var(Y_{t-1}) + var(u_t) \\ &= \beta_1^2 var(Y_t) + \sigma_u^2\end{aligned},$$
  - $\Leftrightarrow var(Y_t) = \frac{\sigma_u^2}{1-\beta_1^2}$.
- Then, given that $var(Y_t)$ is non-negative, if $\{Y_t\}$ is a stationary $AR(1)$ process, then $\beta_1 \in (-1, 1)$. In other words $\beta_1 \in (-1, 1)$ is a necessary (and almost sufficient) condition for stationarity if $\{Y_t\}$ is an $AR(1)$ process.
- $\beta_1 \in (-1, 1)$, and $\mathbb{E}Y_0 = \frac{\beta_0}{1-\beta_1}$ and $var(Y_0) = \frac{\sigma_u^2}{1-\beta_1^2}$ are collectively sufficient for stationarity. The former condition is the substantive condition, whereas the latter two are understood as a mathematical technicality.
  - For such a process,
  - $$\begin{aligned}cov(Y_t, Y_{t-h}) &= cov(\beta_0 + \beta_1 Y_{t-1} + u_t, Y_{t-h}) \\ &= \beta_1 cov(Y_{t-1}, Y_{t-h}) \\ &\vdots \\ &= \beta_1^h cov(Y_{t-h}, Y_{t-h}) \\ &= \beta_1^h \sigma_Y^2\end{aligned},$$
  - $\Rightarrow \rho_h = \beta_1^h$.
- Generally, the substantive condition for the stationarity of an $AR(p)$ process (i.e. a time series that satisfies $Y_t = \beta_0 + \sum_{i=1}^p \beta_i Y_{t-i} + u_t$, where $\mathbb{E}[u_t|\mathcal{Y}_{t-1}] = 0$) is that $(\sum_{i=1}^p \beta_i) < 1$.

Forecasting

- The optimal 1-step ahead forecast of time series $\{Y_t\}$ is $Y_{T+1|T} := \mathbb{E}[Y_{T+1}|\mathcal{Y}_T]$.
- The optimal h-step ahead forecast of time series $\{Y_t\}$ is $Y_{T+h|T} := \mathbb{E}[Y_{T+h}|\mathcal{Y}_T]$.
  - It can be verified that this minimises the mean-squared forecast error, i.e. $Y_{T+h|T} := \mathbb{E}[Y_{T+h}|\mathcal{Y}_T] = \arg\min_m \mathbb{E}[Y_{T-h} - m(\mathcal{Y}_T)]^2$.
- The optimal 1-step ahead forecast of an $AR(p)$ process $\{Y_t\}$ is
  - $$\begin{aligned}Y_{T+1|T} &:= \mathbb{E}[Y_{T+1}|\mathcal{Y}_T] \\ &= \mathbb{E}[\beta_0 + \sum_{i=1}^p \beta_i Y_{T+1-i} + u_T|\mathcal{Y}_T] \\ &= \beta_0 + \sum_{i=1}^p \beta_i Y_{T+1-i}\end{aligned}.$$
- The optimal 2-step ahead forecast of $AR(1)$ process $\{Y_t\}$ is
  - $$\begin{aligned}Y_{T+2|T} &:= \mathbb{E}[Y_{T+2}|\mathcal{Y}_T] \\ &= \mathbb{E}[\beta_0 + \beta_1 Y_{T+1} + u_{T+2}|\mathcal{Y}_T] \\ &= \beta_0 + \beta_1 \mathbb{E}[Y_{T+1}|\mathcal{Y}_T] \\ &= \beta_0 + \beta_1 Y_{T+1|T}\end{aligned}.$$
- The optimal h-step ahead forecast of $AR(1)$ process $\{Y_t\}$ is
  - $Y_{T+h|T} = \beta_0 + \beta_1 Y_{T+h-1|T}$, by generalisation.
  - Let $\mu = \frac{\beta_0}{1-\beta_1}$, then $\mu = \beta_0 + \beta_1 \mu$,

