Aula 06 - Análise multivariada

Ciência de Dados e
Inteligência Artificial
Aplicadas

Estatística Descritiva

Stefano Mozart

12/02/2025

CC BY

Sumário

- Regressão linear
- Séries temporais

Regressão linear

Regressão Linear

A regressão linear é uma técnica estatística que auxilia a identificação de padrões de associação entre variáveis amostrais, estabelecendo uma relação linear entre uma variável dependente e uma ou mais variáveis independentes:

- **Estimação dos parâmetros de associação:**
    - Quantifica e resume a relação entre duas variáveis, permitindo visualizar e descrever tendências presentes nos dados.
    - Ao representar graficamente a linha de regressão sobre os dados (dispersão), facilita a interpretação visual de como os dados se distribuem e se ajustam a uma tendência linear.
    - Ao reduzir a complexidade dos dados a uma equação simples, a regressão linear permite resumir informações essenciais de forma clara e objetiva.

- **Predição:** interpolação e/ou extrapolação dos da variável dependente;

- **Não é capaz de inferir/estimar relações de causalidade!**

Regressão Linear

- Regressão linear simples: $Y = \beta_0 + \beta_1 X + \varepsilon$
  - β_o é o intercepto
  - β_1 é o coeficiente angular e
  - ε é o erro
- Regressão linear múltivariáve: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$
  - β_o é o intercepto,
  - β_n é o coeficiente da n-ésima variável independente
  - ε é o erro

Regressão Linear

A regressão linear, independente do método aplicado na estimação dos parâmetros, tem como pressupostos de validade:

- **Linearidade:** relação estritamente linear entre as variáveis;
- **Homoscedasticidade:** a dispersão dos resíduos deve ser constante ao longo de todos os valores da variável independente.
- **Independência dos Resíduos:** os resíduos devem ser independentes entre si, não apresentando padrões ou correlações.
- **Ausência de Multicolinearidade:** em modelos com múltiplas variáveis independentes, estas não devem ser fortemente correlacionadas entre si.
- **Variação Amostral:** deve haver variabilidade nos valores da variável independente (X).

Regressão Linear

Existem várias técnicas para estimar os coeficientes de uma regressão linear simples, entre os quais:

- ❑ Método dos Mínimos Quadrados Ordinários (OLS, na sigla em inglês);
- ❑ Estimação por máximo verossimilhança (MLE)
- ❑ Gradiente decrescente:
- ❑ Equação Linear Normal;
- ❑ Decomposição QR;

Métodos Robustos

Método TELBS (Trimmed Elemental Least Binary Squares):

- Eficiente quando os dados apresentam valores muito discrepantes;
- Produz estimativas próximas aos verdadeiros valores dos coeficientes mesmo na presença de outliers;

Método LAD (Least Absolute Deviation):

- Minimiza a soma dos valores absolutos dos resíduos ao invés dos quadrados;
- Também conhecido como regressão mediana;

Regressão Quantílica:

- estima os parâmetros β minimizando uma função de perda assimétrica para um quantil específico τ (*e.g.* τ=0.5);
- Também conhecido como regressão mediana;

Mínimos Quadrados Ordinários (OLS)

O método OLS encontra os valores de β_0 e β_1 que minimizam a soma dos quadrados dos resíduos.

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

- O método OLS é o mais utilizado devido à sua simplicidade e propriedades ótimas sob os pressupostos clássicos (por exemplo, erro normalmente distribuído, homocedasticidade, independência).

