CIÊNCIA DE DADOS (BIG DATA)

ANÁLISE ESTATÍSTICA

Professor curador: Mário Olímpio de Menezes

EaD Mackenzie

TRILHA 6

PARTE A – MODELOS LINEARES GENERALIZADOS, REGRESSÃO LOGÍSTICA E DE POISSON

PARTE A – MODELOS LINEARES GENERALIZADOS, REGRESSÃO LOGÍSTICA E DE POISSON

REGRESSÃO LOGÍSTICA

MODELOS LINEARES GENERALIZADOS

- Chave para escolher o tipo certo de modelo para diferentes tipos de dados é olhar para a **Variável Dependente.**
- Pode até não ter distribuição normal, mas tem drser *contínua*, *ilimitada* e ser medida em escala *intervalar* ou de *razão*.
- Nem sempre estas condições são verdadeiras!

MODELOS LINEARES GENERALIZADOS

- Variável resposta pode categórica dicotômica (binária):
  - Sim/não, passou/reprovou, viveu/morreu etc.
- Variável resposta também pode ser uma contagem:
  - Número de acidentes de tráfego em uma semana.
  - Número de convulsões por mês de um paciente etc.
- Neste caso, as contagem têm valor limitado, não são negativos, e ainda pode haver um relacionamento entre sua média e variância.

MODELOS LINEARES (RECORDAÇÃO)

- Em um modelo linear padrão, assumimos que *Y* tem uma distribuição normal, e o relacionamento é:

$$\mu_Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

- A média condicional da variável resposta é uma combinação linear das variáveis preditoras.
- A equação é linear nos parâmetros β_j

MODELOS LINEARES GENERALIZADOS

- O modelo é dado pela expressão: $$g(\mu_Y) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

- Em que: $g(\mu_Y)$ é uma função da média condicional (chamada de função *link*).
- Não há suposição de que *Y* tenha distribuição normal.
- Na verdade, *Y* segue uma distribuição da família exponencial.

MODELOS LINEARES GENERALIZADOS

- No **R**, utilizamos a função **glm** para ajustar um modelo linear generalizado.
- A forma da função é similar à `lm()`, mas inclui alguns parâmetros adicionais.

```
glm(formula, family=family(link=function), data=)
```

MODELOS LINEARES GENERALIZADOS

Família	Função **link** padrão
binomial	(link = "logit")
gaussian	(link = "identity")
gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance= "contanst")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

REGRESSÃO LOGÍSTICA

REGRESSÃO LOGÍSTICA

- Quando queremos prever um resultado binário a partir de variáveis preditoras categóricas ou contínuas.
- A variável resposta é dicotômica (0 ou 1).
- O modelo assume que Y segue uma distribuição binomial.

$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

REGRESSÃO LOGÍSTICA

$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

- $p = \mu_Y$ é a média condicional de Y;
- $\left(\frac{p}{1-p}\right)$ é a chance de que Y = 1; e
- $\log\left(\frac{p}{1-p}\right)$ é o log das chances, ou *logit*.
- Neste caso, $\log\left(\frac{p}{1-p}\right)$ é a função **link**, a distribuição de probabilidade é binomial.

PROBABILIDADE E *ODDS* – QUAL É A DIFERENÇA?

- **Probabilidade** é o número de vezes que ocorreu sucesso comparado com o número total de tentativas.
- **Chances** (*odds*) é o número de vezes que ocorreu sucesso comparado com o número de falhas ocorridas.
- Probabilidades iguais: 0.5 → um sucesso a cada 2 tentativas.
- Chances iguais: 1 → um sucesso para cada falha.

PROBABILIDADE E *ODDS* – QUAL É A DIFERENÇA?

Se um evento tem probabilidade de **sucesso** de 0.8, a probabilidade de **falha** é 1 – 0.8 = 0.2

Chances de sucesso é a razão da probabilidade de sucesso pela probabilidade de falha.

Neste caso, temos: $\frac{0.8}{0.2} = 4$

- As chances de sucesso são **4 para 1**

Probabilidade e Chances (*odds*) são termos relacionados, mas não sinônimos!

REGRESSÃO LOGÍSTICA

Como vimos, a regressão logística é dada por:

$$\log_e\left(\frac{p}{1-p}\right) = \text{logit}(p) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

ou

$$\log_e\left(\frac{p}{1-p}\right) = \text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_j X_j$$

Em termos de probabilidades, podemos reescrever assim:

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j)}$$

$$p = \text{P(y=1)}$$


M
Mackenzie