STA457S 2023 Summer

Anton S.

- **Professor:** Esam Mahdi, `e.mahdi@mail.utoronto.ca`
- **Lectures:** MW 6-9 BA1160

Contents

1. Introduction

Time series can be defined as a collection of random variables indexed according to the order they are obtained in time. We can consider time series as a sequence of random variables

$$x_1, x_2, \dots, x_t, \dots$$

where x_t is obtained at t-th time point. In this course, the indexing variable t will typically be discrete and not continuous. I.e. $t \in \mathbb{N}$ or $t \in \mathbb{Z}$. A *time series* is a series of observed values (x_t), we call the unrealized model a *process* in this course.

Definition 1.0.1. A series is *stationary* if it remains around a mean value over time.

Examples: Daily temperature, stock prices, generally measurements

Box-Jenkins Methodology

1. **Identification:** Examine graphs and identify patterns and dependency in an observed time series. We look for: *trend, periodic trend, outliers, irregular change*
2. **Estimation:** Select a suitable fitted model for predicting future values.
3. **Diagnostic checking:** Goodness of fit tests and residual scores to estimate adequacy of the model, determine unaccounted for patterns.
4. **Forecasting:** Use model to forecast the future values.

We say forecasting instead of prediction to indicate foretelling closely into the future.

Financial Time Series

We motivate a lot of this course with financial data, so we define terminology for financial time series.

Definition 1.0.2. The *net return* from the holding period $t-1$ to t is

$$R_t = \frac{x_t - x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - 1$$

i.e. relative percent increase of (x_k) from $t-1$ to t.

Definition 1.0.3. The *simple gross return* from the holding period $t-1$ to t is

$$\frac{x_t}{x_{t-1}} = 1 + R_t$$

Definition 1.0.4. The *gross return over the most recent k periods* is defined as

$$1 + R_t(k) = \frac{x_t}{x_{t-k}} = \prod_{i=0}^{i=k} \frac{x_{t-i}}{x_{t-i-1}} = (1 + R_t) \dots (1 + R_{t-k})$$

Definition 1.0.5. The *log returns* or *continuously compounded returns* are denoted r_t and defined as

$$r_t = \log(1 + R_t) = \log(x_t) - \log(x_{t-1})$$

Returns are scale-free but not unitless since they depend on t.

Definition 1.0.6. The *volatility* is the conditional standard deviation of underlying asset return.

In most financial time series data, the scale of the volatility appears to be the same. Highly volatile periods tend to be clustered together.

We may decompose a financial time series as

$$x_t = \underbrace{T_t}_{\text{trend}} + \underbrace{s_t}_{\text{season}} + \underbrace{c_t}_{\text{cycle}} + \underbrace{I_t}_{\text{irregularity}}$$

If these components are corelated, use a multiplicative decomposition $x_t = T_t s_t c_t I_t$. If only some are corelated, use a mixed model, i.e. $x_t = s_t T_t + c_t + I_t$.

Time Series Models

Definition 1.0.7 (Moving average)**.** The k-th (odd) moving average of a time series (x_t) is defined as the sum of the k values of the time series around x_t. For example, the third moving average series for (x_t) is

$$y_t = \frac{1}{3}(x_{t-1} + x_t + x_{t+1})$$

If k is even, we reindex and define the time of the moving average to be at the middle of the times we evaluate. For example the 4-th moving average of (x_t) is

$$y_t = \frac{1}{4}(x_{t-2} + x_{t-1} + x_{t+1} + x_{t+2})$$

Moving averages allow us to 'smooth' a time series by reducing the noise while maintaining the trend in the series.

Definition 1.0.8 (White noise)**.** A *white noise process* is a collection of uncorrelated and identically distributed random variables (w_t), each with 0 mean and finite variance σ_w^2 for every t. If the white noise follows a normal distribution, i.e.

$$w_t \sim N(0, \sigma_w^2)$$

then it is *Gaussian white noise*. In the Gaussian case, independent and uncorrelated are the same, so w_t are i.i.d.

Definition 1.0.9 (Random walk)**.** A *random walk with drift* (x_t) is a series

$$x_t = \delta + x_{t-1} + w_t$$

where $w_t \sim \mathrm{wn}(0, \sigma^2)$. For $t \geqslant 1$, δ is the *drift*. When $\delta = 0$, the series is simply a random walk:

$$x_t = x_{t-1} + w_t$$

The series is the same as in the previous time step plus a white noise shock. Therefore we may write

$$x_t = \delta t + \sum_{j=1}^{t} w_j, \quad t \geqslant 1$$

If $\delta \neq 0$, the series is not stationary.

Definition 1.0.10 (Signal in noise)**.** Many realistic models for generating time series assume an underlying sinusoidal signal:

$$x_t = A\sin(\omega t + \phi) + \omega_t$$

As a general note, the goal of time series analysis is to apply a series of transformations in order to reduce the remaining model to a white noise series. Through these transformations we address trends in the series, aiming to be left with only a noise series.

2. Characteristics of Time Series

A complete description of time series is provided by the joint distribution function.

Definition 2.0.1. The *mean function* is defined as

$$\mu_t = E(x_t) = \int_{-\infty}^{\infty} x f_t(x)dx$$

μ_t is the expectation of the process at the given t, f_t is probability density of x_t.

Definition 2.0.2. The *autocovariance function* is defined as the second moment product

$$\gamma_x(s,t) = \text{Cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)]$$

for all s, t. Note $\gamma_x(t,t) = \text{Var}(x_t)$.

Covariance measures the 'linear relationship' of random variables (it is an inner product on the space). The following examples can be computed with the bilinearity properties of covariance.

Example 1. Consider white noise $w_t \sim \text{wn}(0, \sigma^2)$. Then we have

$$\gamma_w(s,t) = \text{Cov}(w_s, w_t) = \begin{cases} \sigma^2, & s = t \\ 0, & s \neq t \end{cases}$$

Example 2. Consider moving average $v_t = \frac{1}{3}(w_{t+1} + w_t + w_{t-1})$ with $w_t \sim \text{wn}(0, \sigma^2)$. Then we can verify that

$$\gamma_v(s,t) = \begin{cases} \frac{1}{3}\sigma^2, & s = t \\ \frac{2}{9}\sigma^2, & |s-t| = 1 \\ \frac{1}{9}\sigma^2, & |s-t| = 2 \\ 0, & |s-t| > 2 \end{cases}$$

Note: Prof said this is a great exam question!

Example 3. For a random walk without drift, $x_t = \sum_{j=1}^{t} w_j$ and $w_t \sim \text{wn}(0, \sigma^2)$, we have

$$\gamma_x(s,t) = \min\{s,t\}\,\sigma^2$$

since the w_t are uncorrelated random variables. Note $\text{Var}(x_t) = t\sigma^2$.

Definition 2.0.3. The **autocorrelation function** is defined as

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}$$

The autocorrelation function gives a profile of the linear correlation of the series at time t. Cauchy Schwarz implies $|\gamma(s,t)|^2 \leqslant \gamma(s,s)\gamma(t,t)$.

Definition 2.0.4. For multivariate time series we have the **cross-variance** function

$$\gamma_{xy}(s,t) = \mathrm{Cov}(x_s, y_t)$$

and **cross-correlation** function

$$\rho_{xy}(s,t) = \frac{\sqrt{\gamma_{xy}(s,t)}}{\sqrt{\gamma_x(s,s)}\sqrt{\gamma_y(t,t)}}$$

This can be extended to time series with arbitrary components.

Stationary Models

Definition 2.0.5. A **stationary process** x_t has constant mean and variance for all t.

Stationarity is defined uniquely, so there is only one way for a series to be stationary. It is preferred that estimators of parameters do not changed over time. In many cases, stationary data can be approximated with stationary ARMA models which we discuss later. They also avoid the problem of *spurious regression*.

Definition 2.0.6. A series x_t is **strong stationary** if for any $t_1, t_2, \dots, t_n \in \mathbb{Z}$ where $n \geqslant 1$ and any scalar shift $h \in \mathbb{Z}$, the joint distribution of both series is the same:

$$P(x_{t_1} \leqslant c_1, \cdots, x_{t_n} \leqslant c_n) = P(x_{t_1+h} \leqslant c_1, \cdots, x_{t_n+h} \leqslant c_n)$$

We never actually know the joint distribution, but this definition allows us to make some theoretical observations about time series. The above implies

1. $p(x_t \leqslant c) = p(x_{t+h} \leqslant c)$
2. $\mu_t = \mu_s$ for all s, t
3. $\gamma(s,t) = \gamma(s+h, t+h)$

It cannot be checked whether any observed time series is strong stationary. This motivates *weak stationary*.

Definition 2.0.7. A process is **time invariant** if it does not depend on time.

Definition 2.0.8. A time series is **weak stationary invariant, covariance stationary, second-order stationary** if

1. μ_t is constant
2. $\gamma(s,t) = \mathrm{Cov}(x_s, x_t)$ depends on s, t only by the difference $|s-t|$: $\gamma(t+h, t) = \gamma(h, 0)$.

Proposition 1. A strong stationary series is weakly stationary. The converse is not true.

Definition 2.0.9. The **autocovariance function of a stationary time series** will be written as

$$\gamma(h,0) = \gamma(h) = \mathrm{Cov}(x_{t+h}, x_t)$$

Note $\gamma(h) = \gamma(-h)$.

Definition 2.0.10. The **autocorrelation function of a stationary time series** will be written as

$$\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)}\sqrt{\gamma(t,t)}} = \frac{\gamma(h)}{\gamma(0)}$$

Definition 2.0.11. The stochastic process w_t is a **strong white noise process** with mean zero and variance σ_w^2 and written $w_t \sim \text{wn}(0, \sigma_w^2)$ if and only if it is i.i.d. with zero mean and covariance

$$\gamma_w(h) = E(w_t w_{t+h}) = \begin{cases} \sigma_w^2, & h = 0 \\ 0, & h \neq 0 \end{cases}$$

A weak stationary *Gaussian* white noise process is strongly stationary, due to uncorrelated implying independent in this case.

Example 4. Consider moving average $v_t = \frac{1}{3}(w_{t+1} + w_t + w_{t-1})$ with $w_t \sim \text{wn}(0, \sigma^2)$. It is stationary since $\mu_{v,t} = 0$.

$$\gamma_v(h) = \begin{cases} \frac{1}{3}\sigma^2, & h = 0 \\ \frac{2}{9}\sigma^2, & h = 1 \\ \frac{1}{9}\sigma^2, & h = 2 \\ 0, & h > 2 \end{cases} \qquad \rho_v(h) = \begin{cases} 1 & h = 0 \\ \frac{2}{3} & h = 1 \\ \frac{1}{3} & h = 2 \\ 0, & h > 2 \end{cases}$$

Example 5. $x_t = \varepsilon_t$ where $\varepsilon_t \sim \text{i.i.d}(0,1)$ is weakly stationary.