Estimação por Máxima Verossimilhança (MLE)

A função de Log-verossimilhança é definida por:

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

onde os erros ε_i são assumidos como independentes e identicamente distribuídos segundo uma Normal com média zero e variância σ^2, ou seja:

$$\varepsilon_i \sim N(0, \sigma^2).$$

E para estimar β_0 e β_1, derivamos ℓ em relação a β_0 e β_1, e igualamos a zero.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\,\bar{x}. \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

M-estimador de Huber

é uma abordagem robusta para estimar os parâmetros de um modelo de regressão linear, minimizando o impacto dos outliers. Em uma regressão linear simples, o modelo é:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

- os resíduos são definidos como

$$r_i = y_i - (\beta_0 + \beta_1 x_i)$$

- Em vez de minimizar a soma dos quadrados dos resíduos (como no método dos mínimos quadrados), o M-estimador de Huber minimiza a soma de uma função de perda $\rho(r_i)$, definida por:

$$\rho(r) = \begin{cases} \frac{1}{2} r^2, & \text{se } |r| \le \delta, \\ \delta\,|r| - \frac{1}{2}\delta^2, & \text{se } |r| > \delta, \end{cases}$$

- Esse problema de minimização pode ser resolvido por vários algoritmos, inclusive Equações de Estimação ou o algoritmo Iteratively Reweighted Least Squares (IRLS)

Least Absolute Deviation (LAD)

Também conhecido como regressão L1, busca minimizar a soma dos valores absolutos dos resíduos.

- Essa abordagem torna o método mais robusto a outliers, pois os erros extremos não têm o mesmo peso exagerado que teriam na soma dos quadrados:

- O problema de LAD consiste em encontrar os coeficientes β_0 e β_1 que minimizam a soma das diferenças absolutas entre os valores observados e os valores previstos. Ou seja:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|$$

- Generalizando para um modelo linear multivariado com p preditores, a formulação é:

$$\min_{\beta} \sum_{i=1}^{n} |y_i - \mathbf{x}_i^\top \beta|$$

- Onde $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ e $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^\top$.

Regressão Quantílica

Estima os parâmetros β minimizando uma função de perda assimétrica (a "check function") para um quantil específico τ (comum τ=0.5);

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \beta_1 x_i),$$

Com

$$\rho_\tau(u) = u(\tau - I(u < 0)),$$

- onde $I(u<0)$ é a função indicadora que vale 1 se $u<0$ e 0 caso contrário;

OLS Multivariável

No cenário multivariável, OLS também é o método mais utilizado para estimar os coeficientes da regressão linear. O método consiste em minimizar a seguinte expressão quadrática:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

- Em notação matricial, se definirmos y como o vetor de respostas e X como a matriz de regressores (onde a primeira coluna é composta de 1's para o intercepto), os estimadores OLS são dados por:

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

MLE Multivariável

Quando assumimos que os erros seguem uma distribuição normal, o método de máxima verossimilhança (MLE) fornece os mesmos estimadores que o OLS para os coeficientes. Nesse caso, a log-verossimilhança dos dados é dada por:

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

- Maximizando ℓ em relação a β e σ^2 obtemos, sob os pressupostos normais, a mesma solução obtida pelo OLS.

Regressão Quantílica Multivariável

Essa abordagem estima os coeficientes para um determinado quantil da distribuição condicional de Y, minimizando uma função de perda assimétrica (*check function*), por exemplo, para a mediana (quantil 0.5).

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)$$

Onde

$$\rho_\tau(u) = u\,(\tau - I(u < 0))$$

Essa técnica é particularmente robusta contra outliers, pois a mediana (e outros quantis centrais) não são influenciados por valores extremos.

M-Estimador de Huber Multivariável

Em vez de minimizar a soma dos quadrados dos resíduos, minimiza-se uma função de perda $\rho(e_i)$ que é menos sensível a outliers.

$$\min_{\beta} \sum_{i=1}^{n} \rho\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)$$

A função de Huber é, então, definida como:

$$\rho(e) = \begin{cases} \frac{1}{2}e^2, & \text{se } |e| \leq \delta, \\ \delta\left(|e| - \frac{1}{2}\delta\right), & \text{se } |e| > \delta. \end{cases}$$

Assim como na regressão simples, utiliza-se o algoritmo Iteratively Reweighted Least Squares (IRLS) para resolver a minimização, onde a cada iteração os resíduos são ponderados de acordo com o seu tamanho.

Prática

Exercícios:

1. Selecione duas variáveis contínuas no seu dataset e calcule os parâmetros de uma regressão simples utilizando o os métodos descritos nesta aula (OLS, MLE, Huber, Quantílico).

