Regression and Clinical prediction models

Seção 14
Especificação e Estimação – Parte 2

Marcel de Souza Borges Quintana
marcel.Quintana@ini.fiocruz.br
2019

Estrutura

1. **Escolhendo entre modelos alternativos (Cap 6)**
   - Introdução
   - Sugestões gerais
   - Tipos de modelos e alternativas
     - Variável resposta continua
     - Variável resposta dicotômica
     - Variável resposta de tempo (com censuras)
2. **Codificação de preditores categóricos e contínuos (Cap 9)**
   - Categóricos
   - Contínuos
3. **Restrições em candidatos a preditores (Cap 10)**
   - Introdução de Modelos Multivariados
   - Restrição para número de preditores
   - Combinando variáveis similares
4. **Seleção de efeitos principais (Cap 11)**
   - Introdução
   - Seleção por Stepwise
   - Análise univariada e especificação do modelo
5. **Verificando pressupostos em modelos de regressão (Cap 12)**
   - Aditividade
     - Interações
   - Não-linearidade

2. Codificação de preditores categóricos e contínuos (Cap 9)

Categóricos

- Em um modelo linear, preditores categóricos se tornam variáveis dummyes (0 ou 1) para cada categoria. No *R* basta aplicar a função `factor` na variável.
- Por exemplo:
  - Categorias da variável "esquema antirretroviral":
    - Não-HAART
    - IP
    - ITRNN
  - Se escolhendo o tratamento "Não-HAART" como referência teremos em um modelo logístico:

$$log(\frac{p}{1-p}) = \alpha + \text{IP}\beta_1 + \text{ITRNN}\beta_2$$

{ IP = 1 se o tratamento for IP
  IP = 0 caso contrário

{ ITRNN = 1 se o tratamento for ITRNN
  ITRNN = 0 caso contrário

Categóricos

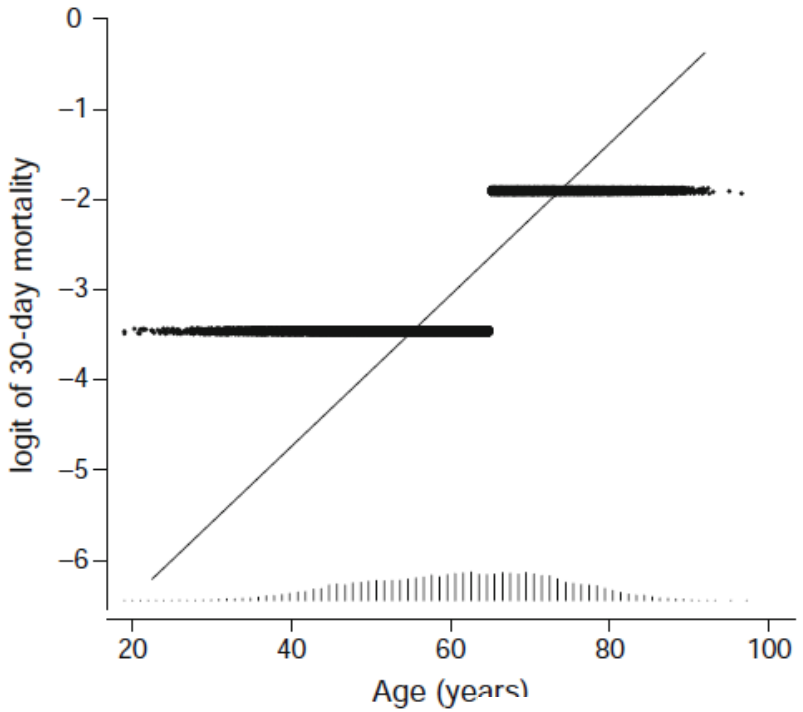
- Cada categoria deve ter um tamanho amostral razoável (ideal, >15 em cada)
  - Caso não haja, devem ser agrupadas as categorias com menor tamanho de amostra
- A escolha da categoria de referência não muda a predição das probabilidades do evento.
- Caso as variáveis categóricas sejam ordinais (e. g. leve, moderado e grave), tratá-la como contínua pode ser uma boa alternativa.

