IA e ML Aplicados a Finanças

Prof. Leandro Maciel

AULA 3: Regressão Linear

Agenda

1. Estrutura e Estimação;
2. Inferência;
3. Análise de diagnóstico;
4. Bibliografia.

1. Estrutura e Estimação

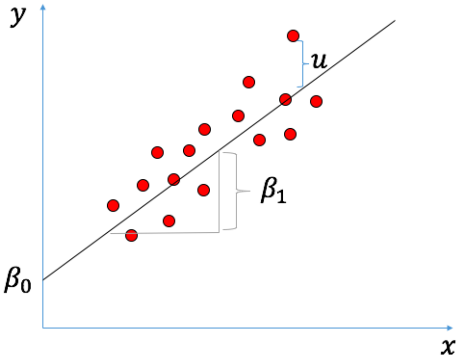
- Regressão linear (simples/múltipla):
  - modelo mais simples de ML (mas excelente!);
  - relação entre variável resposta e suas covariadas;
  - assume hipóteses restritivas.
- Limitações:
  - problemas com dados de elevada dimensão;
  - relações não lineares não capturadas;
  - inferências válidas sob quando hipóteses satisfeitas.

- **Modelo (matemático) de regressão linear múltipla populacional**:

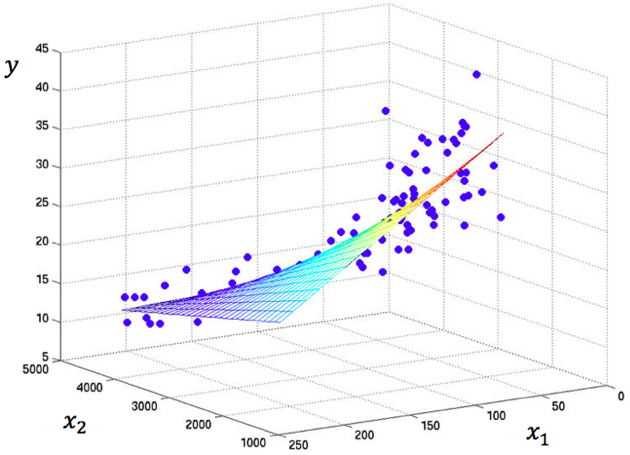
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- $\beta_0 \rightarrow$ intercepto ou constante;
- $\beta_1, \beta_2, \ldots, \beta_k \rightarrow$ parâmetros livres de inclinação das variáveis explicativas;
- $u \rightarrow$ termo de erro ou perturbação (fatores que afetam y não incluídos no modelo, erro de medida, inadequação da forma funcional, etc);
- $k + 1$ parâmetros populacionais desconhecidos $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$;
- representação do mapeamento entrada-saída.

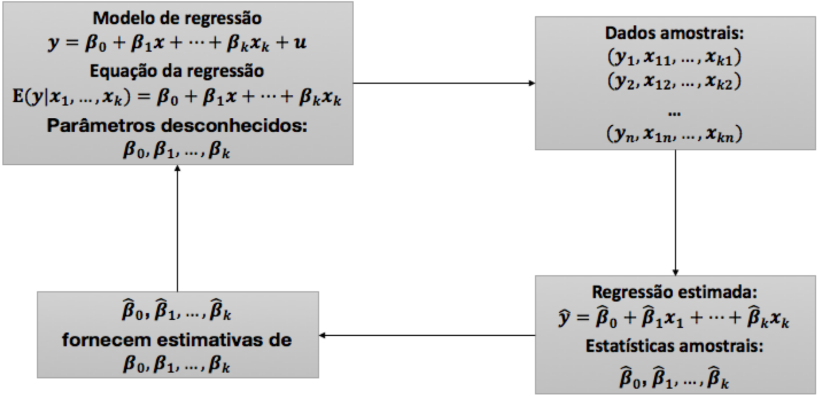
- **Regressão linear simples (RLS)**:



- **Regressão linear múltipla (RLM)**:



- **Identificação regressão linear múltipla**:



- Técnica de aprendizagem: **Mínimos quadrados ordinários** (MQO);

- Estimativas que minimizam a SQR (função custo):

$$SQR = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_k x_{ki})^2$$

- **Estimação** $\rightarrow$ solução do seguinte problema de otimização irrestrita:

$$\min_{(\beta_0, \beta_1, \ldots, \beta_k)} SQR = \min_{(\beta_0, \beta_1, \ldots, \beta_k)} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_k x_{ki})^2$$