Example 6. $x_t = t + \varepsilon_t$ where $\varepsilon_t \sim \text{i.i.d}(0,1)$ is not weakly stationary since μ_t depends on t.

Example 7. Suppose $X_t = A\sin(t+B)$ where $A \sim r.v.(0,1)$, $B \sim U([-\pi,\pi])$. This process is stationary.

$$E(X_t) = E(A\sin(t+B)) = E(A)E(\sin(t+B)) = 0$$
$$\gamma(h) = \frac{1}{2}\cos(h)$$

$\gamma(h)$ can be verified by integrating.

Transforming Nonstationary Series

The random walk process $x_t = \delta t + \sum_{j=1}^{t} w_j$ is not stationary if it has drift, since $E(x_t) = \delta t$ depends on time. Suppose $\delta = 0$ so the mean function is constant. In this case

$$\gamma(h) = \text{Cov}(x_t, x_{t+h}) = t\sigma^2 \text{ and } \rho(h) = \frac{\text{Cov}(x_t, x_{t+h})}{\sqrt{\text{Var}(x_t)\text{Var}(x_{t+h})}} = \frac{1}{\sqrt{1+h/t}}$$

For large t and h much smaller than t, get $\gamma(h)$ is very close to 1. We can eliminate the stationarity in a random walk process by taking the difference of the x_t:

$$\nabla x_t = x_t - x_{t-1} = \varepsilon_t \sim \text{wn}(0, \sigma_w^2)$$

In the presence of d unit rots, we apply d differences to x_t:

$$\nabla^d x_t = (1-B)^d x_t = \varepsilon_t$$

Where B is the backwards shift $Bx_t = x_{t-1}$. For example, consider $x_t = a + bt + ct^2$. Then we may take second order differences:

$$z_t = \nabla^2 x_t = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) = 2c$$

for $t \geqslant 3$. The R function `diff(x, lag, differences)` can be used for this. The series ∇x_t can be used to transform the time series into stationarity.

Definition 2.0.12. A series is **jointly stationary** if they are each stationary and

$$\gamma_{xy}(h) = \text{Cov}(x_{t+h}, y_t) = E(x_{t+h} - \mu_x)(y_t - \mu_y)$$

is only a function of the lag. The **cross correlation function** of two jointly stationary series is

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}$$

We again have $-1 \leqslant \rho_{xy}(h) \leqslant 1$.

Example 8. Consider two series $x_t = w_t + w_{t-1}$ and $y_t = w_t - w_{t-1}$. We find the cross correlation

Definition 2.0.13. A **linear process** x_t is defined to be a linear combination

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_j, \qquad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty$$

We may verify $\gamma_x(h) = \sum_{j=-\infty}^{\infty} \psi_{t+h}\psi_t$. Only need $\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty$ for process to have finite variance. Note that the moving average is an example of a linear process.

If a time series is stationary, we may estimate the mean with $\overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$. In this case,

$$\text{Var}(\overline{x}) = \frac{\sigma_x^2}{n}\left(1 + \sum_{h=1}^{n-1}(1 - h/n)\rho(h)\right)$$

Estimators

Definition 2.0.14. The **sample autocovariance** is defined as

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \overline{x})(x_t - \overline{x})$$

The sum is restricted since x_{t+h} is not available for $t + h > n$. This estimator is preferred than the one dividing by $n - h$ since it is a non-negative definite function. The **sample autocorrelation** is defined as

$$\hat{\rho}(0) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{\sum_{t=1}^{n-h}(x_{t+h} - \overline{x})(x_t - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})^2}$$

This allows us to test whether the autocorrelation is statistically significant at some lags: for n sufficiently large, approximately we have $\hat{\rho}(h) \sim N(0, \frac{1}{n})$. I.e. the estimator is normally distributed with

$$\mu_{\hat{\rho}(h)} = 0 \text{ and } \sigma_{\hat{\rho}(h)} = \frac{1}{\sqrt{n}}$$

We can test $H_0 : \rho(h) = 0$, $H_a : \rho(h) \neq 0$. For $\alpha = 0.05$, have $|\hat{\rho}(h)| \geqslant \frac{2}{\sqrt{n}}$.

- The ACF **cuts off at lag** h if there no spikes at lags $> h$ in the ACF plot.
- The ACF **dies down** if it decreases in a steady fashion.
- If ACF dies down quickly, then the data is stationary. If it dies down very slowly, it is not stationary.

Vector Valued Time Series

Same as regular time series, except

$$\mathbf{x}_t = (x_{t,1}, \ldots, x_{t,p}) \in \mathbb{R}^p$$

The transpose is denoted $\mathbf{x}_t$. The mean is $\mu_t = E(x_t) = (\mu_{t,1}, \ldots, \mu_{t,p})$. If the process is stationary, $E(x_t) = \mu$, and has autocovariance matrix

$$\Gamma(h) = E(x_{t+h} - \mu)(x_t - \mu)'$$

with cross covariance functions $\gamma_{ij}(h) = E(x_{t+h,i} - \mu_i)(x_{t,j} - \mu_j)$. Note $\gamma_{ij}(h) = \gamma_{ji}(-h)$.

Definition 2.0.15. The sample autocovariance matrix

$$\hat{\Gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \overline{x})(x_t - \overline{x})'$$

where $\overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$. The symmetry property holds: $\hat{\Gamma}(h) = \hat{\Gamma}(-h)'$.

3. Time Series Regression and Exploratory Data Analysis

We develop regression models in univariate and multiple time series analyis. We calculate least squares estimators of regression parameters, do ANOVA, and assess our parameters. Then we perform lagged regression, and do transformations of time series to stationarity.

The multiple linear regression model relates the response x to independent variables z_i with the relationship

$$x = \beta_0 + \beta_1 z_1 + \ldots + \beta_q z_q + \varepsilon$$

where ε is some error term. We model

$$E(x \mid z_1, \ldots, z_q) = \beta_0 + \beta_1 z_1 + \ldots + \beta_q z_q$$

The linear model is *linear in the coefficients* β_1, not in z_i.

Definition 3.0.1. The multiple linear regression model in time series is modelled with

$$x_t = \beta_0 + \beta_{t,1} + \ldots + \beta_q z_{t,q} + w_t$$

1. x_t is the **dependent time series**
2. $z_{t,1}, \ldots, z_{t,q}$ are **independent series**.
3. w_t for different t are iid, $\text{wn}(0, \sigma_w^2)$. Note this is stronger than the usual assumption.

We collect $n > q$ observations of the time series, at various time points and predict $\hat{x}_t = \hat{\beta}_0 + \hat{\beta}_1 z_{t1} + \ldots \hat{\beta}_q z_{tq}$. We describe x_t as a linear combination of the other time series. We minimize the error via least squares:

$$Q(\beta_0, \ldots, \beta_q) = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n}(x_t - \hat{x}_t)^2$$

Then differentiate and minimize by setting

$$\frac{\partial Q}{\partial \beta_i}\bigg|_{\beta_0,\dots,\beta_q} = 0$$

When $q = 1$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_t - \overline{x})(z_t - \overline{z})}{\sum_{i=1}^n (z_t - \overline{z})^2}, \quad \hat{\beta}_0 = \overline{x} - \hat{\beta}_z \overline{z}$$

Exam: Should be on reference sheet.

Matrix Form

We can write the multiple linear regression model in terms of vector valued time series/matrix form. Consider $z_t \in \mathbb{R}^q$ with component-wise independent time series $z_{t,i}$. Each z_t can be seen as a column vector of z. Then for the model

$$x_t = \beta' z_t + w_t \quad w_t \sim \text{iid}(0, \sigma_w^2)$$

the least squares estimate is given by

$$\hat{\beta} = (z'z)^{-1} z'x = \left(\sum_{t=1}^n z_t z_t' \right)^{-1} \sum_{t=1}^n z_t x_t$$

The minimized **sum squared errors** can be written

$$\mathsf{SSE} = \sum_{t=1}^n (x_t - \hat{\beta}' z_t)^2$$

The covariance matrix is given by

$$\text{Cov}(\hat{\beta}) = \sigma_w^2 C, \quad C = (zz')^{-1}$$

i.e. the exterior product. The **mean squared error** is

$$\mathsf{MSE} = s_w^2 = \frac{\mathsf{SSE}}{n - (q+1)}$$

which is an unbiased estimator for σ_w^2.

Hypothesis Testing and Model Selection

We may test the hypothesis $\beta_i = 0$ for $i > 0$ with the test statistic

$$t = \frac{\hat{\beta}_i \beta_i}{s_w \sqrt{c_{i,i}}} \sim t_{n-(q+1)}$$

where $c_{i,i}$ is the i-th diagonal element of the covariance matrix C. We can also test whether a subset of z_i influences x_t. The **reduced model** is

$$x_t = \beta_0 + \beta_1 z_{t1} + \ldots + \beta_r + z_{tr} + w_t$$

where $\beta_0, \ldots, \beta_r$ are a subset of the original coefficients. Our null hypothesis is $\beta_{r+1} = \cdots = \beta_q = 0$. We are testing whether the SSE deviates statistically significantly once we reduce the model, since it will always reduce somewhat. Our null is that the subset model is correct, since we prefer more parsimonious models.

$$F = \frac{(SSE_R - SSE)/(q-r)}{SSE/(n-q-1)} = \frac{MSR}{MSE} \sim F_{q-r,n-q-1}$$

Note: $n - q - 1 - (n - r - 1) = q - r$ which gives the above degrees of freedom. Reject the more parsimonious model at level α in favor of H_a if $F \geqslant F_\alpha$.

Sources of Variation	Df	Sum Sq	Mean Sq	F value	Pr(>F)
$z_{t;r+1;q}$	$q-r$	SSR	$MSR = \frac{SSR}{q-r}$	$F_0 = \frac{MSR}{MSE}$	etc
Error	$n-q-1$	SSE	$MSE = \frac{SSE}{n-q-1}$		
Total	$n-r-1$	SSE_0			

- The **sum of squares contributed by regression** (explained variation): $SSR = \sum_{t=1}^{n}(\hat{x}_t - \overline{x})^2$.
- The **sum of squares contributed by error** (unexplained variation): $SSE = \sum_{t=1}^{n}(\hat{x}_t - \hat{x}_t)^2$.
- The **total** sum squared is $SSE_0 = SSR + SSE$.
- The **coefficient of determination** is the proportion of total explained variation is

$$R^2 = SSR/SSE_0 = 1 - SSE/SSE_0$$

.

In order to compare models we also consider the adjusted R^2, which account for the number of predictors.

$$R_a^2 = \left(R - \frac{q}{n-1}\right)\left(\frac{n-1}{n-q-1}\right)$$

For a model with $k+1$ parameters, the least squares estimator of the variance is $\hat{\sigma}_k^2 = SSE(k)/n$, where $SSE(k)$ comes from the model without intercept. Frequently we use *information criteria* to select the best model with k predictors

- $AIC = n \log \hat{\sigma}_k^2 + 2(k+2)$
- $AIC_C = AIC + \frac{2(k+2)(k+3)}{n-k-3}$
- $BIC = \log \hat{\sigma}_k^2 + (k+2)\log n$

We prefer models with minimal information criteria.

