AnSeDa - Résumé Olivier D'Ancona

Espérance

Soit X une variable aléatoire à valeurs dans $\mathbb{R}$, on définit l'espérance de X par :

Cas discret : $\mathrm{E}[X] = \sum_{i=1}^{\infty} x_i \cdot p_i$

Cas continu : $\mathrm{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$ où $f(x)$ est la densité de probabilité de X.

Propriétés

- $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$
- $\mathrm{E}[aX] = a\mathrm{E}[X]$
- $\mathrm{E}[X^2] = \mathrm{Var}[X] + \mathrm{E}[X]^2$
- $\mathrm{E}[XY] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$ if X and Y are independent.
- $\mathrm{E}[c] = c$ if c is a constant.

Variance

Soit X une variable aléatoire à valeurs dans $\mathbb{R}$. Cas discret // continu :

$$\mathrm{Var}[X] = \sum_{i=1}^{\infty} (x_i - \mathrm{E}[X])^2 \cdot p_i$$
$$= \mathrm{E}[(X_i - \mu)^2] = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$\mathrm{Var}[X] = \int_{-\infty}^{\infty} (x - \mathrm{E}[X])^2 \cdot f(x)\, dx$$

$\mathrm{Var}[X] = \sigma^2$

Propriétés

- $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$
- $\mathrm{Var}[aX] = a^2 \mathrm{Var}[X]$
- $\mathrm{Var}[XY] = \mathrm{Var}[X] \cdot \mathrm{Var}[Y] + \mathrm{Var}[X] \cdot \mathrm{E}[Y]^2 + \mathrm{Var}[Y] \cdot \mathrm{E}[X]^2$ if X and Y are independent.
- $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$
- $\mathrm{Var}[c] = 0$ if c is a constant.

Covariance

Soit X et Y deux variables aléatoires, la covariance entre X et Y est définie par :

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] \end{aligned}$$

Propriétés

- $\mathrm{Cov}(X,c) = 0$ si c est une constante
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(aX+b, cY+d) = ac \cdot \mathrm{Cov}(X,Y)$
- $\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$
- $\mathrm{Cov}(X,Y) = 0$ Si X et Y sont indépendants

Corrélation

fonction de corrélation

$\mathrm{E}[XY]$

coefficient de corrélation

Soit X et Y deux variables aléatoires, le coefficient de corrélation entre X et Y est défini par :

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Propriétés

- $\rho_{XY} = \rho_{YX}$
- $-1 \le \rho_{XY} \le 1$
- $\rho_{XY} = 1$ ou -1 indique une corrélation linéaire parfaite
- $\rho_{XY} = 0$ si X et Y sont indépendants linéairement
- Invariance des changements linéaires : $\rho_{aX+b,cY+d} = \rho_{XY}$ pour tout a,b,c,d réels avec $ac \neq 0$

Autocovariance

The autocovariance function measures the linear dependency between different values of a stochastic process at different times. For a stationary time series $\{X_t\}$, the autocovariance function at lag j is given by :

$$\begin{aligned} \gamma(j) &= \mathrm{Cov}[X_t, X_{t+j}] \\ &= \mathrm{E}[(X_t - \mu)(X_{t+j} - \mu)] \end{aligned}$$

It is only dependent on the lag j and not on the specific time t, reflecting the stationarity of the process. For lags $j > 0$, the autocovariance is also known as the auto-correlation function (ACF) and is given by :

$$\rho(j) = \frac{\gamma(j)}{\gamma(0)}$$

Loi de probabilité

Loi	Param	$\mathrm{E}[X]$	$\mathrm{Var}(X)$	Support
Bernoulli	p	p	$p(1-p)$	$\{0,1\}$
Binomiale	n,p	np	$np(1-p)$	$\{0,\ldots,n\}$
Uniforme	a,b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$[a,b]$
Normale	μ,σ	μ	σ^2	$\mathbb{R}$
Exponentielle	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$[0,\infty)$
Poisson	λ	λ	λ	$\{0,1,2,\ldots\}$
Géométrique	p	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\{0,1,2,\ldots\}$

Likelihood

Soit X une variable aléatoire à valeurs dans $\mathbb{R}$, on définit la likelihood de θ par :

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

où $f(x)$ est la densité de probabilité de X. Souvent, on utilise le logarithme de la likelihood :

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

Stationarité

La stationnarité d'un processus aléatoire décrit comment ses propriétés statistiques restent constantes au fil du temps. Pour qu'une série temporelle soit stationnaire, elle doit présenter quatre propriétés constantes dans le temps :

1. Moyenne constante
2. Variance constante
3. Structure d'autocorrélation constante
4. Aucun composant périodique (saisonnalité)

L'autocorrélation signifie que la mesure actuelle de la série temporelle est corrélée à une mesure passée.

