Autocovariância

Prof. Renzo Flores-Ortiz

Covariância

- Covariância é uma medida estatística que indica **o grau em que duas variáveis variam conjuntamente**.
    - Se elas **aumentam ou diminuem juntas**, a covariância será **positiva**.
    - Se **uma aumenta enquanto a outra diminui**, a covariância será **negativa**.
    - Se **não há relação clara**, a covariância será próxima de **zero**.

Definição teórica de covariância populacional para qualquer tipo de variável aleatória — discreta ou contínua:

$$COV(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

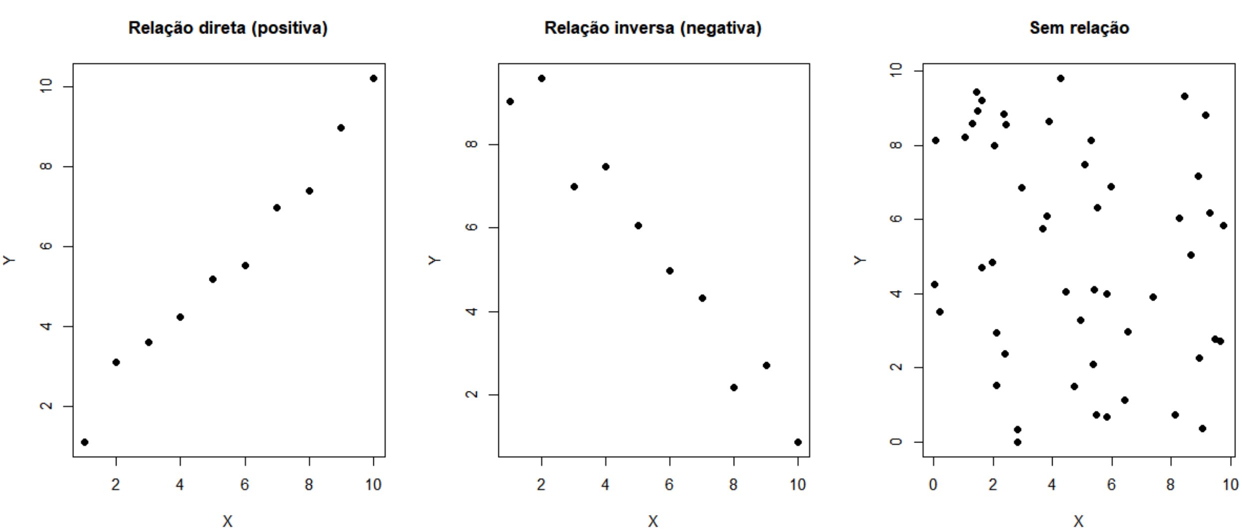
Quando estamos lidando com uma população finita e conhecida (por exemplo, uma lista completa de N pessoas, empresas, etc.), a esperança $E[\cdot]$ pode ser calculada diretamente como uma média aritmética:

$$E[(X-\mu_X)(Y-\mu_Y)] = \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu_X)(Y_i-\mu_Y) \Rightarrow COV(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu_X)(Y_i-\mu_Y)$$

Quando se têm dados amostrais, utiliza-se a **covariância amostral**:

$$COV(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

Tipos de relações entre variáveis



$$COV(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- A unidade de medida da covariância é uma combinação das unidades de x e y.
- Por exemplo, se x representa anos de estudo e y representa renda mensal em reais, a unidade de medida da covariância entre x e y é anos x reais.
- A unidade de medida da covariância não é muito intuitiva de interpretar.

Correlação

- Para **interpretar melhor a relação entre variáveis com unidades de medida diferentes**, usamos o **coeficiente de correlação de Pearson**, que é **adimensional**:

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$$

- Trata-se de uma versão padronizada da covariância, obtida ao dividir a covariância pelos desvios padrão das variáveis.
- O resultado é um número entre **-1 e 1**, que **não depende das unidades de medida originais**.
- Isso facilita muito a **comparação** e a **interpretação** da força e do sentido da relação linear entre as variáveis.