Example 9. Consider a time series M_t which is modelled as depending on other series T_t, P_t.

- **Trend-only**: $M_t = \beta_0 + \beta_1 t + w_t$
- **Linear**: $M_t = \beta_0 + \beta_1 t + +\beta_1 T_t + w_t$ or $M_t = \beta_0 + \beta_1 t + +\beta_1 P_t + w_t$ etc.
- **Curvilinear**: $M_t = \beta_0 + \beta_1 t + +\beta_1 (T_t - \overline{T})^2 + \beta_2 P_t + w_t$

The model simultaneously minimizing AIC and BIC is best. Note that the quadratic term in the curvilinear model is centered, probably to account for average temperature in β_0. Given observations for these models, we perform F-test to see whether we can drop some predictors.

When dealing with temporal data, we also need to consider **lagged variables**. This predicts values of x_t from possible lags in z_t. Lagged regression can be done using `dynlm` in R.

Transformations to Stationarity

In order to satisfy many of our assumptions, it is necessary for a series to be stationary. This is often not the case and we often want to transform our data. To remove any change in the mean function μ_t, we detrend the model by decomposing it into

$$x_t = \mu_t + y_t$$

where μ_t is a fitted mean function, y_t is the residual series. Our assumption about errors is that they follows $iid(0, \sigma^2)$, which makes y_t stationary.

The backshift, forward, and difference operators act on time series by

- **Backshift:** $B^h(x_t) = x_{t-h}$
- **Forward:** $B^{-h}(x_t) = x_{t+h}$
- **Difference:** $\nabla^h(x_t) = (1 - B)^h(x_t)$

Often taking the first difference is more effective than detrending in order to make the series stationary. ACF plots end up much better.

Example 10. Suppose that after differencing, the ACF plot had a significant value at $h = 4$. Then we model

$$X_t = \theta X_{t-4} + w_t$$

We see later that this is a "$MA(4) = ARMA(0, 4)$" model.

When a model has drift, for example $X_t = \delta + X_{t-1} + w_t$, then differencing makes complete sense in order to get a stationary series.

Fractional differencing extends the notion of the difference operator $\nabla^d = (1 - B)^d$ to fractional powers of $d \in \left(-\frac{1}{2}, \frac{1}{2}\right)$ which still define stationary processes, especially for **long memory time series**.

A method to suppress large fluctuations of x_t is through the **Box-Cox** transformations:

$$y_t = \begin{cases} \frac{x_t^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log x_t, & \lambda = 0 \end{cases}$$

which is a method of selecting the best non-linear transformation of x_t in order to minimize the variance of the errors.

Harmonic regression is used when a model contrains a **periodic** trend, allowing us to use trigonometric functions of t to do detrending.

$$x_t = \beta_0 + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{L}\right) + \beta_3 \cos\left(\frac{2\pi t}{L}\right) + w_t$$

Adding trigonometric terms with different frequencies can help with complex seasonality patterns.

Filtering and Smoothing

Filtering/smoothing helps discover useful trends and seasonal components.

Definition 3.0.2. The **moving average smoother** is

$$m_t = \sum_{j=-k}^{k} a_j x_{t-j}$$

where $\sum_{j=-k}^{j} a_j = 1, a_j > 0$ makes a symmetric weighted moving average.

Definition 3.0.3. The **kernel smoothing** is

$$m_t = \sum_{i=1}^{n} w_i(t) x_i$$

where $w_i = K(\frac{t-i}{b}) / \sum_{j=1}^{n} K(\frac{j-i}{b})$ are weights, K is some kernel function. The wider the **bandwidth** b, the smoother the model.

4. ARIMA Models

We move into the core of time series analysis. ARMA models are defined, autocorrelation functions are derived, and stationarity, causality, and invertibility of series are evaluated. The Box-Jenkins methodology requires that the model used in describing and forecasting a series is stationary and invertible

Definition 4.0.1. x_t is **stationary** if it remains in statistical equilibrium with properties that do not change over time. x_t is **invertible** if its weights do not depend on time, and x_t can be expressed as a function of previous observations $x_{t-1}, \ldots$.

Definition 4.0.2. The **partial correlation** at lag k of x_t is

$$\mathrm{Corr}(x_{t+k} - \hat{x}_{t+k}, x_t - \hat{x}_t)$$

where $\hat{x}_{t+k} = \beta_1 x_{t+k-1} + \beta_{k-1} z_{t+1}$ and $\hat{x}_t = \beta_1 x_{t+1} + \beta_{k-1} z_{t+k-1}$. Note coefficients are same but reversed. The partial autocorrelation allows us to detect whether a dependence at lag k is appropriate, and is part of the Box-Jenkins methodology.

Auto-Regressive Models

Once trends and seasonal effects are removed from a model, we might construct a linear model for a series with autocorrelation.

Definition 4.0.3. A time series x_t with zero mean is **autoregressive process of order** p, denoted $AR(p)$ if it can be written

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots \phi_p x_{t-p} + w_t$$

for $\phi_p \neq 0$, $w_t \sim wn(0, \sigma_w^2)$. With backshift operator, we can write this as a polynomial of order p in B,

$$\Phi_p(B) x_t = w_t$$

and $\Phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$, $\phi_p \neq 0$. This is the **characteristic polynomial** of order p.

The second expresssion in terms of characteristic polynomial is preferred, we will see it simplifies our understanding later. If the mean μ of x_t, we may replace x_t by $x_t - \mu$, and rewrite as

$$x_t = \delta + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots \phi_p x_{t-p} + w_t$$

with $\delta = \mu(1 - \phi_1 - \phi_2 - \dots - \phi_p)$.

Example 11. The $\mathrm{AR}(2)$ model $x_t = 1.5 + 1.2x_{t-1} - 0.5x_{t-2} + w_t$ is

$$x_t - \mu = 1.2(x_t - \mu) - 0.5(x_t - \mu) + w_t$$

Solving for μ with $1.5 = \mu(1 - 1.2 - 0.5)$, we see $\mu = 5$.

Suppose we fit an $\mathrm{AR}(h)$ model. In order to decide whether the fit model is a good fit, we check:

- The plot of the time series does not show any increase in variance or trend.
- The ACF plot must decay exponentially, have a wavelet form, or be oscillating (i.e. sign alternates) about 0.
- The PACF plot can be used to detect the correct order for the autoregressive model.

Causal Conditions

We study whether a process can be completely described by its previous values.

Definition 4.0.4 (Causal conditions for $\mathrm{AR}(1)$)**.** The autoregressive process of order 1, $\mathrm{AR}(1)$, $x_t = \phi x_{t-1} + w_t$ is a **causal process** if it is stationary with values that are not depending on the future. In this case, the absolute value of the root of $1 - \phi z = 0$ must lie outside the unit circle. $\mathrm{AR}(1)$ process is causal if

$$|z| = \left|\frac{1}{\phi}\right| > 1 \iff |\phi| < 1$$

A causal process is stationary, but a stationary process is not necessarily causal.

Example 12.

1. $(1-0.4B)x_t = w_t$ is causal since the root of $(1-0.4z) = 0$ satisfies $|z| = |1/0.4| > 1$.
2. $(1 + 1.8B)x_t = w_t$ is not causal since $|1/\phi| < 1$.

Definition 4.0.5 (Causal conditions for $\mathrm{AR}(2)$)**.** The $\mathrm{AR}(2)$ model

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$$

is **causal** when the roots of the characteristic polynomial

$$\Phi_2(z) = 1 - \phi_1 z - \phi_2 z^2$$

lie outside the unit circle

$$\left|\frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}\right| > 1$$

Necessary and sufficient conditions for this are

$$|\phi_2| < 1 \quad \phi_1 + \phi_2 < 1 \quad \phi_2 - \phi_1 < 1$$

Example 13. 1. $x_t = 1.1x_{t-1} - 0.4x_{t-2}$ is causal.

2. $x + t = 0.6x_{t-1} - 1.3x_{t-2} + w_t$ is not stationary (necessary and sufficient conditons).

Definition 4.0.6 (Causal conditions for $\mathrm{AR}(p)$)**.** The autoregressive process of order p, $\mathrm{AR}(p)$,

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots \phi_p x_{t-p} + w_t$$

is a **causal process** if *all* roots of the characteristic polynomial

$$\Phi_p(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p,\ \phi_p \neq 0$$

lie outside the unit circle.

The function `polyroot(a)`, where `a` is a vector with polynomial coefficients, can be used to find the roots.

Moving Average Models

These are analogous to autoregressive models, except moving average models depend on white noise terms instead of terms of the series itself. There is an analogous characteristic polynomial $\Theta_q(B)$, with the same root condition on *invertibility* instead of causality.

Definition 4.0.7. A time series x_t with zero mean is a **moving average process** of order q, denoted $\mathrm{MA}(q)$, if it can be written

$$x_t = \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots \theta_q w_{t-q}$$

where $w_t \sim \mathrm{wn}(0, \sigma_w^2)$ and $\theta_q \neq 0$. This process has characteristic polynomial $x_t = \Theta_q(B)w_t$ where

$$\Theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots \theta_q B^q,\ \theta_q \neq 0$$

If the roots z_i of the polynomial $\Theta_q(z)$ satisfy $|z_i| > 1$ for all i, then the process $\mathrm{MA}(q)$ is **invertible**.

Consider the $\mathrm{MA}(1)$ process. The autocorrelation function $\rho(h) = \dfrac{\theta}{1+\theta^2}$ does not change after replacing θ by $1/\theta$. That is

$$x_t = w_t + \theta w_{t-1} \quad \text{and} \quad x_t = w_t + \frac{1}{\theta} w_{t-1}$$

have the exact same autocorrelation function is $\rho(h)$ (show later). This is why invertibility matters: if the polynomial $\Theta_q(z)$ has all roots lying outside the unit circle, then the noise coefficients $\theta_1, \ldots, \theta_q$ are uniquely identified.
Compare the two models:

- $\mathrm{AR}(p)$: $\Phi_p(B)x_t = w_t$
- Autoregressive process is always invertible, but not always causal.
- $\mathrm{MA}(q)$: $x_t = \Theta_q(B)w_t$
- Moving average process is always causal, but not always invertible.

We check partial autocorrelation, autocorrelation plots. Out of a set of candidate models, we use AIC and BIC in order to perform model selection for AR and MA.