- Utilizando cálculo multivariado, derivamos a função e igualamos a zero;
- Obtemos as **equções normais** de mínimos quadrados:

$$\frac{\partial SQR(\beta_0, \beta_1, \ldots, \beta_k)}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_0 - \hat{\beta}_k x_{ki}) = 0$$

$$\frac{\partial SQR(\beta_0, \beta_1, \ldots, \beta_k)}{\partial \beta_1} = -2\sum_{i=1}^{n}\Big[(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_0 - \hat{\beta}_k x_{ki})x_{1i}\Big] = 0$$

$$\vdots \qquad\qquad \vdots$$

$$\frac{\partial SQR(\beta_0, \beta_1, \ldots, \beta_k)}{\partial \beta_k} = -2\sum_{i=1}^{n}\Big[(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_0 - \hat{\beta}_k x_{ki})x_{ki}\Big] = 0$$

- **Interpretação** dos parâmetros:

- $\hat{\beta}_0 \rightarrow$ valor previsto de y quando $x_1 = \ldots = x_k = 0$;

- Demais estimativas mensuram o **efeito parcial** ou *ceteris paribus* sobre y:

$$\frac{\partial y}{\partial x_j} \approx \frac{\Delta y}{\Delta x}$$

$$\hat{\beta}_j \approx \frac{\Delta y}{\Delta x}$$

$$\hat{\beta}_j \approx \Delta y, \text{ quando } \Delta x = 1$$

- Derivada parcial $\rightarrow$ mantidas as demais variáveis constantes.

- Qualidade do ajuste → coeficiente de determinação R^2:

$$R^2 \equiv 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SQR}{SQT}, \quad 0 \leq R^2 \leq 1$$

- Proporção da variação amostral em y que é explicada pela regressão;

- R^2 ajustado (penaliza número de parâmetros estimados):

$$\bar{R}^2 = 1 - \frac{SQR/(n-k-1)}{SQT/(n-1)} = 1 - \frac{\hat{\sigma}^2}{SQT/(n-1)} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

- Como estimar uma regressão no R?
- Vamos considerar base de dados "*beauty*" do pacote *wooldridge*;
- Objetivo: verificar se a aparência influencia nível salarial;
- Desejamos estimar a seguinte regressão:

$$lwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 belavg + \beta_4 abvavg + \beta_5 female + u$$

1. Estrutura e Estimação

- As variáveis do modelo são:

  1. **lwage** - logaritmo natural salários em USD - variável dependente;
  2. **educ** - número de anos de estudo;
  3. **exper** - número de anos de experiência;
  4. **belavg** - binária, 1 para abaixo da média (menos atraente);
  5. **abvavg** - binária, 1 para acima da média (mais atraente);
  6. **female** - binária, 1 para mulher.

1. Estrutura e Estimação

- Estimação usamos a função **lm()**:

> **lm(lwage ∼ educ + exper + belavg + abvavg + female, data = beauty)**

- Símbolo "+" só indica a inclusão de uma nova variável;
- Função tem essencialmente 2 argumentos: fórmula e a base de dados;
- Para estimar o modelo sem intercepto:

> **lm(lwage ∼ 0 + educ + . . .)**

- Regressão estimada:

$$lwage = 0.69 + 0.07educ + 0.01exper - 0.15belavg - 0.01abvavg - 0.46female$$

- Interpretação dos coeficientes? Sinal e magnitude;
- $R^2 = 0.3362$;
- Qual o salário médio para um conjunto de covariadas?

1. Estrutura e Estimação

- Como selecionar covariadas (regressores)?
  - conhecimento teórico do problema;
  - correlação e experiência;
  - testes para encontrar relações escondidas (*data-driven*);
  - cuidado com relações espúrias.
- Qual período (dados) considerar?
  - relações são dinâmicas;
  - contexto que se objetiva realizar inferências.

2. Inferência

- Testar hipóteses sobre os parâmetros do modelo populacional;
- Teste de significância individual dos parâmetros:

$$H_0 : \beta_j = 0 \text{ contra } H_1 : \beta_j \neq 0$$

- Estatística *t-Student* associada:

$$t_{\hat{\beta}_j} \equiv \frac{\hat{\beta}_j}{ep(\hat{\beta}_j)} \sim t_{n-k-1}$$

- Decisão: região crítica ou p-valor, ambos para nível significância α.

- Teste de significância conjunta dos parâmetros:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \text{ contra } H_1 : \beta_1 \neq \beta_2 \neq \ldots \neq \beta_k \neq 0$$

- Estatística F associada:

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \sim F_{k,n-k-1}$$

- Decisão: região crítica ou p-valor, ambos para nível significância α;

- Estatísticas t e F são reportadas usando a função **summary()**.

