fsp.usp.br Faculdade de Saúde Pública Universidade de São Paulo

Modelos de Regressão

HEP0184 – Bioestatística II

Prof. Gizelton P Alencar, Profa. Denise P Bergamaschi

www.fsp.usp.br/~gizelton

– FSP - 2022 –

Copyright

Regressão linear simples vs múltipla

- Exemplo 1: lbw.xlsx

Regressão linear simples:

1) gráfico de dispersão e boxplots
2) estimar a reta, modelar e fazer inferências
3) avaliar o modelo (coef de determinação R2)
4) gráfico de resíduos: *outlier*, homocedasticidade e transformações necessárias
5) normalidade: Q-Q plot e teste de Shapiro-Wilk
6) conclusão

Regressão linear múltipla

- O que é?
- O que é diferente da regressão linear simples?

Regressão linear múltipla

- Modelo de regressão linear múltipla

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i$$

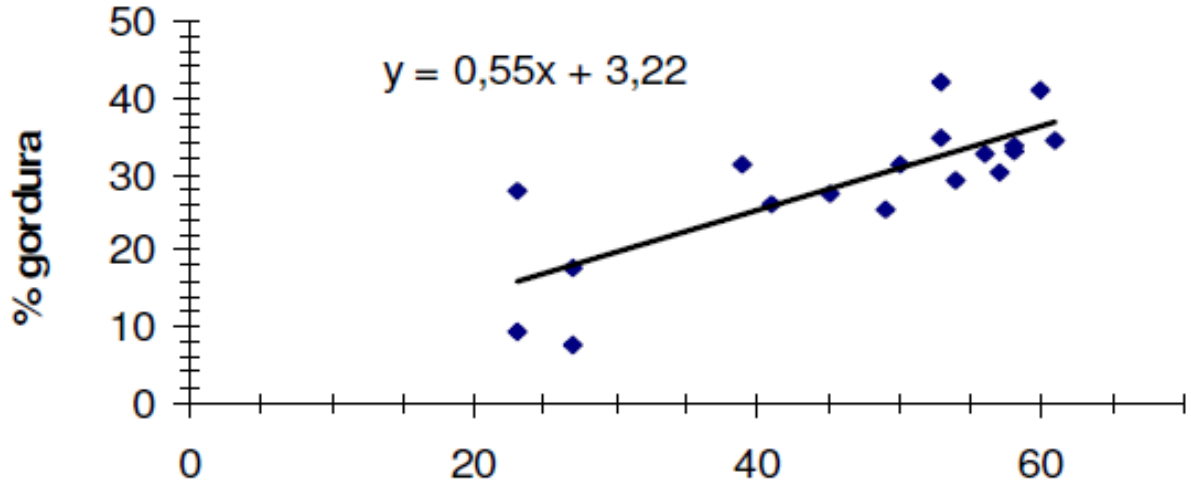
- Y_i: variável resposta (desfecho)
- X_i: variável explicativa
- ε_i representa o termo de erro $\varepsilon_i \sim N(0, \sigma^2)$, independentes
- i é o índice para o indivíduo
- k é o número de variáveis explicativas

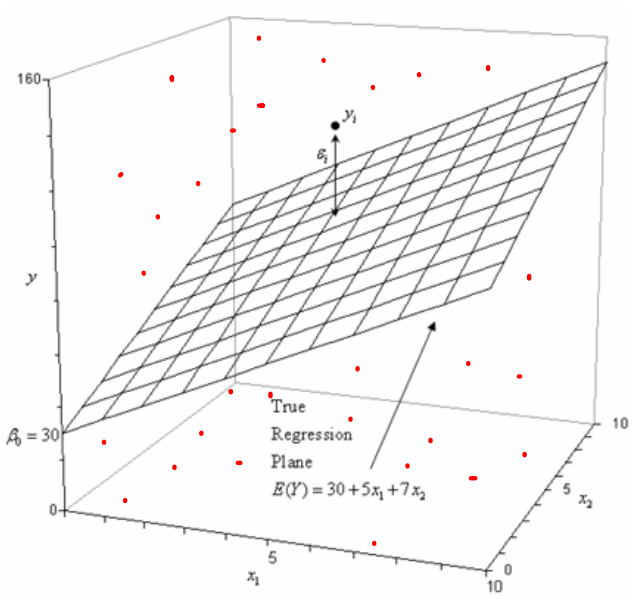
- A estimativa do modelo é: $\hat{Y}_i = b_0 + b_1 X_i$