Auto-Regressive Moving Average Models

Definition 4.0.8. A time series x_t is an **auto-regressive moving average (ARMA)** of order (p, q) if it can be written

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \sum_{j=1}^{q} \theta_q x_{t-q}$$

also written as

$$\Phi_p(B)x_t = \Theta_q(B)w_t$$

If x_t has non-zero mean, we can rewrite the above with $\Phi_p(B)(x_t - \mu) = \Theta_q(B)w_t$. Can also be written in summation form with a constant term $\delta = \mu(1 - \phi_1 - \cdots - \phi_p)$. This is the `intercept` from `arima()`.

The ARMA satifies stationarity, invertibility, identifiability conditions if

- **Stationary:** Same condition as for AR(p) on $\Phi_p(z)$.
- **Invertible:** Same condition as for MA(q) on $\Theta_q(z)$.
- **Identifiable:** The model is not redundant. $\Phi_p(z)$ and $\Theta_q(z)$ have no common roots.

Example 14. The ARMA$(1, 2)$ model $x_t = 0.2x_{t-1} + w_t - 1.1w_t + 0.18w_{t-2}$ can be written as

$$(1 - 0.2B)x_t = (1 - 1.1B + 0.18B^2)w_t \implies x_t = (1 - 0.9B)w_t$$

which is really an ARMA$(0, 1)$ = MA(1) model. That is we can find the non-redundant expression by removing common roots of the characteristic polynomials.

Definition 4.0.9. The MA(q) process $x_t = \Theta_q(B)w_t$ where

$$\Theta_q(B) = 1 + \sum_{j=1}^{q} \theta_j B^j$$

and $w_t \sim \text{wn}(0, \sigma_w^2)$ is **invertible** if it can be represented as a convergent infinite AR form: AR(∞). Multiply both sides of above by $\Theta_q(B)^{-1}$ to get

$$w_t = \Theta_q(B)^{-1}x_t$$

Recall combinatorics and writing the above as a product of geometric series (factor the polynomial). We denote

$$w_t\Theta_q(B)^{-1}x_t = \Pi_\infty(B)x_t = 1 - \sum_{i=1}^{\infty} \pi_i B^i = -\sum_{i=0}^{\infty} \pi_i B^i$$

Note we are ensured that $\sum_{i=0}^{\infty} |\pi_i| < \infty$ with $\pi_0 = -1$.

Recall the definition of a *linear process* as defined in Section 2. Above we have shown that x_t can be written as an infinite sum of white noise series, and is therefore a linear process.

Example 15. Consider $x_t = (1 + \theta B)w_t$. Then we have the geometric series

$$w_t = \frac{1}{1 - (-\theta B)}x_t = \sum_{k=0}^{\infty} (-1)^k \theta^k B^k x_t = \Pi_\infty(B)x_t$$

This gives the expression

$$\pi_i = (-1)^{i+1}\theta^i$$

and particularly

$$x_t = \sum_{k=1}^{\infty} (-1)^{i+1}\theta^i B^i x_t + w_t$$

Note *why* we need the condition for all the roots of Θ_p to be within the unit circle: we want each geometric series in the product to converge absolutely.

Example 16. Suppose $x = w_t + 0.4w_{t-1}$. This is invertible since $|\theta| = 0.4 < 1$. We can then write

$$x_t = w_t + 0.4x_{t-1} - 0.4^2 x_{t-2} + \cdots$$

In general we know $\Pi_\infty(B) = \Theta_q(B)^{-1}$, so the coefficients π_i can be obtained by equating

$$\begin{aligned} 1 &= \Pi_\infty(B)\Theta_q(B) \\ &= 1 - (\pi_1 - \theta_1)B - (\pi_2 + \theta_1\pi_1 - \theta_2)B^2 \cdots \\ &- (\pi_j + \theta_1\pi_{j-1} + \cdots + \theta_{q-1}\pi_{j-q+1} + \theta_q\pi_{j-q})B^j \end{aligned}$$

All non-constant coefficients are 0,

$$\pi_j = -\theta_1\pi_{j-1} - \cdots - \theta_q\pi_{j-q}$$

Now what if we reverse this and do the same for a causal process?

Definition 4.0.10. The AR(p) process

$$\Phi_p(B)x_t = w_t$$

where $\Phi_p(B) = 1 - \sum_{j=1}^{p} \phi_j B^j$, $w_t \sim \text{wn}(0, \sigma_w^2)$ is **causal** if it can be represented as a convergent infinite MA(∞) form:

$$x_t = \Phi_p(B)^{-1}w_t = \Psi_\infty(B)w_t$$

where $\Phi_p(B)^{-1} = \Psi_\infty(B) = 1 + \sum_{k=1}^{\infty} \psi_k B^k$.

Using the same condition as before,

$$1 = \Psi_\infty(B)\Phi_p(B)$$

gives us $\psi_j = \phi_1\psi_{j-1} + \ldots + \phi_p\psi_{j-p}$. This Ψ is known as the **impulse response sequence**.

5. ARIMA Models Continued

Last lecture we discussed the models

- $\mathrm{ARMA}(p, 0) = \mathrm{AR}(p)$: $\Phi_p(B)x_t = w_t$ and $x_t = \Psi_\infty(B)w_t$ if this process is **causal**. The process is causal if the series representation of $1/\Phi_p(B)$ converges absolutely, which occurs when the roots of $\Phi_p(z)$ lie outside the unit circle. In this case, it is also denoted as $\mathrm{MA}(\infty)$.

$$\Psi_\infty(B) = 1 + \psi_1 B + \psi_2 B^2 + \dots$$

- $\mathrm{ARMA}(0, q) = \mathrm{MA}(p)$: $x_t = \Theta_q(B)w_t$ and $w_t = \Pi_\infty(B)x_t$ if this process is **invertible**. The process is invertible if the series representation of $1/\Theta_q(B)$ converges absolutely, which occurs when the roots of $\Theta_q(z)$ lie outside the unit circle. In this case, it is also denoted as $\mathrm{AR}(\infty)$.

$$\Pi_\infty(B) = 1 - \pi_1 B - \pi_2 B^2 - \dots$$

- $\mathrm{ARMA}(p, q)$ means that
$$\Phi_p(B)x_t = \Theta_q(B)w_t$$

The convergence follows from the partial fraction decomposition of the reciprocal of the characteristic polynomials.

Consider **causal** conditions for the $\mathrm{ARMA}(p, q)$ model. We may write

$$x_t = \frac{\Theta_q(B)}{\Phi_p(B)} w_t = \Psi_\infty(B)w_t$$

Similar to the pure AR model situation, the ψ_i coefficients may be calculated by

$$\Theta_q(B) = \Phi_p(B)\Psi_\infty(B)$$

and equating coefficients, we are left with

$$\psi_j = \phi_1\psi_{j-1} + \dots + \phi_p\psi_{j-p} + \theta_j$$

Now consider similar **invertible** conditions. Then

$$w_t = \frac{\Phi_p(B)}{\Theta_q(B)} x_t = \Pi_\infty(B)x_t$$

Using the similar equality
$$\Theta_q(B)\Pi_\infty(B) = \Phi_p(B)$$
we find
$$\pi_j = -\theta_1\pi_{j-1} - \dots - \theta_p\pi_{j-p} + \phi_j$$
The coefficients of $\Phi_\infty(B)$ are called the **impulse response coefficients**.

The ACF of an Autoregressive Process

ACF of AR(1)

Suppose we have $(1-\phi B)x_t = w_t$. When $|\phi| < 1$ we may write

$$x_t = (1+\phi B+\phi^2 B^2+\ldots)w_t = \sum_{j=1}^{\infty} \phi^j w_{t-j}$$

which is the **MA**(∞) **Wold representation**. This representation is useful because white noise variables are easy to deal with: each of the terms are uncorrelated. The below holds with series manipulations justified by $|\phi| < 1$.

1. $E(x_t) = \sum_{i=0}^{\infty} \phi^i E(x_{t-i}) = 0$
2. $\gamma(0) = \text{Var}(x_t) = \sum_{i=0}^{\infty} E(\phi^{2i}x_{t-i}^2) = \sigma_w^2 \sum_{i=0}^{\infty} \phi^{2i} = \dfrac{\sigma_w^2}{1-\phi^2}$
3. $\gamma(h) = E(x_t x_{t+h}) = \sigma_2^2 \sum_{i=0}^{\infty} \phi^{i+h}\phi^i = \sigma_w^2 \dfrac{\phi^h}{1-\phi^2}$
4. $\rho(h) = \phi(h)/\phi(0) = \phi^h$ and $\rho(h) = \phi\rho(h-1)$

These highlight some of our stationarity checks during model diagnostics. The ACF plot should have exponential decay towards 0, oscillating decay, or sine/cosine like decay.

ACF of AR(2)

Consider the AR(2) process

$$x_t = \phi_1 x_{t-1}\phi_2 x_{t-2} + w_t$$

Multiply the sides by x_{t-h} and use linearity of expectation to get

$$\begin{aligned}\gamma(h) &= E(x_t x_{t-h}) = \phi_1 E(x_{t-1}x_{t-h}) + \phi_2 E(x_{t-2}x_{t-h})\\ &= \phi_1\gamma(h-1) + \phi_2\gamma(h-2)\end{aligned}$$

where we used $E(w_t x_{t-h}) = E(w_t \sum_{j=0}^{\infty} \psi_j w_{t-h-j}) = 0$. Dividing through by $\gamma(0)$ we get the difference equation

$$\rho(h) - \phi_1\rho(h-1) - \phi_2\rho(h-2) = 0$$

Using $\rho(0) = 1,\ \rho(1) = \rho(-1)$ we have initial conditions

$$\rho(1) = \frac{\phi_1}{1-\phi_2} \qquad \rho(2) = \frac{\phi_1^2}{1-\phi_2} + \phi_2$$

ACF of AR(p)

For a general AR(p) process, we can calculate the **autocorrelation function** solving

$$\rho(h) - \phi_1\rho(h-1) - \cdots - \phi_p\rho(h-p) = 0,\ h \geqslant p$$

by following the same process as the A(2) case. For the AR(p) process we may describe the above as $D_p(B)\rho(h) = 0$ for some p-th order polynomial D_p. Consider $D_p(z) = 0$ (z can be thought of as an initial state)

- If all roots are real, $\rho(h)$ dampens exponentially as $h \to \infty$.
- If some roots are complex, then they will be in conjugate pairs and $\rho(h)$ will dampen exponentially in a sinusoidal fashion as $h \to \infty$.
- If roots are only complex, the time series will appear to be cyclic.

FINISH ! with difference equation approach. ALSO eigenvalues !

Partial Autocorrelation Function

Definition 5.0.1. Consider random variables X, Y, Z. The **partial correlation** between X, Y given Z is the correlation of the residuals of X, Y regressed on Z. That is, for $\hat{X}$ (X regress on Z) and $\hat{Y}$ (Y regress on Z), it is the correlation of

$$\rho_{XY|Z} = \text{Corr}(X - \hat{X}, Y - \hat{Y})$$

We are "removing the effect of Z".