        - $Y_{T+h|T} - \mu = \beta_0 + \beta_1 Y_{T+h-1|T} - \beta_0 - \beta_1\mu$,
        - $Y_{T+h|T} - \mu = \beta_1(Y_{T+h-1|T} - \mu)$.
    - By recursive substitution, $Y_{T+h|T} - \mu = \beta_1^h(Y_T - \mu)$.
    - Supposing that $\{Y_t\}$ is stationary, $\beta_1 \in (-1, 1)$, then as h approaches ∞, $Y_{T+h|T}$ converges to μ.
- The result that a stationary $AR(1)$ process is mean-reverting (in the sense that the optimal h-step ahead forecast converges to $\mathbb{E}\mu$ as h becomes large) generalises to $AR(p)$ processes for $p > 1$.

Estimation

- If it is supposed that time series $\{Y_t\}$ is an $AR(p)$ process, i.e. that it satisfies$Y_t = \beta_0 + \sum_{i=1}^p \beta_i Y_{t-i} + u_t$, where $\mathbb{E}[u_t|\mathcal{Y}_{t-1}] = 0$, then $\mathbb{E}u_t = cov(u_t, Y_{t-1} = \ldots = cov(u_t, Y_{t-p}) = 0$, i.e. the model $Y_t = \beta_0 + \sum_{i=1}^p \beta_i Y_{t-i} + u_t$ satisfies orthogonality, hence its parameters can be consistently estimated by OLS regression of Y_t on $Y_{t-1}, \ldots, Y_{t-p}$.
    - Supposing that $\{Y_t\}$ is stationary, the OLS estimators for $\beta_0, \ldots, \beta_p$ are asymptotically normal, and the familiar methods for performing hypothesis tests and constructing confidence intervals apply.
- Note here a minor complication in the OLS estimation of an $AR(p)$ model that supposedly generates some time series $\{Y_t\}$.
    - The analogue from the cross-sectional case of the OLS regression problem is
    - $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \arg\min_{b_0, b_1, \ldots, b_p} \sum_{t=p+1}^T (Y_t - b_0 - b_1 Y_{t-1} - \ldots - b_p Y_{t-p})^2$.
    - This is based on the understanding that given T observations of Y_t, there are only $T - p$ observations of $(Y_t, Y_{t-1}, \ldots, Y_{t-p})$ on which the OLS model can be fit.
    - We make the simplifying assumption that there are always sufficient observations $Y_0, Y_{-1}, \ldots$ such that there are T observations on which the OLS model can be fit. Then, the OLS regression problem is instead
    - $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \arg\min_{b_0, b_1, \ldots, b_p} \sum_{t=1}^T (Y_t - b_0 - b_1 Y_{t-1} - \ldots - b_p Y_{t-p})^2$.
    - Presumably, it remains the case that $\hat{\beta}_0 = \overline{Y}_T - \hat{\beta}_1 \overline{Y}_{T-1} - \ldots - \hat{\beta}_p \overline{Y}_{T-p}$ and $\hat{\beta}_i = \frac{c\hat{o}v(Y_T, Y_{T-i})}{v\hat{a}r(Y_{T-i})}$.