Categóricos

- Exemplo: Classe de Killip para predição de infarto (pág 160)
  - Grau I=1; Grau II=2; Grau III=3; Grau IV=4.
  - Neste caso, vale a pena testar a variável ao quadrado.
    - A estatística χ^2 é substancialmente maior para o modelo considerando a Classe de Killip como ordinal (χ^2 1.388 vs 861).
    - Obs: χ^2 maior indica um melhor ajuste do modelo.

Contínuos

- Exemplo: Idade no estudo de Gusto-I para predição de mortalidade em 30 dias após infarto.
  - Inicialmente, idade pode ser usada como contínua ou categórica (<=65 ou >65)



- Problemas da dicotomização:
  1. Perda de informação;
  2. Precisa haver respaldo para escolha de ponto de corte;
  3. Valores na fronteira abaixo e acima, em geral, não devem apresentar riscos tão diferentes.
- $\chi 2$ para idade categorizada menor do que para contínua (1.463 vs 2.099).

Contínuos

- Quando não há linearidade do efeito existem algumas metodologias alternativas:
  1. Polinômios
     - Colocar, por exemplo, $idade + idade^2$ no modelo (grau 2).
  2. Polinômios fracionais
     - Testa vários graus e opções de polinômios mais complexos.
  3. Splines
     - Método mais complexo em que define-se os graus de liberdade para a função que será ajustada.
       - Quanto maior o número de graus de liberdade maior o viés;
       - Quanto menor o número de graus de liberdade menor é a verossimilhança do ajuste dos dados;
         - » AIC é um critério que auxilia na escolha do número de graus de liberdade da função.

Contínuos

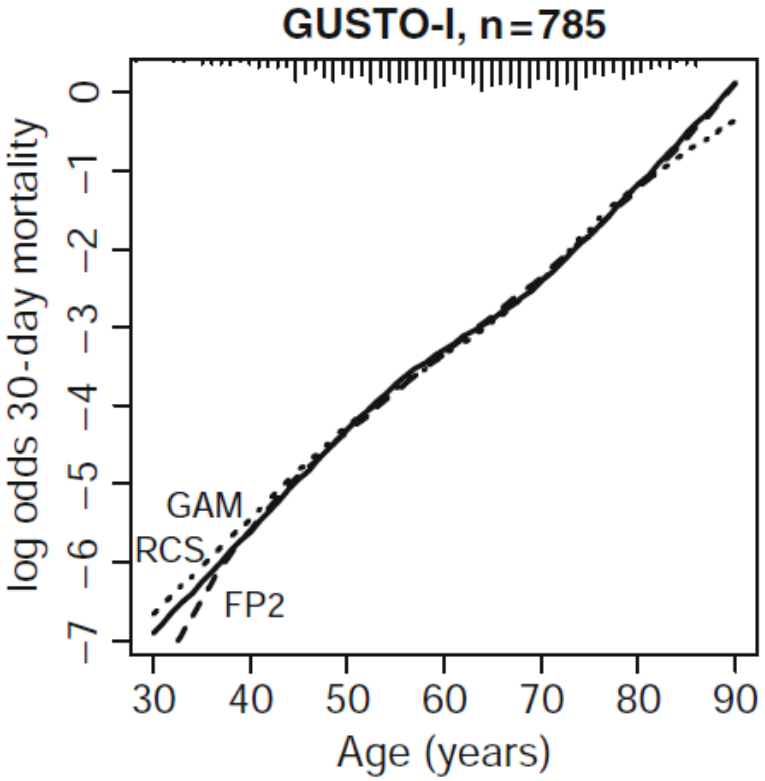


Fig. 9.3 FP, RCS (4 knots, 3 *df*), and GAM (3 *df*) functions for age in a subsample of GUSTO-I