Covariância em séries temporais

Na prática com séries temporais, a covariância pode ser usada para:

1. Verificar relação entre duas séries temporais (ex: vendas vs. publicidade).
2. Calcular autocovariância (covariância da série com ela mesma em diferentes defasagens).
3. É a base para funções de autocorrelação (ACF), muito usadas em modelos ARIMA.

Autocovariância

- Em séries temporais, a autocovariância mede o grau de dependência linear entre os valores da série em dois instantes distintos no tempo.

- Definição formal: Seja $\{X_t\}$ um processo estocástico com média constante $\mu = \mathrm{E}[X_t]$, a autocovariância de ordem h (ou defasagem h) é definida por:

$$\gamma(h) = COV(X_t, X_{t+h}) = E[(X_t - \mu)(X_{t+h} - \mu)]$$

Onde

- $\mu = E[X_t]$ é a média do processo estocástico (assume-se constante no tempo, conforme requisito de estacionaridade),
- $\gamma(0)$ corresponde à variância do processo estocástico.

Função de autocovariância

- A função de autocovariância é a função que associa a cada defasagem $h \in \mathbb{Z}$ (inteiros) o valor da autocovariância correspondente. Ou seja, é a função:

$$\gamma: h \mapsto \gamma(h) = COV(X_t, X_{t+h})$$

- Essa função descreve como os valores da série se relacionam com suas versões defasadas, ao longo de diferentes valores de h. A função é fundamental para entender a estrutura de dependência temporal de um processo.

- A autocovariância depende apenas da defasagem h, não do tempo t, quando o processo é estacionário em sentido estrito ou fraco.

Autocovariância amostral

- A autocovariância amostral com defasagem h (ou lag h) estima a autocovariância populacional a partir de uma série temporal finita $\{x_1, x_2, \dots, x_T\}$. A autocovariância amostral com defasagem h é definida como:

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

Onde

- T é o número total de observações da série temporal;
- $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$ é a média amostral da série;
- $h \in \{0, 1, 2, \dots\}$ é o valor da defasagem;
- x_t é o valor observado no tempo t;
- x_{t+h} é o valor observado h períodos à frente (no futuro).

- Essa fórmula calcula o grau de dependência linear entre os valores da série no tempo t e no tempo $t+h$, considerando o deslocamento de h unidades de tempo.

Função de autocovariância amostral

- A função de autocovariância amostral associa a cada valor de defasagem $h \in \{0,1,\dots,H\}$ o valor correspondente de $\hat{\gamma}(h)$, para algum valor máximo de defasagem H escolhido. Ou seja, é a função:

$$\hat{\gamma}: h \mapsto \hat{\gamma}(h)$$

- Ela permite visualizar como a dependência entre os valores da série diminui (ou se comporta) à medida que o tempo entre as observações aumenta.

Exemplo: calculo da autocovariância amostral com defasagem $h = 1$

- Considere a série temporal de $T = 5$ observações igualmente espaçadas no tempo: $\{3, 5, 7, 6, 8\}$

- Passo 1: Calcular a média amostral: $\bar{x} = \frac{3+5+7+6+8}{5} = 5.8$

t	x_t	x_{t+1}
1	3	5
2	5	7
3	7	6
4	6	8

- Passo 2: Aplicar a fórmula da autocorrelação amostral

A fórmula da autocorrelação amostral com defasagem h é:

$$r(h) = \frac{\sum_{t=1}^{T-h}(x_t - \bar{x})(x_{t+h} - \bar{x})}{\sum_{t=1}^{T-h}(x_t - \bar{x})^2}$$

- Para *h*=1:

$$r(1) = \frac{1}{5-1}[(3-5.8)(5-5.8) + (5-5.8)(7-5.8) + (7-5.8)(6-5.8) + (6-5.8)(8-5.8)]$$

$$\hat{\gamma}(1) = \frac{1}{4}(1.96) = 0.49$$

- <u>Interpretação</u>: O valor positivo da autocovariância com defasagem 1 indica que os valores consecutivos da série tendem a variar na mesma direção, ou seja, há uma associação linear positiva entre os valores separados por um período.