Estimation of the Optimal Forecast

- Because in general, even if it is supposed that $\{Y_t\}$ is an $AR(p)$ process, the parameters of the associated model are not known, the optimal forecast is not known, and must be estimated.
- The estimated optimal 1-step ahead forecast of an $AR(p)$ process $\{Y_t\}$ is $\hat{Y}_{T+1|T} := \hat{\beta}_0 + \sum_{i=1}^p \hat{\beta}_i Y_{t+1-i}$, where $\hat{\beta}_0, \ldots, \hat{\beta}_p$ are the OLS estimators for $\beta_0, \ldots, \beta_p$.
- The error of this estimated optimal 1-step ahead forecast is $\hat{e}_{T+1|T} := Y_{T+1} - \hat{Y}_{T+1|T}$, and can be decomposed as follows.
    - Let $e_{T+1|T} = Y_{T+1} - Y_{T+1|T}$.
    - $$\begin{aligned} \hat{e}_{T+1|T} &:= Y_{T+1} - \hat{Y}_{T+1|T} \\ &= (Y_{T+1} - Y_{T+1|T}) + (Y_{T+1|T} - \hat{Y}_{T+1|T}) \\ &= e_{T+1|T} + (Y_{T+1|T} - \hat{Y}_{T+1|T}) \\ &= [\beta_0 + \sum_{i=1}^p \beta_i Y_{T+1-i} + u_{T+1}] - [\beta_0 + \sum_{i=1}^p \beta_i Y_{T+1-i}] + (Y_{T+1|T} - \hat{Y}_{T+1|T}) \\ &= u_{T+1} + (Y_{T+1|T} - \hat{Y}_{T+1|T}) \\ &= u_{T+1} + [\beta_0 + \sum_{i=1}^p \beta_i Y_{T+1-i}] - [\hat{\beta}_0 + \sum_{i=1}^p \hat{\beta}_i Y_{T+1-i}] \\ &= u_{T+1} + [(\beta_0 - \hat{\beta}_0) + \sum_{i=1}^p (\beta_i - \hat{\beta}_i) Y_{T+1-i}] \end{aligned}.$$
    - The error of the estimated optimal 1-step ahead forecast is decomposed into:
        - the error of the optimal forecast $e_{T+1|T}$ which is equal to the unforecastable shock component $u_{T+1}$of Y_{T+1}, and
        - the error introduced by OLS estimation of the model parameters.
    - Note that the two components are uncorrelated because the latter component is a function of past observations of the time series and OLS estimators which are themselves computed as a function of past observations of the time series, and the unforecastable shock component is entirely uncorrelated with past observations of the time series.
- Then, the mean-squared forecast error $MSFE(\hat{Y}_{T+1|T}) := \mathbb{E}\hat{e}_{T+1|T}^2$ can be decomposed as follows.
    - $$\begin{aligned} MSFE(\hat{Y}_{T+1|T}) &:= \mathbb{E}\hat{e}_{T+1|T}^2 \\ &= \mathbb{E}u_{T+1}^2 + \mathbb{E}(\hat{Y}_{T+1|T} - Y_{T+1|T})^2 \\ &= var(u) + \mathbb{E}(\hat{Y}_{T+1|T} - Y_{T+1|T})^2 \end{aligned}.$$

- So the variance of the unforecastable shock u_T (and its sample counterpart, the residual $\hat{u}_t = Y_t - \hat{Y}_t$) is a poor estimate of the mean-squared forecast error that systematically underestimates the mean-squared forecast error. $\hat{u}_t$ is unsuitable for evaluation of forecast performance.