Contínuos

Table 9.4 Options for dealing with continuous predictors in prediction models

Procedure	Characteristics	Recommendation
Dichotomization	Simple, easy interpretation	Bad idea
More categories	Categories capture prognostic information, better but are not smooth, sensitive to choice of cut-points and hence instable	Primarily for illustration
Linear	Simple	Often reasonable as a start
Polynomials	Square, cubic terms added; tails may behave unstable	Reasonable as checks for non-linearity
Transformations	Log, square root, inverse, exponent, etc.	May provide robust summaries of non-linearity
Fractional polynomials	Flexible combinations of polynomials; tails may behave unstable	Flexible descriptions of non-linearity
Restricted cubic splines	Flexible functions with robust behaviour at the tails of predictor distributions	Flexible descriptions of non-linearity
Splines in GAM	Highly flexible functions with smoothness set by penalty terms	Highly flexible descriptions of non-linearity

- Para extrapolação, as transformações polinomiais podem ser mais robustas.
- As maiores diferenças entre as transformações polinomiais e os splines são nas caudas
  1. Se tiver hipótese de inflexão na cauda -> dar preferência polinomiais
  2. Se não tiver -> dar preferência para splines

Contínuos – valores extremos

- Nos casos de valores extremos:
  - Ver se faz sentido o valor
    - Se não houver sentido, rever o prontuário ou colocar como missing.
    - Se houver, a variável pode ser truncada em algum valor (com base na distribuição da variável).

$$\text{If } X > X_{max} \text{ then } X = X_{max};$$
$$\text{If } X < X_{min} \text{ then } X = X_{min};$$
$$\text{else } X = X$$

Contínuos – interpretações de variáveis

- Para o exemplo de predição da probabilidade de morte em 30 dias (modelo logístico)
  - OR(idade)=$\exp(\beta)$=1.088
- Interpretações:
  1. "Para cada ano a mais de idade a chance de morte em 30 dias aumenta em 1.088".
  2. Se quisermos para cada 10 anos de idade
     - Ajustar o modelo para idade10=idade/10
     - OR(idade10)=$\exp(\beta)$=2.32
     - Obs: podemos dividir por IQR, DP, etc.
  3. Para efeitos de variáves não-lineares
     - Comparar o predito para os 1º e 3º quartis
     - q_1=45 e q_3=75 -> p_{q1}=0.2 e p_{q2}=0.5

3. Restrições em candidatos a preditores (Cap 10)

Introdução de Modelos Multivariados

$$f\big(E(Y)\big) = \alpha + X_1\beta_1 + X_2\beta_2 + \cdots + X_q\beta_q$$

- "Quanto mais preditores melhor o modelo" – FALSO
  - Margem para mais confundimento;
  - Perda de dados (graus de liberdade);
  - Total de dados baixo em subcategorias;
    - Interpretação de modelos multivariados.
- Por isso vamos:
  1. Restringir o número de preditores
  2. Combinar variáveis similares

Restrição para número de preditores

1. Seleção baseada na literatura e consultar experts
   - Meta-análise
   - Escolher entre variáveis parecidas (colinearidade, [VIF, R^2])
2. Seleção baseada nas distribuições
   - Retirar variáveis importantes que tiverem número excessivo de NA;
   - Retirar variáveis com *n* pequeno em alguma das categorias.
     - Ou recategorizar
     - Se for uma variável muito importante podemos:
       1. Incluir
       2. Excluir e justificar
       3. Analisar apenas as que não tiveram

Combinando variáveis similares

- Exemplo 1: vários sintomas para uma doença.
  - Criar variável “total de sintomas”
    - Pode ser não linear -> testar termos não-lineares.
- Exemplo 2: fuma atualmente vs tempo de fumo
  - Carga tabágica
- Criar componentes principais.
  - Variáveis demográficas: A, B, C.
  - Variáveis clínicas: D e E.
    - se reduzem para F e G componentes principais.

