Estatística Aplicada à Epidemiologia 2
Introdução à modelagem estatística

Leo Bastos – leonardo.bastos@fiocruz.br

PROCC – Fundação Oswaldo Cruz

Modelos estatísticos

- O que é um modelo estatístico?
  - É uma representação matemática da relação entre uma variável **resposta**, e uma ou mais variáveis **explicativas**.
- Variável resposta:
  É o **desfecho** de interesse, usualmente denotado pela letra Y.
- Variáveis explicativas:
  É a variável de **exposição**, que pode ser mais de uma, e as variáveis de **controle**, usualmente denotadas pela letra X.

FIOCRUZ

Exemplo

- No problema onde se queira encontrar a relação entre o número de cigarros consumidos por um fumante, e a quantidade de nicotina no sangue de seu(sua) companheiro(a) não fumante.
- Quem é o desfecho e quem é a exposição?

Formulação matemática

- Muitos modelos tem a seguinte forma simplificada:

$$Y = f(X) + \epsilon,$$

  - $f(X)$ é a componente sistemática
  - ϵ é o erro aleatório
- A componente sistemática é uma função matemática das variáveis explicativas.
- O primeiro objetivo da modelagem estatística é estimar o componente sistemático.
- Isso é alcançado analisando dados de vários indivíduos (no exemplo, vários casais com um parceiro fumante e outro não fumante)

FIOCRUZ

Formulação matemática

- Após estimarmos o componente sistemático, temos os chamados valores **ajustados**, denotados por $\hat{Y}$, ou seja,

$$Y = \hat{Y} + \epsilon.$$

- Os valores ajustados permitem que façamos afirmações epidemiologicas sobre uma aparente relação entre o desfecho Y e as variáveis explicativas X.
- No exemplo, podemos dizer qual a quantidade esperada de nicotina no sangue para qualquer quantidade de cigarros fumados pelo seu(sua) parceiro(a).
- Essa 'previsão' claramente é imperfeita, e essa variação entre o esperado e o observado é dada pela aleatoriedade.

FIOCRUZ

- É extremamente importante que o que 'sobra' para o erro aleatório seja realmente aleatório.
- Que não tenha nenhum outro componente sistemático que possa ser removido ou incorporado ao valor ajustado $\hat{Y}$.
- No exemplo, a quantidade de nicotina no sangue de um conjugê não fumante com um companheiro fumante depende **somente** do número de cigarros que o companheiro fuma?

FIOCRUZ

Suposições

- Como definir o componente sistemático? Qual a função matemática mais simples que poderíamos usar?

- O que poderiamos assumir para o componente aleatório?
  - Qual o valor médio poderíamos esperar?
  - Como esses valores poderiam estar distribuídos?

- E quanto aos dados? Será que temos que considerar algo ao coletar os dados?

FIOCRUZ

Modelagem estatística, de uma forma mais geral

- Defina seu desfecho, Y, entenda bem sua natureza.
  - Y é a quantidade de nicotina no sangue do parceiro não fumante.

- Represente a aleatoriedade associada ao desfecho usando uma distribuição de probabilidades.
  - $Y \sim N(\mu, \sigma^2)$

- Defina o valor médio do desfecho como uma função que dependa da exposição e potencialmente de outras variáveis de controle.
  - $\mu = \alpha + \beta X$

- Isso é equivalente a: $Y = \alpha + X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2)$

FIOCRUZ

- OK, mas quais são os valores de $(\alpha, \beta, \sigma^2)$?
- Precisamos agora de uma amostra de tamanho n com observações **independentes** da população de interesse (casais na qual apenas uma das pessoas fuma).
- De posse dessa amostra, denotada por $(Y_1, X_1), (Y_2, X_2), (Y_3, X_3), \ldots, (Y_n, X_n)$, estimamos $(\alpha, \beta, \sigma^2)$.
- Como?

- Usando a distribuição de probabilidades assumida para Y, e a **indepedência** entre as observações construimos o que chamamos de verossimilhança:

$$L(\alpha, \beta, \sigma^2) = \prod_{i=1}^{n} dnorm(y_i, \alpha + x_i\beta, \sigma)$$

- A partir dessa função, estimamos os parâmetros $(\alpha, \beta, \sigma^2)$ basicamente segundo duas abordagens:
  - Abordagem frequentista: maximizando $L(\alpha, \beta, \sigma^2)$
  - Abordagem bayesiana: obtendo a posteriori de $(\alpha, \beta, \sigma^2)$

- Em ambas abordagens temos estimativas pontuais e intervalares para $(\alpha, \beta, \sigma^2)$

Nessa primeira metade do curso:

- Não vamos nos preocupar com a parte matemática **associada a inferência**, vamos deixar isso para o software. Os scripts para análises estarão sempre disponíveis.
- Usaremos a abordagem frequentista, exceto quando for dito o contrário.
- Vamos explorar somente dois tipos de desfechos: desfecho contínuo e desfecho binário.