Exemplo: calculo da autocovariância amostral com defasagem $h = 2$

- Considere a série temporal de $T = 5$ observações igualmente espaçadas no tempo: $\{3, 5, 7, 6, 8\}$
- Passo 1: Calcular a média amostral: $\bar{x} = \frac{3+5+7+6+8}{5} = 5.8$
- Passo 2: Aplicar a fórmula da autocovariância amostral

A fórmula da autocovariância amostral com defasagem h é:

t	x_t	x_{t+2}
1	3	7
2	5	6
3	7	8

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

- Para *h*=2:

$$\hat{\gamma}(2) = \frac{1}{5-2}[(3-5.8)(7-5.8) + (5-5.8)(6-5.8) + (7-5.8)(8-5.8)]$$

$$\hat{\gamma}(2) = \frac{1}{3}(-0.88) = -0.293$$

- Interpretação: O valor negativo da autocovariância com defasagem 2 indica que os valores separados por dois períodos tendem a variar em direções opostas, ou seja, há uma associação linear negativa entre os valores separados por dois períodos.

Exemplo: calculo da autocovariância amostral com defasagem $h = 0$

- Considere a série temporal de $T = 5$ observações igualmente espaçadas no tempo: $\{3, 5, 7, 6, 8\}$
- Passo 1: Calcular a média amostral: $\bar{x} = \frac{3+5+7+6+8}{5} = 5.8$
- Passo 2: Aplicar a fórmula da autocovariância amostral

A fórmula da autocovariância amostral com defasagem h é:

t	x_t	x_{t+0}
1	3	3
2	5	5
3	7	7
4	6	6
5	8	8

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

- Para *h*=0:

$$\hat{\gamma}(0) = \frac{1}{5-0}[(3-5.8)^2 + (5-5.8)^2 + (7-5.8)^2 + (6-5.8)^2 + (8-5.8)^2]$$

$$\hat{\gamma}(0) = \frac{1}{5}(14.8) = 2.96$$

- Interpretação: A autocovariância com defasagem $h = 0$ mede a variância amostral não corrigida da série — ou seja, ela expressa o grau de dispersão dos valores da série em torno da média. Portanto, o valor $\hat{\gamma}(0) = 2.96$ indica que, em média, os valores da série se desviam da média em um quadrado médio de 2.96 unidades.

Observação: cálculo da autocovariância amostral no R

- A função acf(x, type = "covariance"), a mais comum para calcular a autocovariância amostral, utiliza a seguinte fórmula para calcular a autocovariância amostral:

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

- Essa fórmula é diferente da que foi apresentada:

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

- Por que essa diferença?

Por que a função acf() usa T no denominador?

1. Consistência entre lags:

- Usar o mesmo denominador T para todos os lags facilita a comparação direta dos valores da autocovariância entre diferentes defasagens.
- Se cada lag tivesse um denominador diferente ($T - h$), os valores não estariam exatamente na mesma escala, o que dificultaria comparações visuais e análises em gráficos.

2. Compatibilidade com estimativas de autocorrelação:

- A autocorrelação de Pearson (também computada por acf()) é obtida dividindo a autocovariância do lag h pela autocovariância do lag 0.
- Usar o mesmo denominador T em todos os lags faz com que esses cancelamentos no cálculo da autocorrelação sejam diretos e fáceis de interpretar.

3. Eficiência computacional e estabilidade:

- Esse padrão é usado por muitos softwares estatísticos (como R, Python, MATLAB) por fornecer resultados mais estáveis numericamente, especialmente com séries longas.

4. Simplicidade de implementação:

- Um denominador fixo T é mais simples de aplicar em vetores grandes e evita a checagem de comprimento restante para cada lag.

Calcular a autocovariância amostral com T ou $T - h$ no denominador?

- Dividir por T (como acf() faz) é útil quando o foco é análise exploratória e visualização, e se deseja manter uma base comum para comparação.

- Dividir por $T - h$ é mais apropriado se seu objetivo for estimar a autocovariância amostral com precisão estatística.
  - Dividir por $T - h$ retorna estimativas mais próximas do valor populacional para cada h.
  - Dividir por T pode ser mais viesado em pequenos T.