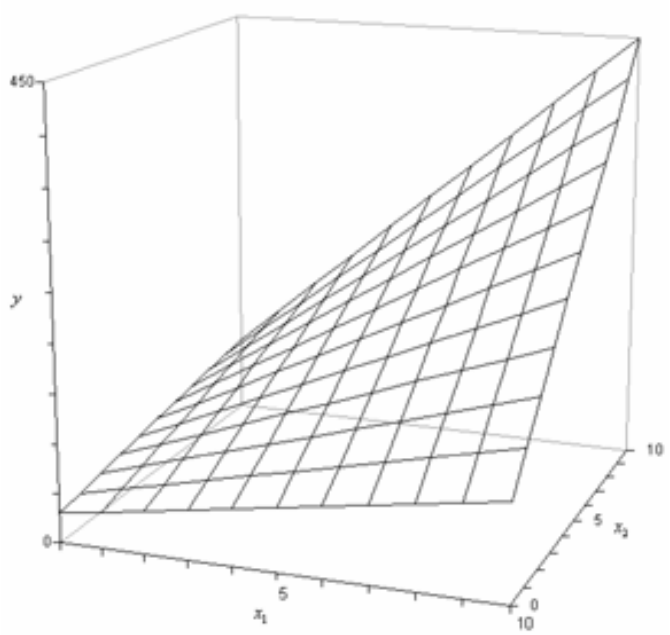
- Os resíduos do modelo são: $e_i = Y_i - \hat{Y}_i$

Regressão linear múltipla

- Exemplo:







Regressão linear múltipla

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2), \text{ independentes}$$

Ex. (cirurgia corretiva no joelho): Tempo de Fisioterapia (dias) até a reabilitação bem sucedida, segundo o escore aptidão física.

Y :

X :

Abaixo da média	Na média	Acima da média
29	30	26
42	35	32
38	39	21
40	28	20
43	31	23
40	31	22
30	29	
42	35	
	29	
	33	

Regressão linear múltipla

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2), \text{ independentes}$$

- Teste F:

$$F = \frac{QME}{QMR} \sim F_{k-1,n-k}$$

Fonte de Variação	*g.l.*	*SQ*	*QM*	*Fo*	*p-valor*
Entre tratamentos	*2*	*672*	*336*	*19,96*	*<0,001*
Resíduos (dentro de tratamentos)	*21*	*416*	*19,81*		
Total	*23*	*1088*			

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$
vs
H_1: $\beta_1 \neq 0$ ou $\beta_2 \neq 0$ ou ... ou $\beta_p \neq 0$.
(pelo menos um)

Se $F > F_{p-1,\, n-p(\alpha)}$ então, rejeita H_0
=> o efeito global de pelo menos uma variável explica a variabilidade de Y

Regressão linear múltipla

Estimação

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2), \text{ independentes}$$

- Métodos de estimação:
  1. mínimos quadrados (*least squares*)
     - regressão linear simples
  2. máxima verossimilhança (*maximum likelihood*)

Regressão linear múltipla – estimação

Método de máxima verossimilhança para estimar β

Verossimilhança é a probabilidade de acontecerem os desfechos (respostas) observados dado um conjunto de parâmetros.

A função de verossimilhança é:

$$L(\beta) = \prod_{i=1}^{n} f(Yi)$$

Regressão linear múltipla – estimação

- não há solução analítica para os valores β_0 e β_1 que maximizam a função de verossimilhança
- métodos numéricos são necessários para encontrar as estimativas de máxima verossimilhança, b_0 e b_1.
- para encontrar os valores ajustados, substitui-se as estimativas b_0 e b_1.

Regressão linear múltipla

Princípio da verossimilhança

- Qual é a função que se ajusta melhor aos dados?

- Função: $Xi \sim F(\Theta)$, $i = 1 \ldots n$
- Maximizar: $L(\{Xi\}, \Theta) = \text{Somatório } F(Xi ; \Theta)$

- P. ex., quer-se prever a variável resposta
- Bom para predição de novos indivíduos

Regressão linear múltipla

Estimadores de máxima verossimilhança

- β_0 -> b_0 (mesmo que mínimos quadrados)
- β_1 -> b_1 (mesmo que mínimos quadrados)
- σ^2 -> s^2: EQM

- Pelo método de mínimos quadrados, o erro ao quadrado é minimizado - entre o predito (y^) e o y observado (y_i) -.