2. Selecione três variáveis contínuas, de preferência mantendo a mesma variável dependente do exercício anterior, e calcule os parâmetros de regressão utilizando os métodos de regressão multivariada apresentados;

Séries temporais

Séries temporais

Uma série temporal é uma amostra com uma dimensão temporal, geralmente coletada em intervalos regulares.

- O objetivo da análise de séries temporais é compreender os padrões, identificar tendências, sazonalidades, ciclos e ruídos, além de possibilitar previsões futuras;

- Essas técnicas geralmente são aplicadas a medições de fenômenos naturais, indicadores econômicos e sinais de sensores, entre outros;

- As medidas estatísticas mais comumente utilizadas são:
  - Média;
  - Variância;
  - Média móvel;
  - Desvio padrão móvel;
  - Autocorrelação;

Séries temporais

Tendência:

- Refere-se à direção de longo prazo dos dados. Pode ser ascendente, descendente ou constante.
- Exemplo: Aumento contínuo do Produto Interno Bruto (PIB) ao longo dos anos.

Sazonalidade:

- São padrões periódicos e previsíveis que ocorrem em intervalos regulares (por exemplo, variações mensais ou sazonais).
- Exemplo: Aumento nas vendas de determinado produto durante as festas de final de ano.

Ruído/Resíduo:

- Refere-se a flutuações irregulares, que podem representar erros pontuais de coleta ou até mesmo choques externos (*e.g.* fatores econômicos ou políticos).
- Exemplo: Guerras, Ciclos econômicos de expansão e recessão.

Análise exploratória

Gráfico de linha: Permitem visualizar a evolução dos dados ao longo do tempo.

Boxplot por período: Úteis para identificar sazonalidade e outliers comparando distribuições em períodos (por exemplo, meses ou trimestres).

Correlogramas:

- **Função de autocorrelação (ACF):** Mede a correlação entre os valores da série e suas defasagens.

$$\rho(k) = \frac{\mathrm{Cov}(X_t, X_{t-k})}{\sigma^2}$$

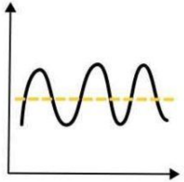
- **Função de Autocorrelação Parcial (PACF):** Mede a correlação entre X_t e X_{t-k}, removendo o efeito das defasagens intermediárias.

Estacionariedade

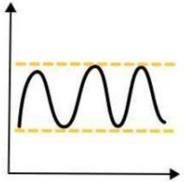
Propriedade fundamental em séries temporais que ocorre quando as características estatísticas da série permanecem constantes ao longo do tempo.

Uma série temporal é considerada estacionária quando apresenta:

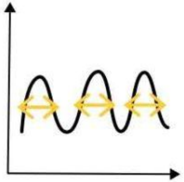
❏ Média constante ao longo do tempo



❏ Variância constante ao longo do tempo



❏ Autocovariância dependente apenas da distância amostral (e não da dimensão temporal);



Estacionariedade

Estacionariedade Fraca:
Ocorre quando a média e a variância são constantes ao longo do tempo, mas a autocovariância pode depender da distância entre as observações.

Estacionariedade Forte:
É mais rigorosa e exige que todas as propriedades estatísticas, incluindo média, variância e autocovariância, permaneçam constantes ao longo do tempo.

Estacionariedade

Teste de Dickey-Fuller Aumentado:

- Hipótese nula (H0): A série temporal não é estacionária porque há uma raiz unitária (se o valor-p > 0,05);
- Hipótese alternativa (H1): A série temporal é estacionária porque não há raiz unitária (se o valor-p ≤ 0,05);

Teste de Kwiatkowski-Phillips-Schmidt-Shin:

- Hipótese nula (H0): A série temporal é estacionária porque não há raiz unitária (se o valor-p > 0,05);
- Hipótese alternativa (H1): A série temporal não é estacionária porque há uma raiz unitária (se o valor-p ≤ 0,05) Quanto mais positiva for essa estatística, maior a probabilidade de rejeitarmos a hipótese nula (temos uma série temporal não estacionária).