Stationarité au sens strict

Un processus $\{X_i\}$ est strictement stationnaire si :

- La distribution de X_i est la même que celle de X_j pour $i \neq j$.
- Les implications sont que $\mathrm{E}[X_i]$ et $\mathrm{Var}(X_i)$ sont constants et ne dépendent pas de i.
- La distribution conjointe de (X_i, X_j) est la même que celle de (X_{i+k}, X_{j+k}).
- L'autocovariance $\mathrm{Cov}(X_i, X_{i+j})$ ne dépend que de j.

Weak Stationarity

Un processus $\{X_i\}$ est largement stationnaire si :

- $\mathrm{E}[X_i]$ et $\mathrm{Var}(X_i)$ sont constants et ne dépendent pas de i.
- L'autocovariance $\mathrm{Cov}(X_i, X_{i+j}) = \mathrm{E}[X_i X_{i+j}] - \mathrm{E}[X_i]\mathrm{E}[X_{i+j}]$ ne dépend que de j.

Modélisation

Analyse exploratoire:

Examiner les données pour comprendre les caractéristiques clés, telles que les tendances, la saisonnalité, et les anomalies.

Vérification de la stationnarité:

Tester la stationnarité de la série temporelle à l'aide de tests statistiques, tels que le test de Dickey-Fuller augmenté.

Transformation et différenciation:

Appliquer des transformations, telles que la différenciation ou le logarithme, pour stabiliser la variance et atteindre la stationnarité.

Identification du modèle:

Utiliser l'ACF et la PACF pour identifier les structures potentielles de modèle AR, MA, ou ARMA. Le choix du modèle doit être basé sur les caractéristiques des données.

Estimation et ajustement du modèle:

Estimer les paramètres du modèle choisi et ajuster le modèle aux données.

Diagnostic du modèle:

Analyser les résidus pour vérifier l'absence de structure autocorrélative et la normalité. Utiliser des graphiques tels que les QQ plots et les résidus ACF/PACF.

Validation et amélioration du modèle:

Valider le modèle à l'aide de données de test et ajuster le modèle au besoin en ajoutant ou en modifiant des termes.

Interprétation et utilisation:

Interpréter les résultats en contexte et les utiliser pour des prévisions ou des analyses plus poussées.

Régression Linéaire

Soit un tableau de données :
$x = \text{Soap(g)}$, $y = \text{Height(cm)}$, $x \cdot y$, x^2

$$X = [1, Soap]$$

$$X^T X = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = \begin{bmatrix} 7 & 38.5 \\ 38.5 & 218.95 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \begin{bmatrix} 348 \\ 1975 \end{bmatrix}$$

$$\hat{\theta} = (X^T X)^{-1} X^T y = \begin{bmatrix} -2.67 \\ 9.51 \end{bmatrix}$$

Inverse d'une matrice 2x2 :

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Normale

1. $P(X \leq x) = \Phi(x) = \int_{-\infty}^{x} N(0,1)dx$
2. $\Phi(x) = 1 - \Phi(-x)$
3. $F_x(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ est une loi normale centrée réduite

White Noise

A stationary time series or a stationary random process with zero autocorrelation.

Backshift Operator

$BX_i = X_{i-1}$ has the effect of shifting the data back one period. For instance : $B^2 X_i = B(BX_i) = B(X_{i-1}) = X_{i-2}$.

différence d'ordre d

$\Delta^d X_i = (1-B)^d X_i = X_i - X_{i-d}$

MA

Let $\{X_t\}$ be a time series that follows an MA(q) process, defined by the equation :

$$X_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

where

- μ is the mean of the series,
- ϵ_t are white noise error terms with zero mean and constant variance σ^2,
- $\theta_1, \ldots, \theta_q$ are the parameters of the MA model.

The autocovariance function of the MA(q) process is given by :

$$\gamma(h) = \begin{cases} \sigma^2 \sum_{i=0}^{q-j} \theta_i \theta_{i+j} & \text{for } h = 1, 2, \ldots, q \\ 0 & \text{for } h > q \end{cases}$$

Autoregressive Processes

An autoregressive model of order p, denoted as AR(p), posits that the current value of a series X_i can be expressed as a linear combination of its p past values and a white noise term Z_i with zero mean and variance σ^2. The AR(p) model is represented by :

$$X_i = \lambda_1 X_{i-1} + \lambda_2 X_{i-2} + \ldots + \lambda_p X_{i-p} + Z_i$$

Using the backshift operator B, where $B^k X_i = X_{i-k}$, the model can be succinctly written as :

$$\Phi(B) X_i = Z_i$$

with the autoregressive operator $\Phi(B)$ defined as :

$$\Phi(B) = 1 - \lambda_1 B - \lambda_2 B^2 - \ldots - \lambda_p B^p$$

The AR(p) model captures the dependence of X_i on its own past values, and it is used for forecasting future values in time series analysis.