3. Análise de diagnóstico

Hipótese RLM 1 - Linearidade

Modelo populacional é linear nos parâmetros: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$.

Hipótese RLM 2 - Amostragem Aleatória

Amostra aleatória de n observações: $\{(y_i, x_{1i}, x_{2i}, \ldots, x_{ki}), i = 1, \ldots, n\}$. As estimativas dos parâmetros populacionais dependem da amostra particular.

Hipótese RLM 3 - Colinearidade não perfeira

Na amostra (e, portanto, na população), nenhuma das variáveis independentes é constante, e não há relações **lineares exatas** entre as variáveis independentes.

Hipótese RLM 4 - Média condicional zero

O erro u tem valor esperado igual a zero, para quaisquer valores das variáveis independentes: $E(u|x_1, \ldots, x_k) = 0$.

TEOREMA - Inexistência de viés de MQO

Se as hipóteses RLM 1-4 são satisfeitas: $E(\hat{\beta}_j) = \beta_j, j = 1, \ldots, k$, para quaisquer valores dos parâmetros populacionais $\rightarrow$ **estimadores não viesados**.

Hipótese RLM 5 - Homocedasticidade

O erro u tem mesma variância para quaisquer valores das variáveis explicativas: $Var(u|x_1, x_2, \ldots, x_k) = \sigma^2$.

Teorema de Gauss-Markov

Sob as hipóteses de RLM 1-5, os estimadores de MQO de $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ são os melhores estimadores lineares não viesados (BLUEs) de $\beta_0, \beta_1, \ldots, \beta_k$, respectivamente.

- H1 de linearidade $\rightarrow$ estrutura linear é imposta aos dados;
- H2 de amostra aleatória $\rightarrow$ problema de amostragem;
- H3 de colinearidade perfeita $\rightarrow$ difícil violar na prática;
- H4 de média condicional nula $\rightarrow$ problema de endogeneidade (omissão);
- H5 de homocedasticidade $\rightarrow$ problema de heterocedasticidade.

3. Análise de diagnóstico

- **Heterocedasticidade**: viola hipótese variância constante do erro;
- Parâmetros não viesados, mas variâncias dos estimadores incorretas;
- Inferência (testes) inválida na presença de heterocedasticidade;
- Teste de heterocedasticidade de Breusch-Pagan:

$$H_0 : Var(u|x_1, \ldots, x_k) = \sigma^2 \text{ (homocedasticidade)}$$

- Função **ncvTest()** faz o teste de Breusch-Pagan de heterocedasticidade:

  $>$ ncvTest(*NomeModelo*)

- p-valor $\leq \alpha$ (nível de significância) $\rightarrow$ rejeita H_0 (homocedasticidade);
- Correção das variâncias dos estimadores (**variâncias robustas**);
- Função **hccm()** calcula matriz de variância robusta a heterocedasticidade:

  $>$ hccm(*NomeModelo*) # ou diag(hccm(*NomeModelo*))

3. Análise de diagnóstico

- Análise para modelos com dados de corte transversal;
- Endogeneidade $\rightarrow$ mais complexo, solução variáveis instrumentais;
- Dados de séries temporais $\rightarrow$ problema de correlação dos resíduos;
- Modelagem para avaliar relações entre variáveis;
- Para previsão temos técnicas mais sofisticadas e adequadas.

Problema

Reestime a regressão considerada nessa aula com a inclusão das variáveis: **union** - binária igual a 1 se membro de sindicato; **goodhlth** - binária igual a 1 se tem boa saúde; **black** - binária igual a 1 se negro; **married** - binária igual a 1 se casado; **bigcity** - binária igual a 1 se trabalha em cidade grande; e outras que achar apropriadas. Analise os resultados, interprete, e verifique se o modelo apresenta problemas de heterocedasticidade.

- Próxima aula...
  - RL e precificação de ativos;
  - seleção de modelos por encolhimento (*shrinkage*)...

4. Bibliografia

WOOLDRIDGE, Jeffrey M. **Introdução à Econometria** - Uma abordagem Moderna. Tradução da 4ª Ed. Sâo Paulo: CENGAGE Learning, 2011. Capítulos 3, 4 e 8.

JAMES, Gareth, et al. **An Introduction to Statistical Learning** - With Applications in R. New York: Springer, 2013. Capítulo 3.

Prof. Leandro Maciel

leandromaciel@usp.br