Definition 5.0.2. The **partial autocorrelation** of stationary process x_t denoted ϕ_{hh} for $h = 1, 2, \dots$ is

$$\phi_{hh} = \text{Corr}(x_{t+h} - \hat{x}_{t+h}, x_t - \hat{x}_t),\ h \geqslant 2$$

and $\phi_{11} = \text{Corr}(x_{t+1}, x_t)$. The values are regressed on $x_{t+1}, \dots, x_{t+h-1}$, i.e. the linear dependence on these is removed. If the process is Gaussian:

$$\phi_{hh} = \text{Corr}\left(x_{t+h}, x_t \mid x_{t+1}, \dots, x_{t+h-1}\right)$$

PACF of AR(1) Process

Consider $x_t = \phi x_{t-1} + w_t$, $|\phi| < 1$. By definition, $\rho(1) = \phi$. Calculate ϕ_{22}:

1. Consider the regression x_{t+2} on x_{t+1}, say $\hat{x}_{t+2} = \beta x_{t+1}$.
2. Minimize $\hat{\beta} = \arg\min E(x_{t+2} - \hat{x}_{t+2})^2$.

$$E(x_{t+2} - \hat{x}_{t+2})^2 = \gamma(0) - 2\beta\gamma(1) + \beta^2\gamma(0) \implies \hat{\beta} = \gamma(1)/\gamma(0) = \phi$$

3. Analogously, consider the regression of x_t on x_{t+1}, $\hat{x}_t = \beta x_{t+1}$. Minimizing $E(x_t - \hat{x}_t)^2$,

$$\hat{\beta} = \phi$$

   as well.

4. By causality,

$$\begin{aligned} \phi_{22} &= \text{Corr}(x_{t+2} - \hat{x}_{t+2}, x_t - \hat{x}_t) \\ &= \text{Corr}(x_{t+2} - \phi x_{t+1}, x_t - \phi x_t) \\ &= \text{Corr}(w_t, x_t - \phi x_{t-1}) && \text{(uncorrelated noise)} \\ &= 0 \end{aligned}$$

For a given lag h, a general method for finding the autocorrelation function ϕ_{hh} for any stationary process with autocorrelation function $\rho(h)$ satisfy the **Yule-Walker** equations.

$$\rho(j) = \phi_{h1}\rho(j-1) + \phi_{h2}\rho(j-2) + \cdots + \phi_{hh}\rho(j-h)$$

and

$$\phi_{h,j} = \phi_{h-1,j} - \phi_{h,h}\phi_{h-1,j}$$

$j = 0, \dots h-1$,giving a system of h linear equations. Solving these equations gives ϕ_{hh} for any stationary process.

Proposition 2. For an AR(p) process,

$$\phi_{hh} = \begin{cases} \phi_h, & h \leqslant p \\ 0, & h > p \end{cases}$$

The sample autocorrelation is calculated by **Levinson-Durbin** equations[1]

$$\hat{\phi}_{hh} = \frac{\rho(h) - \sum_{j=1}^{h-1} \phi_{h-1,j}\rho(h-j)}{1 - \sum_{j=1}^{h-1} \phi_{h-1,j}\rho(j)}$$

Using $\phi_{11} = \rho(1)$, get

$$\phi_{22} = \frac{\rho_{22} - \phi_{11}\rho(1)}{1 - \phi_{11}\rho(1)} = \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2}$$

We can iterate to get ϕ_{hh}. Replacing ρ with $\hat{\rho}$ we may find $\hat{\phi}_{hh}$. Under the assumption that AR(p) is the correct model, then

$$\hat{\phi}_{hh} \sim N(0, 1/n)$$

The estimator is actually t-distributed but we approximate as normal for large enough n.

Autocorrelation of a Moving Average Process

ACF of MA(1)

Consider the process $x_t = w_t + \theta w_{t-1}$.

1. $E(x_t) = 0$
2. $\gamma(0) = \text{Var}(x_t) = E(w_t^2 + 2\theta w_t w_{t-1} \theta^2 w_{t-1}^2) = \sigma_w^2 (1 + \theta^2)$
3. The autocovariance is independent of h:

$$\gamma(h) = E(w_t w_{t+h}) + \theta E(w_{t-1} w_{t+h}) + \theta E(w_t w_{t+h-1}) + +\theta^2 E(w_{t-1} w_{t+h-1})$$

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma_w^2, & h = 0 \\ \theta\sigma_w^2, & h = \pm 1 \\ 0, & \text{else} \end{cases}$$

4. The autocorrelation is also independent of t.

$$\rho(h) = \begin{cases} 1, & h = 0 \\ \frac{\theta}{1+\theta^2}, & h = \pm 1 \\ 0, & \text{else} \end{cases}$$

[1] Will show up on exam.

ACF of MA(q)

For a general MA(q) process,

$$x_t = \Theta_q(B)w_t$$

with Θ_q having coefficients θ_i.

1. $E(x_t) = 0$
2. $\gamma(0) = \text{Var}(x_t) = \sigma_w^2 \left(1 + \theta_1^2 + \cdots \theta_q^2\right) = \sigma_w^2 \sum_{i=0}^q \theta_i^2$
3. The autocovariance is independent of h:
$$\gamma(h) = \begin{cases} \sigma_w^2(\sum_{i=h}^q \theta_i\theta_{i-h}), & h = 0, \pm 1, \dots, \pm q \\ 0, & \text{else} \end{cases}$$
4. The autocorrelation is also independent of t.
$$\rho(h) = \begin{cases} \frac{\sum_{i=h}^q \theta_i\theta_{i-h}}{\sum_{i=0}^q \theta_i^2}, & h = 0, \pm 1, \dots, \pm q \\ 0, & \text{else} \end{cases}$$

PACF of MA(1)

Consider $x_t = w_t + \theta w_{t-1}$, $|\theta| < 1$. The partial autocorrelation function is given by

$$\phi_{hh} = -\frac{(-\theta)^h(1-\theta^2)}{1-\theta^{2(h+1)}}$$

Therefore the theoretical PACF will have one of

- Damped exponential decay.
- Damped oscillating exponential decay.
- Damped sinusoidal exponential decay.

Note: PACF for MA models behaves like ACF for AR models. ACF for MA models behaves like PACF for AR models. Since an invertible ARMA model has an infinite AR representation,the PACF for MA models will not cut off.

A question about ACF/PACF of MA/AR or an ARMA$(1,1)$ model will probably show up on the final.

ACF of ARMA$(1,1)$

Consider the causal ARMA $(1,1)^2$ model $x_t = \phi x_{t-1} + w_t + \theta w_{t-1}$, $|\phi| < 1$. Then

$$\gamma(h) = \text{Cov}(x_{t+h}, x_t) = \phi E(x_{t+h-1}x_t) + E(w_{t+h}x_t) + \theta E(w_{t+h-1}, x_t)$$

[2] Slide 71 of Module 4 will appear on final.

Recall

$$E(w_{t+h}x_t) = \begin{cases} \psi_0\sigma_w^2, & h=0 \\ 0 & \text{else} \end{cases} \qquad E(w_{t+h-1}x_t) = \begin{cases} \psi_1\sigma_w^2, & h=0 \\ \psi_0\sigma_w^2, & h=1 \\ 0 & \text{else} \end{cases}$$

Therefore

$$\gamma(h) = \begin{cases} \phi\gamma(1) + \sigma_w^2(1+\phi\theta+\theta^2) & h=0 \\ \phi\gamma(0) + \sigma_w^2\theta & h=1 \\ \phi\gamma(h-1) & h \geqslant 2 \end{cases}$$

Note the iterative form $\gamma(h) = \phi^{h-1}\gamma(1)$ for $h \geqslant 2$, with initial conditions given by the system of equations

$$\begin{cases} \gamma(0) & = \phi\gamma(1) + \sigma_w^2(1+\phi\theta+\theta^2) \\ \gamma(1) & = \phi\gamma(0) + \sigma_w^2\theta \end{cases}$$

This gives

$$\gamma(0) = \sigma_2^2\frac{1+2\phi\theta+\theta^2}{1-\phi^2} \qquad \gamma(1) = \sigma_2^2\frac{(1+\phi\theta)(\phi+\theta)}{1-\phi^2}$$

and for $h \geqslant 1$

$$\gamma(h) = \sigma_2^2\frac{(1+\phi\theta)(\phi+\theta)}{1-\phi^2}\phi^{h-1}$$

Summary for Model Diagnostics

These ACF/PACF results we have shown in the past two classes can be summarized in the below table. The model diagnostics we discussed make sense in this context.

Model	ACF	PACF
White noise	All zeros	All zeros
$AR(p)$	Tails off as exponential decay	Spikes through lag p, cuts off
$MA(q)$	Spikes through lag p, cuts off	Tails off as exponential decay
$ARMA(p, q)$	Decay beginning at lag q	Decay beginning at lag p
Random walk	No decay to zero	All zero after lag 1

6. Time Series Diagnostics

We perform Dickey-Fuller tests for non-stationarity, address regression with autcorrelated errors, compute forecasts for ARMA models, and diagnose fitted models.

Test Statistics for Time Series Models

For a simple AR(1) model $x_t = \phi x_{t-1} + w_t$, the model is stationary when $|\phi| < 1$ and non-stationary when $|\phi = 1|$. In order to avoid over-differencing, we might want to do a hypothesis tesk of whether this is a random walk. Overdifferencing AR(1) may lead to ARMA$(1,1)$. The regression model can be written with first difference operator

$$\Delta x_t = (\phi - 1)x_{t-1} + w_t = \delta x_{t-1} + w_t$$

The model can be estimated and testing for a unit root (i.e. random walk when $\delta = 0$) by

$$H_0 : \delta = 0 \text{ or } H_1 : \delta < 0$$

I.e. the null hypothesis is that the series is non-stationary. This is the **Dickey-Fuller** unit root test. There are three versions:

1. $\Delta x_t = \delta x_{t-1} + w_t$: unit root without drift and without trend
2. $\Delta x_t = a_0 + \delta x_{t-1} + w_t$: unit root test with drift and without trend
3. $\Delta x_t = a_0 + a_1 t + \delta x_{t-1} + w_t$

Under the null hypothesis, then it can be shown

$$\hat{\phi} \sim N\left(\phi, \frac{1}{n}(1 - \phi^2)\right)$$

under the null this gives $\hat{\phi} \sim N(1, 0)$ which does not make sense. Philips showed:

$$n(\hat{\phi} - 1) \to^d \frac{(\chi_1^2 - 1)/2}{\int_0^1 W^2(t)dt}$$

where $W(t)$ is Brownian motion on [0, 1].

We **reject** H_0 if $n(\hat{\phi} - 1) \leqslant d$ for d being the tabeled value of the Dickey Fuller unit toot test statistics.