Modelagem preditiva

Modelos ARIMA (AutoRegressive Integrated Moving Average):
Modelos que combinam componentes autorregressivos (AR), de média móvel (MA) e integração (I) para séries não estacionárias.

Exemplo de Modelo ARIMA(p, d, q): $\phi(B)(1-B)^d X_t = \theta(B)\varepsilon t$,

onde B é o operador defasador, d o número de diferenças necessárias para estacionaridade, e p e q representam as ordens dos componentes AR e MA, respectivamente.

Modelos SARIMA (Seasonal ARIMA):
Extensão do ARIMA para lidar com a sazonalidade, incorporando termos sazonais.

Suavização Exponencial (Exponential Smoothing):
Métodos que ponderam os dados passados com decaimento exponencial, como o modelo de Holt-Winters, que lida com tendência e sazonalidade.

Modelos de Cointegração

A cointegração ocorre quando duas ou mais séries temporais, individualmente não estacionárias, compartilham uma relação de equilíbrio de longo prazo – ou seja, existe uma combinação linear delas que é estacionária.

Modelo Engle–Granger:
É um método em dois passos para testar e modelar a cointegration entre duas séries, primeiro estimando uma relação de equilíbrio e, depois, ajustando um modelo de correção de erro (ECM) para capturar as dinâmicas de curto prazo.

Modelos Estruturais/VECM:
Generalizam a ideia para múltiplas séries, permitindo modelar simultaneamente os efeitos de longo prazo (por meio dos vetores de cointegracão) e as respostas de curto prazo às perturbações no equilíbrio.

Modelos Engle-Granger

O método Engle–Granger é uma abordagem em dois passos para testar e estimar relações de cointegração entre duas séries:

Regressão de Equilíbrio:

Regressa-se uma série sobre a outra (por exemplo, $y_t = \beta_0 + \beta_1 x_t + u_t$).
Se x_t e y_t forem I(1), mas o resíduo u_t for estacionário (I(0)), então as séries são cointegradas e u_t representa o "erro de equilíbrio" ou a "lacuna" de cointegração.

Modelo de Correção de Erro (ECM):

Em seguida, utiliza-se o resíduo defasado u_{t-1} para modelar as variações de curto prazo em y_t (ou x_t). Um ECM típico para y_t pode ser expresso como:

$$\Delta y_t = \alpha\left(y_{t-1} - \beta_0 - \beta_1 x_{t-1}\right) + \gamma\,\Delta x_t + \varepsilon_t$$

Onde:

- $\Delta y_t = y_t - y_{t-1}$ e $\Delta x_t = x_t - x_{t-1}$ representam as mudanças de curto prazo;
- $y_{t\_1} - \beta_0 - \beta_1 x_{t-1}$ é o termo de erro (o desequilíbrio) da relação de longo prazo;
- α indica a velocidade de ajuste de y_t de volta ao equilíbrio.

Modelos Estruturais VECM

Quando há mais de duas séries cointegradas, o método Engle–Granger é generalizado pelo Vector Error Correction Model (VECM):

Um VECM para k k séries pode ser escrito na forma:

$$\Delta \mathbf{y}_t = \Pi \, \mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta \mathbf{y}_{t-i} + \varepsilon_t$$

Onde:

- $\mathbf{y}_t$ é um vetor k×1 de séries;
- $\Delta\mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$
- Π é a matriz de cointegração, que pode ser fatorada como Π = $\alpha\beta^\top$, com β representando os vetores de cointegração (relações de equilíbrio) e α os coeficientes de ajuste;
- Γ_i captura as dinâmicas de curto prazo;
- p é o número de defasagens utilizado no modelo;
- ε_t é o termo de erro multivariado.

Prática

Exercícios:

1. Selecione uma série temporal, decomponha e analise os componentes (tendência, sazonalidade e ruído).

2. Exiba a média móvel da sua série;

Obrigado

Stefano Mozart
linkedin.com/in/stefano-mozart/
github.com/stefanomozart

CC BY