Regressão linear múltipla

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i$$

- Inferências para **cada** coeficiente da regressão
- Teste de Wald

$H_0: \beta_i = \beta_{i0}$ vs $H_1: \beta_i \neq \beta_{i0}$

Teste T: $T_0 = \frac{\widehat{\beta}_i - \beta_{i0}}{\widehat{EP}(\widehat{\beta}_i)}$

Rejeita-se a hipótese, se $T_0 > t_{\alpha,k,v}$ *(k+1)*

Regressão linear múltipla

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i$$

- Inferências para **cada** coeficiente da regressão

1) $H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$

Teste T: $T_0 = \frac{\hat{\beta}_1 - 0}{\widehat{EP}(\hat{\beta}_1)}$

Rejeita-se a hipótese, se $T_0 > t_{\alpha,v}$

2) $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

Teste T: $T_0 = \frac{\hat{\beta}_1 - 0}{\widehat{EP}(\hat{\beta}_1)}$

Rejeita-se a hipótese, se $T_0 > t_{\alpha,v}$

3) $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$

Teste T: $T_0 = \frac{\hat{\beta}_2 - 0}{\widehat{EP}(\hat{\beta}_2)}$

Rejeita-se a hipótese, se $T_0 > t_{\alpha,v}$

Regressão linear múltipla

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i$$

- Inferência para **os coeficientes** da regressão

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k$ vs H_1: $\beta_j \neq 0$, para pelo menos algum j

Teste F: $$F_0 = \frac{MSReg}{MSE}$$

Rejeita-se a hipótese, se $F_0 > f_{\alpha,k,v}$

Regressão linear múltipla

- Exemplo 1: lbw.xlsx

100 bebês nascidos com baixo peso no Hospital de Boston (Pagano 2015)

headcirc: circunferência da cabeça

gestage: idade gestacional

birthwt: peso ao nascer

Vamos analisar a relação entre circunferência da cabeça em relação à idade gestacional e ao peso ao nascer.

Regressão linear simples vs múltipla

- Exemplo 1: lbw.xlsx

Regressão linear múltipla:

0) analisar para cada variável:

1) gráfico de dispersão

2) estimar a reta, modelar e fazer inferências

3) avaliar o modelo (coef de determinação R2)

4) gráfico de resíduos: *outlier*, homocedasticidade e transformações necessárias

5) normalidade: Q-Q plot e teste de Shapiro-Wilk

6*) seleção de variáveis para o modelo final

7) conclusão

Regressão linear

Métodos de seleção de variáveis

`Jr Harrell FE. Regression Models Strategies. Springer. 2001. 566 p.`

O critério para a adição ou remoção de covariáveis é geralmente baseado na estatística F, comparando modelos com e sem as variáveis em questão.

Existem outros métodos, como usar o AIC (*Akaike Information Criteria*).

Existem alguns procedimentos automáticos:

Método *Forward*

Método *Backward*

Regressão linear múltipla

- Exemplo 1: lbw.xlsx

100 bebês nascidos com baixo peso no Hospital de Boston (Pagano 2015)

headcirc: circunferência da cabeça

gestage: idade gestacional

birthwt: peso ao nascer

Vamos analisar a relação entre circunferência da cabeça em relação à idade gestacional e ao peso ao nascer.

Regressão linear múltipla

- Exemplo 1: lbw.xlsx

Vamos ao jamovi

Arquivo word

1. Estudar as variáveis
2. Trabalhar uma variável explicativa
3. Depois a outra variável explicativa

Regressão linear múltipla

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2), \text{ independentes}$$

Ex. (cirurgia corretiva no **joelho**): Tempo de Fisioterapia (dias) até a reabilitação bem sucedida, segundo o escore aptidão física.

Y :

X :

Abaixo da média	Na média	Acima da média
29	30	26
42	35	32
38	39	21
40	28	20
43	31	23
40	31	22
30	29	
42	35	
	29	
	33	

Variável indicadora (*dummy**)

- Variável indicadora (*dummy**)
- Exemplo:
  - Níveis de consumo de fibras (X_1):

    alto vs médio vs baixo
- Quantas variáveis *dummies* são necessárias para uma variável com três categorias?