Example 17. The AR(1) model $\overline{x}_t = 0.946x_{t-1}$ where $n = 34$. Then the Dickey-Fuller test statistic is

$$n(\hat{\phi} - 1) = 34(0.946 - 1) = -1.836$$

The d statistic is $\alpha = -1.95$. Since $-1.836 > d$, we do not reject H_0, so there exists a unit root.

If x_t has a unit root, then $\Delta x_t = x_t - x_{t-1}$ will be stationary (think of random walk).

Forecasting

Forecasting is probably the most important topic in time series. The goal is to predict future values of x_t assuming we know $x_{1:n} = \{x_1, \ldots, x_n\}$. We first consider

$$x_{n+m}^n = E(x_{n+m} \mid x_1, \ldots, x_n) = \sum_{k=0}^{n} \alpha_k x_k$$

the notation x_{n+m}^n means "given n observations, predict $n+m$-th observation". The α_i depend on n, m but we do not include this in the notation for now. For example, if $n = m = 1$ then

$$x_2^1 = \alpha_0 + \alpha_1 x_1$$

Linear predictors of this form that minimize

$$Q = E(x_{n+m} - x_{n+m}^n)^2 = E\left(x_{n+m} - \sum_{k=0}^{n} \alpha_k x_k\right)^2$$

are **best linear predictors**.

Proposition 3. Given $x_1, \ldots, x_n$, the best linear predictor $x_{n+m}^m = \sum_{k=0}^n \alpha_k x_k$ is found by solving

$$E\left((x_{n+m} - x_{n+m}^n)x_k\right) = 0 \text{ for each } k = 0, 1, \ldots$$

These are the **prediction equations** and are used to solve for coefficients $\alpha_0, \ldots, \alpha_n$. The proposition is shown by minimizing with $\partial Q / \partial a_j = 0$. If the series is stationary, and $E(x_t) = \mu = E(x_{n+m}^n)$, then by taking expectations we see

$$\mu = \alpha_0 + \sum_{k=1}^{n} \alpha_k \mu$$

so

$$x_{n+m}^n = \mu + \sum_{k=1}^{n} \alpha_k (x_k - \mu)$$

The α_k can then be though of as the weight of the standard error at each observed time step x_k.

1 Step Ahead Prediction

Definition 6.0.1. The BLP of the **one step ahead predictor** can be written

$$x_{n+1}^n = \phi_{n1} x_n + \phi_{n2} x_{n-1} + \cdots + \phi_{nn} x_1 = \phi_n' \cdot x$$

The dependence of the coefficients on n is shown.

These coefficients satisfy

$$E\left(\left(x_{n+1} - \sum_{j=1}^{n} \phi_{nj} x_{n+1-j}\right) x_{n+1-k}\right) = 0$$

by Proposition 3 and since $\mathsf{E}(x_i) = 0$, since we can absorb the mean into a constant term. This can be expanded and expectations taken in order to be written

$$\sum_{j=1}^{n} \phi_{nj}\gamma(k-j) = \gamma(k)$$

As a matrix, [3]

$$\Gamma_n \phi_n = \gamma_n$$

where $\Gamma_n = \{\gamma(k-j)\}_{j,k=1}^{n}$ is an $n \times n$ symmetric matrix and $\gamma_n = (\gamma(1), \dots, \gamma(n))$. There are the **Yule-Walker** equations.

$$\begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \gamma(3) \\ \vdots \\ \gamma(n) \end{bmatrix} = \begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \cdots & \gamma(n-2) \\ \gamma(2) & \gamma(1) & \gamma(0) & \cdots & \gamma(n-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \gamma(n-3) & \cdots & \gamma(0) \end{bmatrix} \cdot \begin{bmatrix} \phi_{n1} \\ \phi_{n2} \\ \phi_{n3} \\ \vdots \\ \phi_{nn} \end{bmatrix}$$

We may therefore estimate $\hat{\phi}_n = \Gamma_n^{-1}\gamma_n$. The **mean one step ahead predictor** is

$$P_{n+1}^{n} = \mathsf{E}(x_{n+1} - x_{n+1}^{n})^2 = \gamma(0) - \gamma_n^T \Gamma_n^{-1} \gamma_n$$

since $x_{n+1}^{n} = \phi_n^T x$

Example 18 (Prediction for an AR(2))**.** Consider the AR(2) process $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$. with observation x_1.

- With one observation x_1, the one step ahead prediction of x_2 is

$$x_2^1 = \phi_{11}x_1 = \underbrace{\gamma(1)/\gamma(0)}_{\text{1D Yule-Walker}} x_1 = \rho(1)x_1$$

- With two observations x_2, the one step ahead prediction of x_2 is given by solving

$$\phi_{21}\gamma(0) + \phi_{22}\gamma(1) = \gamma(2) \text{ and } \phi_{21}\gamma(1) + \phi(22)\gamma(0) = \gamma(2)$$

  Then

$$\phi_2 = \begin{bmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{bmatrix}^{-1} \begin{bmatrix} \gamma(1) \\ \gamma(2) \end{bmatrix}$$

Since $\mathsf{E}[(x_3 - \phi_1 x_2 + \phi_2 x_1)x_k] = \mathsf{E}(w_3 x_k) = 0$ for $k = 1, 2$, we have $\phi_{21} = \phi_1, \phi_{22} = \phi_2$. We can also verify $\phi_{n1} = \phi_1, \phi_{n2} = \phi_2$. Therefore

$$x_{n+1}^{n} = \phi_1 x_n + \phi_2 x_{n-1} \text{ for } n \geqslant 2$$

If the series is a causal AR(p) process then for $n \geqslant p$ we have

$$x_{n+1}^{n} = \phi_1 x_n + \phi_2 x_{n-1} + \ldots + \phi_p x_{n-p+1}$$

where the justification is the same as the above example.

[3] Bring on formula sheet.

The Levinson-Durbin Algorithm

Inverting Γ is computationally expensive for large n. We can use the Levinson-Durbin Algorithm which is an iterative approach for computing this value.

$$\phi_{00} = 0, R_1^0 = \gamma(0)$$

For $n \geqslant 1$,

$$\phi_{nn} = \frac{\rho(n) - \sum_{k=1}^{n-1} \phi_{n-1,k}\rho(n-k)}{1 - \sum_{k=1}^{n-1} \phi_{n-1,k}\rho(k)}$$

and $P_{n+1}^n = P_n^{n-1}(1 - \phi_{nn}^2)$. In general the standard error of the one step ahead forecast is the square root of

$$P_{n+1}^n = \gamma(0) \prod_{j=1}^{n} (1 - \phi_{jj}^2)$$

m Step Ahead Prediction

Definition 6.0.2. The m **step ahead prediction** where $m \geqslant 1$ is

$$x_{n+m}^n = \phi_{n1}^{(m)} x_n + \phi_{n2}^{(m)} x_{n-1} + \cdots + \phi_{nn}^{(m)} x_1 = \phi_n^{(m)'} \cdot x$$

all results for this are very similar to the one step ahead case.

Innovations Algorithm

For a time series process x_t, the **innovation** is defined as a residual for the one step ahead estimator:

$$x_t - x_t^{t-1}, \text{ for } t = 1, 2, \ldots n$$

For the MA(n) process, $x_t = \sum_{j=1}^{n} \theta_j w_{n-j}$ where $w_t \sim \text{wn}(0, \sigma_w^2)$. The one step ahead predictors x_{t+1}^t and their mean squared errors P_{t+1}^t can be calculated iteratively as

$$x_1^0 = 0, P_1^0 = \gamma(0)$$

$$x_{t+1}^t = \sum_{j=1}^{t} \theta_{tj}(x_{t+1-j} - x_{t+1-j}^{t-j})$$

$$P_{t+1}^t = \gamma(0) - \sum_{j=0}^{t-1} \theta_{t,t-j}^2 P_{j+1}^j$$

7. ARMA Forecasting

Assume x_t is a causal, invertible ARMA(p, q) process

$$\Phi_p(B)x_t = \Theta_q(B)w_t$$

where $E(x_t) = 0$, since we may replace $x_t' = x_t - \mu$. There are two types of forecasts

1. $x^n_{n+m} = E(x_{n+m} \mid x_n, x_{n-1}, \ldots, x_1)$, the minimum mean square predictor based on $x_1, \ldots, x_n$.
2. $\tilde{x}_{n+m} = E(x_{n+m} \mid x_n, x_{n-1}, \ldots, x_1, x_0, \ldots)$, the x_{n+m} predictor based on infinite past data.

These are generally not the same, and for large amounts of data, $\tilde{x}_{n+m}$ will provide a good approximation. To see this, write x_{n+m} in the causal form

$$x_{n+m} = \sum_{j=0}^{\infty} \psi_j w_{n+m-j}$$

Taking conditional expectations, and $w_t = 0$ when $t > n$,

$$\tilde{x}_{n+m} = \sum_{j=m}^{\infty} \psi_j w_{n+m-j}$$

The residual satisfies

$$x_{n+m} - \tilde{x}_{n+m} = \sum_{j=0}^{m-1} \psi_j w_{n+m-j} \qquad \Longrightarrow \qquad P^n_{n+m} = E(x_{n+m} - \tilde{x}_{n+m})^2 = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2$$

The covariance satisfies

$$E[(x_{n+m} - \tilde{x}_{n+m})(x_{n+m+h} - \tilde{x}_{n+m+h})]^2 = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j \psi_{j+h}$$

As $m \to \infty$, the mean square prediction satisfies

$$P^n_{n+m} \to \sigma_w^2 \sum_{j=0}^{\infty} \psi_j^2 = \gamma_x(0) = \sigma_x^2$$

From the model in its invertible form

$$w_{n+m} = \sum_{j=0}^{\infty} \pi_j x_{n+m-j}$$

We have

$$\tilde{x}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{x}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j x_{n+m-j}$$

which may be calculated recursively in m.

Definition 7.0.1. The **truncated** predictor is written as

$$\tilde{x}^n_{n+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{x}^n_{n+m-j} - \sum_{j=m}^{n+m-1} \pi_j x_{n+m-j}$$

and may also be calculated recursively in m.

Definition 7.0.2. For an ARMA(p,q) model the truncated predictor is written

$$\tilde{x}^n_{n+m} = \phi_1 \tilde{x}^n_{n+m-1} + \ldots + \phi_p \tilde{x}^n_{n+m-p} + \theta_1 \tilde{w}^n_{n+m-1} + \ldots + \theta_q \tilde{w}^n_{n+m-q}$$

where we consider

- $\tilde{x}^n_t = x_t$ for $1 \leqslant t \leqslant n$
- $\tilde{x}^n_t = 0$ for $t \leqslant 0$
- $\tilde{w}^n_t = 0$ for $t \leqslant 0, \quad t > n$
- $w^n_t = \Phi_p(B)\tilde{x}^n_t - \theta_1 \tilde{w}^n_{t-1} - \ldots - \theta_q \tilde{w}^n_{t-q}$ for $1 \leqslant t \leqslant n$.