Forecast Evaluation

- Suppose that the performance of a forecasting procedure is evaluated by computing its mean-squared forecast error $MSFE(\hat{Y}_{T+1|T}) := \mathbb{E}\hat{e}^2_{T+1|T} := \mathbb{E}(Y_{T+1} - \hat{Y}_{T+1|T})^2$. Because Y_{T+1} is not known, the MSFE cannot be directly computed and must be estimated.
- The general difficulty with such estimation is that the residual $\hat{e}_{T+1|T} := Y_{T+1} - \hat{Y}_{T+1|T}$ is an out-of-sample residual and not an in-sample residual, unlike cross-sectional residuals, which are in-sample. Then, the variance of the in-sample residual $\hat{u}_t = Y_t - \hat{Y}_t$, by the above decomposition, systematically underestimates the MSFE.
- The procedure for estimating MSFE is as follows.
  - Suppose that the estimated optimal forecast is obtained by fitting an $AR(p)$ model.
  - Select some number of periods P. By convention, $P \in [0.1T, 0.2T]$.
  - For each $s \in \{T-P, \dots, T-1\}$, compute the pseudo out-of-sample forecast error $\hat{e}_{s+1|s} := Y_{s+1} - \hat{Y}_{s+1|s}$, where $\hat{Y}_{s+1|s}$ is the estimated optimal forecast of Y_{s+1} obtained by fitting an $AR(p)$ model on $\{Y_t\}_{t=1}^{s}$ (i.e. by estimating the parameters of the $AR(p)$ model by OLS regression of Y_t on $Y_{t-1}, \dots, Y_{t-p}$ for $t \in \{1, \dots, s\}$ and computing $\hat{Y}_{s+1|s}$).
  - Then, the estimator for $MSFE(\hat{Y}_{T+1|T})$ is $M\hat{S}FE(\hat{Y}_{T+1|T}) := \frac{1}{P}\sum_{s=T-P}^{T-1} \hat{e}^2_{s+1|s}$, i.e. the average of the squared pseudo out-of-sample forecast errors.

Model Selection

- In the above, the order p of the $AR(p)$ model that generated time series data $\{Y_t\}$ was treated as a known quantity. In practice, this quantity is unknown, so in estimating an optimal forecast, researchers must choose the order of the model to be estimated.
- The choice of the model to be estimated involves the bias-variance trade-off.
  - A higher order $AR(p)$ model is more flexible, and can accommodate richer dynamics, so is potentially better able to capture the true dynamics of the process that generated $\{Y_t\}$. This reduces the bias of the estimated optimal forecast.
  - A lower order $AR(p)$ model estimates fewer parameters with a finite amount of data. This reduces the variation in these estimates due to noise in the data, hence reduces the variance of the estimated optimal forecast.
  - The higher the bias and/or variance of some estimator, the more poorly it performs on unseen data. [See James et al., 2014 - An Introduction to Statistical Learning with Applications in R, pp. 33-35].
- In practice, the choice of p is guided by information criteria.
  - The Akaike information criterion is $AIC_m := ln\frac{SSR_m}{T} + m\frac{2}{T}$. The model selected by the Akaike information criterion is the $AR(p)$ model that minimises AIC_m.
  - The Bayesian information criterion is $BIC_m := ln\frac{SSR_m}{T} + m\frac{lnT}{T}$. The model selected by the Bayesian information criterion is the $AR(p)$ model that minimises BIC_m.
  - In each of the above, m is the number of parameters in the regression model (including the constant, so for an $AR(p)$ model, $m = p+1$) and T is the number of observations. SSR_m is the sum of squared residuals in the regression of Y_t on $Y_{t-1}, \dots, Y_{t-p}$.
- Suppose that some time series data is in fact generated by an autoregressive $AR(p_0)$ model, and that the information criterion are used to select p from $\{0, 1, \dots, p_{max}\}$ where $p_{max} > p_0 > 0$.
  - In general, for small T, both information criteria tend to select models with order lower than p_0. As T grows, both information criteria tend to select larger models.
  - As T approaches ∞, the probability that BIC selects $p = p_0$ converges to 1. This is not true of AIC, which even in large samples, selects $p > p_0$ with non-zero probability.

Autoregressive Distributed Lag Models

- An autoregressive distributed lag model $ADL(p, q)$ of order (p, q) is some set of equations that describes the relationship between a time series variable Y_t and (1) its lags $Y_{t-1}, \dots, Y_{t-p}$, and (2) the lags of another time series variable $X_{t-1}, \dots, X_{t-q}$, namely the equation $Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{i=1}^{q} \delta_i X_{t-i} + u_t$, where $\mathbb{E}[u_t | \mathcal{Y}_{t-1}, \mathcal{X}_{t-1}] = 0$, i.e. u_t is mean-zero and is not forecastable on the basis of past realisations of Y, X, Y_0, X_0 is given, and u_t is some stationary shock.