**dummy = manequim, artificial*

Variável indicadora (*dummy*)

- Exemplo: Variável indicadora (*dummy*): lbw.xlsx

- Categorias de peso ao nascer:

* muito baixo peso ao nascer: <1500g

* peso baixo extremo: <1000g

* peso baixo “mais extremo” < 750g

Variável categórica *birthwtcat*:

0: 1000 -| 1500

1: 750 -| 1000

2: <=750

Variável indicadora (*dummy*)

Variável dummy *bw1*:

0: > 1000

1: <=1000

Variável dummy *bw2*:

0: > 750

1: <=750

Regressão linear múltipla

Elementos - pressuposições (*assumptions*):

- Relação linear entre a resposta e a explicativa
- Erros com distribuição normal (média zero e variância constante)
- Variância dos erros é constante ao longo do X
- Atenção à presença de *outliers* no modelo
- Observações são independentes

- Ausência de multicolinearidade e auto-correlação
- Sobreajuste (superajuste) Overftting e Subajuste (underfitting)

Multicolinearidade

Multicolinearidade: Existem variáveis explicativas no modelo que são altamente correlacionadas (importância da variável)

Como tratar:

Retira do modelo uma das variáveis

Alguns modelos podem dar conta disso (*structured models*, *structural equation models* etc)

Análise fatorial (*factor analysis*)

Multicolinearidade

Exemplo (*lbw.xlsx*):

Resposta (*Y*): perímetro encefálico

X_1 e X_2, resp.: Peso ao nascer e Altura da criança

Coeficiente de correlação r = ???

1) Não faz sentido no quadro conceitual

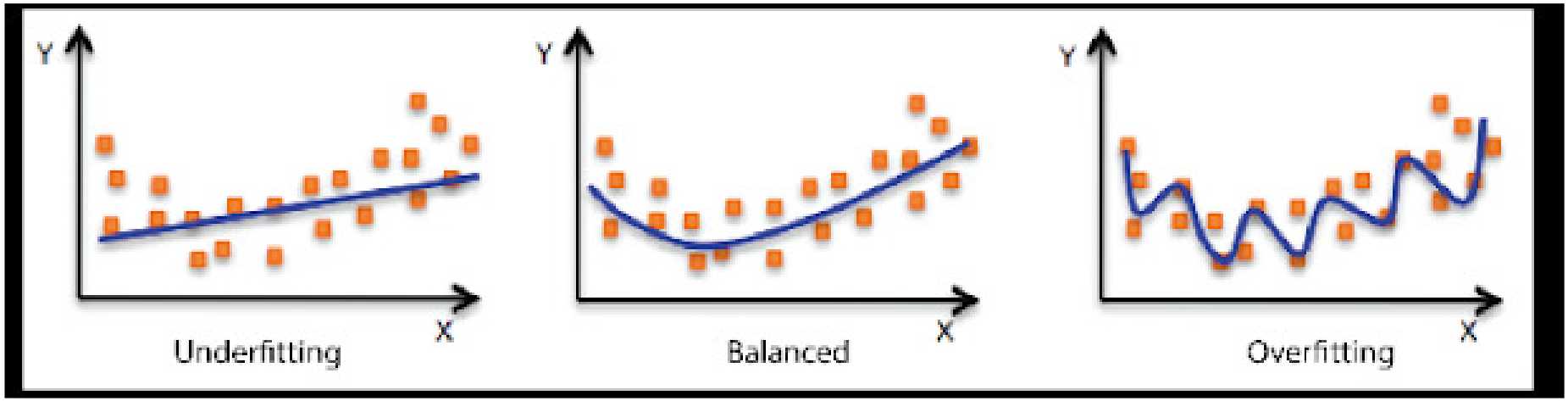
2) Correlação alta

Underfitting e *overfitting*

Subajuste e Sobreajuste (ou superajuste) O que é?

Uso de variáveis explicativas não necessárias

infla o R^2 (e parece que o modelo está melhor)



https://www.listendata.com/2018/03/regression-analysis.html

Referências

- Slides from Jan Wohlfahrt. Department of Epidemiology Research, Statens Serum Institut. Cohort studies: Statistical analysis. http://192.38.117.59/~pka/
- José Maria Pacheco de Souza e Denise Pimentel Bergamaschi. Stata básico. Programa de verão FSP/USP 2015.
- Almeida MF et al. Risk-factors for antepartum fetal deaths in the city of São Paulo, Brazil. Rev. Saúde Pública 2007: 41(1):35-43.
- Almeida S, Barros MBA. Atenção à Saúde e Mortalidade Neonatal: estudo caso-controle realizado em Campinas, SP. Rev. Bras. Epidemiol. 2004.