Ergodicity

Ergodicity for a stationary stochastic process $X = (X_n : n \in \mathbb{Z})$ with finite index set $-m \leq n \leq m$ occurs when :

$$\lim_{m \to \infty} \hat{\mu}_X(m) = E[X_n]$$

where $\hat{\mu}_X(m)$ is the time average defined as :

$$\hat{\mu}_X(m) = \frac{1}{1+2m} \sum_{i=-m}^{m} X_i$$

This signifies that long-term time averages equate to expected values as the sample size increases indefinitely.

ARMA(p,q) Processes

An ARMA(p,q) model combines autoregressive (AR) and moving average (MA) models to describe a time series X_n. The current value of the series is explained by a linear function of p past values (AR part) and q past white noise terms (MA part), represented as :

$$\begin{aligned} X_n &= \lambda_1 X_{n-1} + \lambda_2 X_{n-2} + \ldots + \lambda_p X_{n-p} + \\ &\quad Z_n + \theta_1 Z_{n-1} + \theta_2 Z_{n-2} + \ldots + \theta_q Z_{n-q} \\ &= \sum_{i=1}^{p} \lambda_i X_{n-i} + \sum_{j=1}^{q} \theta_j Z_{n-j} \end{aligned}$$

Utilizing the backward shift operator B, the process can be written as :

$$\Phi(B) X_i = \Psi(B) Z_i$$

where the AR operator $\Phi(B)$ and the MA operator $\Psi(B)$ are defined by :

$$\Phi(B) = 1 - \lambda_1 B - \lambda_2 B^2 - \ldots - \lambda_p B^p$$

$$\Psi(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$$

The ARMA model captures the dynamics of a time series by accounting for both the momentum (AR part) and shocks (MA part).

ARIMA Models

ARIMA (Autoregressive Integrated Moving Average) models extend ARMA models by incorporating differencing to handle non-stationary data. An ARIMA(p, d, q) model has three components :

- AR : p = order of the autoregressive part.
- I : d = degree of differencing.
- MA : q = order of the moving average part.

A time series X_i is ARIMA(p, d, q) if $Y_i = (1-B)^d X_i$ follows an ARMA(p, q) model. The steps for ARIMA model identification are :

1. Choose d based on time series trends and stationarity, typically d = 1 or 2.
2. Determine p and q by plotting ACF and PACF of the d^{th} order differenced data.
3. Estimate the parameters, typically using statistical software.
4. Perform residuals diagnostics to check for remaining trends or correlations, adjusting if necessary.

ARIMA models are powerful tools for forecasting and analyzing time series that exhibit non-stationary behavior.

Invertibility of MA Models

Invertibility of MA models allows us to express past white noise terms based on the observed time series.

Conditions

MA(1) : $-1 < \theta < 1$
MA(2) : $-1 < \theta_2 < 1$, $\theta_1 + \theta_2 > -1$ and $\theta_1 - \theta_2 < 1$

General Case

MA(q) model, represented by

$$X_i = (1 + \theta_1 B + \ldots + \theta_q B^q) Z_i$$

is invertible if the polynomial

$$\Phi(B) = 1 + \theta_1 B + \ldots + \theta_q B^q$$

has roots outside the unit circle in the complex plane. The innovations (or white noise) Z_i can be recovered by inverting the operator : $Z_i = \Phi(B)^{-1} X_i$. Invertibility ensures that the estimated white noise series is unique.

MA and AR Link

Duality between AR and MA processes :

MA, AR link

- $MA(q)$ Invertible, $MA(q) \to AR(\infty)$
- $Ar(p)$ Stationary, $AR(p) \to MA(\infty)$

Observation

For an AR(1) model :

- $\lambda_1 = 0$: X_i is a white noise
- $\lambda_1 = 1$: X_i is a random walk
- $\lambda_1 = 1$ and $c \neq 0$: X_i is a random walk with drift
- $\lambda_1 < 1$: X_i tends to oscillate around the mean

PACF

The Partial Autocorrelation Function (PACF) measures the correlation between observations in a time series separated by k periods, removing the effect of intervening observations. PACF is particularly useful in identifying the order of an autoregressive (AR) process. It helps determine the number of AR terms to be included in an ARIMA model. PACF plots display the partial correlation coefficients for different lags. A significant spike at a specific lag in the PACF plot suggests the need for that number of AR terms in the model.

ACF

$$\frac{\mathrm{Cov}(X_i, X_{i+j})}{\mathrm{Cov}(X_i, X_i)}$$

PACF

$$\frac{\mathrm{Cov}(X_i, X_{i+j} | X_{i+1}, \ldots, X_{i+j-1})}{\mathrm{Cov}(X_i, X_i | X_{i+1}, \ldots, X_{i+j-1})}$$

TOPIC

0