Forecasting

- The optimal 1-step ahead forecast of an $ADL(p,q)$ process $\{Y_t\}$ is
  $$Y_{T+1|T} := \mathbb{E}[Y_{T+1}|\mathcal{Y}_T, \mathcal{X}_T]$$
  - $$= \beta_0 + \sum_{i=1}^{p} \beta_i Y_{T+1-i} + \sum_{i=1}^{q} \delta_q X_{T+1-i}.$$

Estimation

- Orthogonality holds in an $ADL(p,q)$ model (by construction), hence its parameters can be consistently estimated by OLS regression of Y_t on $Y_{t-1}, \dots, Y_{t-p}, X_{t-1}, \dots, X_{t-p}$.
  - Again, supposing that $\{Y_t\}, \{X_t\}$ are jointly stationary, the OLS estimators are asymptotically normal, and the familiar methods for performing hypothesis tests and constructing confidence intervals apply.
- The estimated optimal 1-step ahead forecast is obtained by substituting the OLS estimates for their population counterparts in the optimal 1-step ahead forecast.

Granger Causality

- $\{X_t\}$ does not Granger cause $\{Y_t\}$ iff lags of $\{X_t\}$ carry no useful information about $\{Y_t\}$ in addition to that carried by its own lags. Formally, this is iff
  - $\mathbb{E}[Y_{T+1} - \mathbb{E}[Y_{T+1}|\mathcal{Y}_T, \mathcal{X}_T]]^2 = \mathbb{E}[Y_{T+1} - \mathbb{E}[Y_{T+1}|\mathcal{Y}_T]]^2$,
  - i.e. the mean-squared forecast error on the basis of lags of Y_t and lags of X_t is equal to that on the basis of lags of Y_t alone.
- $\{X_t\}$ Granger causes $\{Y_t\}$ otherwise. Formally, this is iff
  - $\mathbb{E}[Y_{T+1} - \mathbb{E}[Y_{T+1}|\mathcal{Y}_T, \mathcal{X}_T]]^2 < \mathbb{E}[Y_{T+1} - \mathbb{E}[Y_{T+1}|\mathcal{Y}_T]]^2$,
  - i.e. the mean-squared forecast error on the basis of lags of Y_t and lags of X_t is less than that on the basis of lags of Y_t alone.
- Given that the conditional expectation minimises the mean-squared forecast error, i.e. $\mathbb{E}[Y_{T+1} - \mathbb{E}[Y_{T+1}|\mathcal{Y}_T, \mathcal{X}_T]]^2 = \mathbb{E}[Y_{T+1} - \mathbb{E}[Y_{T+1}|\mathcal{Y}_T]]^2$ iff $\mathbb{E}[Y_{T+1}|\mathcal{Y}_T] = \mathbb{E}[Y_{T+1}|\mathcal{Y}_T, \mathcal{X}_T]$.

Granger Causality Test

- Given that $\{Y_t\}$ follows some $ADL(p,q)$ model, where $p=q$ is given, the test of Granger causality is a F test of
  - $H_0 : \delta_1 = \ldots = \delta_p = 0$, against
  - $H_1 : \exists i \in \{1, \ldots, p\} : \delta_i \neq 0$, in
  - $Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{i=1}^{p} \delta_i X_{t-i} + u_t$, where
  - $\mathbb{E}[u_t|\mathcal{Y}_{t-1}, \mathcal{X}_{t-1}] = 0$ by construction.
  - Supposing that $\{X_t\}$ and $\{Y_t\}$ are jointly stationary, the critical values for this test are drawn from the usual $F_{q,\infty}$ distribution for a test of q restrictions.
  - A rejection of the null can be interpreted as a finding that $\{X_t\}$ Granger causes $\{Y_t\}$.