Example 19 (ARMA(1,1) predictor)**.** Consider $x_{n+1} = \phi x_n + w_{n+1} + \theta w_n$. Based on the truncated predictor,

$$\tilde{x}^n_{n+1} = \phi x_n + \theta w_n$$

and for $m \geqslant 2$, $\tilde{x}^n_{n+m} = \phi x_n$. This may be calculated recursively, by setting $\tilde{w}^n_0 = 0, x_0 = 0$ and

$$\tilde{w}^n_t = x_t - \phi x_{t-1} - \theta \tilde{w}^n_{t-1}$$

The approximate forecast variance becomes

$$P^n_{n+m} = \sigma^2_w \left(1 + \frac{(\phi+\theta)^2(1-\phi^{2(m-1)})}{1-\phi^2}\right)$$

8. Estimation

Assume n observations $x_1, \ldots, x_n$ from a causal, invertible, Gaussian ARMA(p,q) process

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}$$

where the parameters p, q are known. Later we discuss how they are determined, which is typically through fitting different orders which minimize AIC, BIC.

Consider the AR(p) model. Multiplying by x_{t-h}, we get the $p+1$ Yule-Walker equations

$$\sigma^2_w = \gamma(0) - \phi_1\gamma(1) - \ldots - \phi_p\gamma(p)$$

$$\gamma(h) = \phi_1\gamma(h-1) + \ldots + \phi_p\gamma(h-p)$$

In matrix notation, this is written

$$\Gamma_p \phi = \gamma_p \quad \sigma^2_w = \gamma(0) - \phi \cdot \gamma_p$$

Using the method of moments we replace $\gamma(h)$ by $\hat{\gamma}(h)$ to get the **Yule-Walker estimators**.

$$\hat{\phi} = \hat{R}_p^{-1}\hat{\rho}_p \quad \hat{\sigma}^2_w = \hat{\gamma}(0)[1 - \hat{\rho}_p \hat{R}_p^{-1}\hat{\rho}_p]$$

Proposition 4. The asymptotic behaviour of the Yule-Walker estimators in the case of causal AR(p) processes is

$$\sqrt{n}(\hat{\phi} - \phi) \to^d N(0, \sigma_w^2 \Gamma_p^{-1}) \quad \text{and} \quad \hat{\sigma}_w^2 \to^p \sigma_w^2$$

Proposition 5. The asymptotic behaviour of the partial autocorrelation satisfies

$$\sqrt{n}\hat{\phi}_{hh} \to N(0, 1)$$

The Levinson-Durbin algorithm can be used to calculate $\hat{\phi}$ without inverting $\hat{\Gamma}_p$ by replacing $\gamma(h)$ with $\hat{\gamma}(h)$. We iteratively calculate $\hat{\phi}$.

Example 20. Suppose $\hat{\gamma}(0) = 8.903$ with $\hat{\rho}(1) = 0.849$ and $\hat{\rho}(2) = 0.519$. Then

$$\hat{\phi} = \begin{bmatrix} 1 & 0.849 \\ 0.849 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0.849 \\ 0.519 \end{bmatrix} = \begin{bmatrix} 1.463 \\ -0.723 \end{bmatrix}$$

Using proposition 3, the covariance matrix of $\hat{\phi}$ is

$$\frac{1}{144} \frac{1.187}{8.903} \begin{bmatrix} 1 & 0.849 \\ 0.849 & 1 \end{bmatrix}^{-1}$$

Method of Moments

Example 21 (MA(1))**.** For the MA(1) process $x_t = w_t + \theta w_{t-1}$, we write

$$x_t = \sum_{j=1}^{\infty} (-\theta)^j x_{t-j} + w_t$$

Then $\hat{\rho}(1) = \hat{\theta}/(1 + \hat{\theta}^2)$, and we estimate $\hat{\theta}$ by solving the above equation. When $|\hat{\rho}(1)| < \frac{1}{2}$,

$$\hat{\theta} = \frac{1 - \sqrt{1 - 4\hat{\rho}(1)^2}}{2\hat{\rho}(1)}$$

Asymptotically,

$$\hat{\theta} \sim \text{AN}\left(\theta, \frac{1 + \theta^2 + 4\theta^4 + \theta^6 + \theta^8}{n(1 - \theta^2)^2}\right)$$

Maximum Likelihood

Consider the causal AR(1) model $x_t = \mu + \phi(x_{t-1} - \mu) + w_t$. Given data $x_1, \ldots, x_n$, then

$$\begin{aligned} L(\mu, \phi, \sigma_w^2) &= f(x_1, \ldots, x_n \mid \mu, \phi, \sigma_w^2) \\ &= f(x_1) f(x_2 \mid x_1) \cdots f(x_n \mid x_{n-1}) \\ &= f(x_1) \prod_{t=2}^{n} f_w((x_t - \mu) - \phi(x_{t-1} - \mu)) \end{aligned}$$

Where f_w is the density of w_t, so $f_w(x_t \mid x_{t-1})$ is a normal density. Expanding this, we see

$$= (2\pi\sigma_w^2)^{-n/2} (1 - \phi^2)^{-1/2} \exp\left(-\frac{S(\phi, \mu)}{2\sigma_2^2}\right)$$

Where S is the sum of squares, or the square prediction error

$$S(\mu, \phi) = (1 - \phi^2)(x_1 - \mu^2) + \sum_{t=2}^{n} [(x_t - \mu) - \phi(x_{t-1} - \mu)]^2$$

$S(\mu, \phi)$ is the **unconditional sum of squares**, and the **unconditional least squares** estimate is obtained by minimizing S. From this, the MLE of σ_w^2 is given by

$$\hat{\sigma}_w^2 = S(\hat{\mu}, \hat{\phi})/n$$

The **conditional likelihood** is taken by conditioning on the initial observation

$$L(\mu, \phi, \sigma_w^2 \mid x_1) = (2\pi\sigma_w^2)^{-(n-1)/2}(1 - \phi^2)^{-1/2} \exp\left(-\frac{S(\phi, \mu)}{2\sigma_2^2}\right)$$

and $\hat{\sigma}_w^2 = S(\hat{\mu}, \hat{\phi})/(n-1)$.

For general $AR(p)$ models the same process for maximum likelihood estimates is followed. For general ARMA models, the likelihood is difficult to derive explicitly, and is typically written as a function of the innovation $x_t - x_t^{t-1}$. A common numerical algorithm for minimizing $S(\mu, \vec{\phi}, \vec{\theta})$ in an $ARMA(p, q)$ model is with the Newton-Raphson algorithm.

Asymptotics of some distributions

SLIDE 46, 47 MODULE 5 ON EXAM

ARIMA Models

Definition 8.0.1. A time series with zero mean x_t is called **Autoregressive-Integrated-Moving average** of order (p, d, q) denoted $ARIMA(p, d, q)$ if the d-th difference of x_t is an ARMA (p, q) process. That is x_t is $ARIMA(p, d, q)$ if

$$\Phi_p(B)\nabla^d x_t = \delta + \Theta_q(B)w_t$$

where $\delta = \mu(1 - \phi_1 - \cdots - \phi_p)$, Φ_p, Θ_q, and $\nabla^d = (1 - B)^d$ as before.

Example 22.

- $ARIMA(1, 1, 0) : x_t = \phi x_{t-1} + x_{t-1} - \phi x_{t-2} + w_t = (1 - \phi B)(1 - B)x_t = w_t$
- $ARIMA(1, 1, 1) : (1 - \phi B)(1 - B)x_t = (1 + \theta B)w_t$
- $ARIMA(1, 2, 2) : (1 - \phi B)(1 - B)^2 x_t = (1 + \theta_1 B + \theta_2 B^2)w_t$

There are two main steps to verify models after applying transformations.

- **Goodness of fit:** t-tests, AIC, BIC, SIC, likelihood ratio tests for adding and removing parameters to the model.
- **Residuals:** We assume the series is stationary with Gaussian white noise innovations. Residuals should look like white noise series.
- If the appropriate ARMA model is chosen there will be theoretically zero autocorrelation in the errors.

- To check the frequency of ARMA fitted model, we can use autocorrelation function (ACF/PACF) of the innovations, or standardized innovations:

  $$e_t = \frac{x_t - \hat{x}_t^{t-1}}{\sqrt{\hat{P}^{t-1}}}$$

  A good check on correlation structure is to plot sample correlations and ensure they do not fall outside of $\pm 2/\sqrt{n}$.

In the (e_t) series,

$$H_0 : \rho(h) = 0, \quad H_a : \rho(h) \neq 0$$

the null hypothesis is that $\rho(h) = 0$, for all $h = 1, \dots, m$. Typically $m \approx \sqrt{n}$. The alternative hypothesis is $\rho(h) \neq 0$. Recall

$$\hat{\rho}(h) = \frac{\sum_{t=h+1}^{n} e_t e_{t-h}}{\sum_{t=1}^{n} e_t^2}$$

Test statistics

In order to test uncorrelatedness at individual lags $h = 1, \dots, m$, there are two *portmanteau* tests that can be used to test all autocorrelations simultaneously.

1. **Box-Pierce:**

   $$Q_m = n \sum_{h=1}^{m} \hat{\rho}(h)^2 \sim \chi^2_{m-p-q}$$

2. **Ljung-Box:** [4]

   $$\tilde{Q}_m = n(n+2) \sum_{h=1}^{m} \frac{\hat{\rho}(h)^2}{n-h} \sim \chi^2_{m-p-q}$$

Generally, if the model fits well there should be no significant patter in e_t.

9. Regression Continued

Autocorrelated errors

We discuss regression models following

$$y_t = \sum_{j=1}^{r} \beta_j z_{t,j} + x_t$$

where x_t has covariance function $\gamma_x(s,t)$. In ordinary least squares, the assumption is that x_t is gaussian white noise w_t, constant variance and independent. Here x_t becomes the error process. If it has non-constant variance, **weighted least squares** can be used. The weighted least squared estimate is used when the covariance matrix has diagonal elements non-zero, all else zero. I.e. x_t are independent but do not have equal variance.

[4] This will be on the final - know how to calculate. Examples on Module 5 slides 58-59.

Suppose the x_t are not independent (i.e. covary in the sample),

$$y = Z\beta + x$$

where Z is the matrix of input variables. We diagonalize the covariance matrix to use weighted least squares again. Letting $\Gamma = \{\gamma_x(s,t)\}$ be the covariance matrix then the transformation

$$\Gamma^{-\frac{1}{2}}y = \Gamma^{-\frac{1}{2}}Z\beta + \Gamma^{-\frac{1}{2}}x$$

gives a new equation

$$y^* = Z^*\beta + \delta$$

which is in the form of the classical linear model. That is, the new linear model has independent errors, and we can use the weighted least squares estimate.

$$\hat{\beta} = (Z^{*T}Z^*)^{-1}Z^{*T}y \implies \text{Cov}(\hat{\beta}) = \sigma^2(Z^{*T}Z^*)^{-1}$$

and in terms of the original model,

$$\hat{\beta} = (Z^T\Gamma^{-1}Z)^{-1}Z^T\Gamma^{-1}y$$

Regression in AR(p)

Consider the AR(p) case $\Phi_p(B)x_t = w_t$ and the regression model

$$y_t = Z\beta + x_t$$

Multiplying through by the characteristic polynomial, we have

$$\underbrace{\Phi_p(B)y_t}_{y_t^*} = \underbrace{\sum_{j=1}^{r}\beta_j\Phi_p(B)z_{t,j}}_{z_{t,j}^*} + \underbrace{\Phi_p(B)x_t}_{w_t}$$

Which gives us the regression model. I.e. if $p = 1$ then $z_{t,j}^* = z_{t,j} - \phi z_{t-1,j}$. Then

$$S(\phi,\beta) = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n}\left(\Phi_p(B)y_t - \sum_{j=1}^{r}\beta_j\Phi_p(B)z_{t,j}\right)$$

with ϕ, β being coefficients of the polynomial, and parameters of linear model respectively.

Regression in AR(p,q)

If we have that

$$\Phi_p(B)x_t = \Theta_q(B)w_t$$

then setting $\Pi(B) = \Phi_p(B)/\Theta_q(B)$ if Θ_q has appropriate roots, we get the same expression

$$w_t = \Pi(B)x_t$$

and minimize the same sum, with parameters for Θ_q

$$S(\phi,\theta,\beta) = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n}\left(\Pi(B)y_t - \sum_{j=1}^{r}\beta_j\Pi(B)z_{t,j}\right)$$

Identification

We do not actually know the behaviour of x_t before we run a regression.

1. Run an ordinary regression of $y_t \sim z_{t,1}, \dots, z_{t,r}$
2. Identify arma models in the residuals $\hat{x}_t = y_t - Z\hat{\beta}$
3. Run weighted least squares or MLE on regression model using previous discussion
4. Inspect $\hat{w}_t$ for whiteness, iterate additional steps if needed

Detecting Autocorrelation

The **Durbin-Watson** test statistic can be used to detect the presence of autocorrelation in a regression model based on assumptions that the observations come from an AR(1) model. This test is used to check $\phi = 0$.

$$H_0 : \phi = 0, x_t = w_t \quad \text{or} \quad H_a : \phi > 0, x_t = \phi x_{t-1} + w_t$$

The **Durbin-Watson** statistic [5] for time ordered residuals $e_1, \dots, e_n$ is given by

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

- If $d < d_{L,\alpha}$, reject H_0
- If $d > d_{U,\alpha}$, do not reject H_0
- If $d_{L,\alpha} < d < d_{U,\alpha}$, inconclusive

Testing the hypothesis

$$H_0 : \phi = 0, x_t = w_t \quad \text{or} \quad H_a : \phi < 0, x_t = \phi x_{t-1} + w_t$$

Can be done with

- If $4 - d < d_{L,\alpha}$, reject H_0
- If $4 - d > d_{U,\alpha}$, do not reject H_0
- If $d_{L,\alpha} < 4 - d < d_{U,\alpha}$, inconclusive

The above conditions for positive or negative correlation can be combined to test $H_a : \phi \neq 0$.

[5] Will appear on final.

Seasonal ARIMA Models

Definition 9.0.1. The **seasonal ARIMA** of order $(p, d, q) \times (P, D, Q)_s$ where s is the number of seasons is written

$$\Phi_p(B)\boldsymbol{\Phi}_P(B^s)(1-B)^d(1-B^s)^D x_t = \delta + \Theta_p(B)\boldsymbol{\Theta}_Q(B^s)w_t$$

The idea is we apply an additional polynomial in B^s in order to capture a seasonal component in the data at lag s. We assume $\boldsymbol{\Phi}_P, \boldsymbol{\Theta}_Q$ have no common roots. (P, D, Q) are the order of the **seasonal autoregressive, seasonal differencing** , **seasonal moving average** models respectively. (p, d, q) are called the non-seasonal orders. We write

$$\nabla_s^D = (1-B^s)^D$$

for simplicity. The **pure seasonal autoregressive model** is denoted

$$\boldsymbol{\Phi}_P(B^s)x_t = \boldsymbol{\Theta}_q(B^s)w_t$$

Similar causality and invertibility conditions hold for these polynomials.

Example 23. [6] For the first order seasonal $s = 12$ moving average model $x_t = w_t + \Theta w_{t-12}$ we have

$$\gamma(0) = (1+\Theta^2)\sigma_w^2, \quad \gamma(\pm 12) = \Theta\sigma_w^2, \quad 0 \text{ otherwise}$$

Example 24. [7] For the first order seasonal $s = 12$ moving autoregressive model $x_t = \Phi x_{t-12} + w_t$ we have

$$\gamma(0) = \sigma_w^2/(1-\Phi^2), \quad \gamma(\pm 12h) = \Phi^h\sigma_w^2/(1-\Phi^2), \quad 0 \text{ otherwise}$$

10. Additional Topics

Mainly extra[8] topics, spectral analysis, fractional differencing and long memory, volatility.

Spectral Analysis

Many time series show complex periodic behaviour. Spectral analysis explains the underlying periodicities, where we decompose a stationary series into sine and cosine waves with uncorrelated coefficients.

Definition 10.0.1. The **spectral density** is a *frequency domain* representation of a time series that is directly related to the autocovariance *time domain* representation: discrete Fourier transform.

Frequency domain approach considers regression on sinusoids, whereas time domain considers regression on past values.

- **Regression:** $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots$
- **Spectral:** $x_t = \sum_{\omega\in\mathbb{N}} A_{\omega,1}\cos(2\pi\omega t) + A_{\omega,2}\sin(2\pi\omega t)$

[6] This will appear on the final as a MCQ.

[7] This will appear on the final as a MCQ.

[8] There will only by 2 MCQ on exam based on this topic.

Consider the periodic process

$$x_t = A\cos(2\pi\omega t + \phi) \text{ for } t = 0, \pm1, \pm2, \cdots$$

where T is the length of one cycle, $\omega = 1/T$ is the frequency, A is the amplitude, ϕ is the phase. We may write

$$x_t = \beta_1\cos(2\pi\omega t) + \beta_2\sin(2\pi\omega t)$$

where $\beta_1 = A\cos(\phi)$, $\beta_2 = -\sin(\phi)$. Therefore $A = \sqrt{\beta_1^2 + \beta_2^2}$ and $\phi = \tan^{-1}(-\beta_2/\beta_1)$.

Proposition 6. A, ϕ are independent random variables where

$$A \sim \chi_2^2,\ \phi \sim U(-\pi, \pi) \iff \beta_1, \beta_2 \sim N(0,1)$$

The autocovariance of uncorrelated sinusoids is the sum of their autocovariances. Therefore for

$$x_t = \sum_{j=1}^{k} A_j\cos(2\pi\omega_j t) + B_j\sin(2\pi\omega_j t)$$

we have $\gamma(h) = \sum_{j=1}^{k}\sigma_j^2\cos(2\pi\omega_j h)$.

Definition 10.0.2. For x_t with autocovariance γ satisfying $\sum_{h=-\infty}^{\infty}|\gamma(h)| < \infty$ we define its **spectral density** [9] as

$$f(\omega) = \sum_{h=-\infty}^{\infty}\gamma(h)e^{-2\pi i\omega h}$$

and the **normalized spectral density** as

$$f^*(\omega) = \sum_{h=-\infty}^{\infty}\rho(h)e^{-2\pi i\omega h}$$

Note that $f(\omega) > 0$, f is periodic, f has period 1: we may restrict the domain of f to $-\frac{1}{2} \leqslant \omega \leqslant \frac{1}{2}$. Inverting the fourier transform gives the autocovariance function

$$\gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i\omega h}f(\omega)d\omega$$

We may split the above summation definition of f as [10]

$$\begin{aligned} f(\omega) &= \gamma(0) + \sum_{h=1}^{\infty}\gamma(h)e^{-2\pi i\omega h} + \sum_{h=-\infty}^{-1}\gamma(h)e^{-2\pi i\omega h} \\ &= \gamma(0) + 2\sum_{h=1}^{\infty}\gamma(h)\cos(2\pi\omega h) \end{aligned}$$

The corresponding normalized spectrum is

$$f^*(\omega) = \rho(0) + 2\sum_{h=1}^{\infty}\rho(h)\cos(2\pi\omega h)$$

[9] Will show up on exam

[10] Will show up on exam, might be asked to show spectral density of MA process

Example 25. We compute the spectral density of an AR(1) process.

$$\begin{aligned} f(\omega) &= \gamma(0) + \sum_{h=1}^{\infty} \gamma(h)e^{-2\pi i\omega h} + \sum_{h=1}^{\infty} \gamma(h)e^{2\pi i\omega h} \\ &= \frac{\sigma^2}{1-\phi^2}\left(\gamma(0) + \sum_{h=1}^{\infty}(\phi e^{2\pi i\omega})^h + \sum_{h=1}^{\infty}(\phi e^{-2\pi i\omega})^h\right) \\ &= \frac{\sigma^2}{1-\phi^2}\left(\frac{1-\phi e^{-2\pi i\omega}\phi e^{2\pi i\omega}}{(1-\phi e^{-2\pi i\omega})(1-\phi e^{2\pi i\omega})}\right) \\ &= \frac{\sigma^2}{1-2\phi\cos(2\pi\omega)+\phi^2} \end{aligned}$$

Long Memory and Fractional Differencing

The ARMA(p, q) is often called a short memory process since the coefficients in Wold representation $x_t = \sum \psi_j w_{t-j}$ decay exponentially. The result implies $\rho(h) \to 0$ exponentially fast as $h \to \infty$.

When the ACF of x_t decays slowly, we may difference the series until it seems stationary

$$\nabla x_t = (1-B)x_t$$

However, this may yield an over-differencing of the model by too strongly modifying it. Such a process is a **long memory** time series. The basic long memory series is a special case of the **autoregressive fractionally integrated model ARIFMA**$(0, d, 0)$ given by

$$(1-B)^d x_t = w_t$$

where $0 < d < 0.5$. When d is not an integer, the d-th difference

$$\nabla^d x_t = (1-B)^d x_t = \left(1 - dB + \frac{d(d-1)}{2!}B^2 - \frac{d(d-1)(d-2)}{3!}B^3 + \cdots\right)x_t$$

Definition 10.0.3. The **auto-regressive fractionally integrated ARFIMA**(p, d, q) is

$$\left(1 - \sum_{i=1}^{p}\phi_i B^i\right)(1-B)^d x_t = \left(1 + \sum_{i=1}^{q}\theta_i B^i\right)w_t$$

where d is the fractional difference, takes a value between $0, 1$ possibly up to $2+$ in more extreme cases.