Exemplo: Predição de mutação no Sistema de reparo de emparelhamentos

- Preditores:
  - Cancer colorretal (CRC)
  - Idade do diagnóstico
- Modelo 1: $\text{Mutation} \sim \text{CRC1}<50 + \text{CRC2}<50$
- Modelo 2: $\text{Mutation} \sim \text{CRC1}<50 + \text{CRC1}\geq 50 + \text{CRC2}<50 + \text{CRC2}\geq 50$

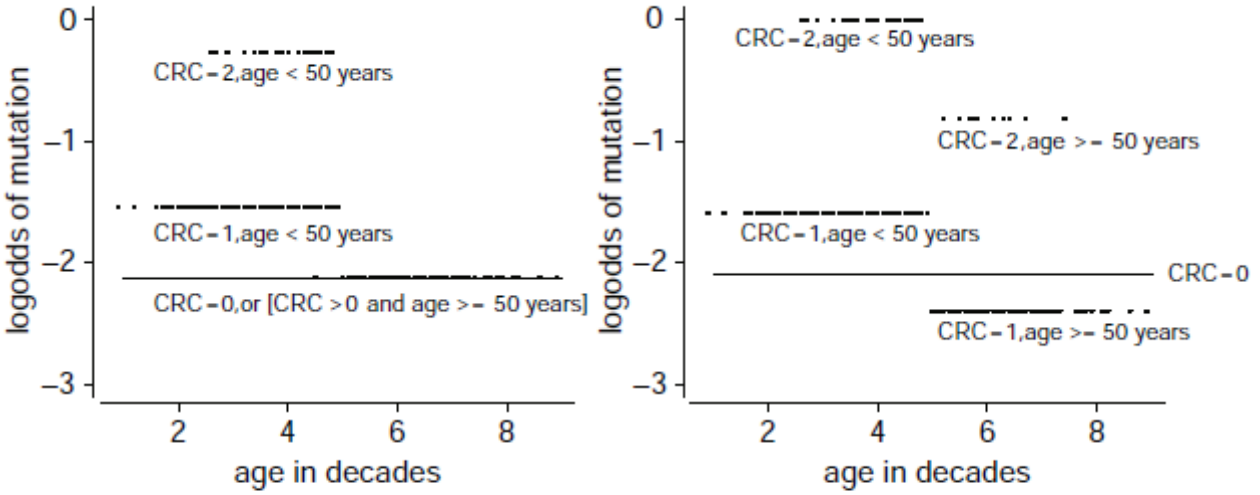


Fig. 10.1 Mutation prevalence in relation to presence of a single or multiple CRCs diagnosed before age 50 in the proband *(left)*, or before or after age 50 *(right)*

3. Restrições em candidatos a preditores (Cap 10)

Exemplo: Predição de mutação no Sistema de reparo de emparelhamentos

- Modelo 3: $\text{Mutation} \sim \text{CRC1} + \text{CRC2} + \text{CRC1age} + \text{CRC2age}$,
- Modelo 4: $\text{Mutation} \sim \text{CRC1} + \text{CRC2} + \text{CRCage}$
  - P-valor=0.80 (Teste de Razão de Verossimilhanças)

Table 10.3 Performance of alternative modes for the predictive effect of CRC and its age of diagnosis in patients tested for mutations in HNPCC (898 patients, 130 mutations). The third coding is preferred (3 *df*), with a single, linear term for the continuous variable "CRCage"

Model	*df*	R^2	*C*
CRC1<50+CRC2<50	2	4.6%	0.602
CRC1<50+CRC2<50+CRC1>=50+CRC2>=50	4	6.9%	0.634
CRC1+CRC2+CRCage	3	7.6%	0.651
CRC1+CRC2+CRC1age+CRC2age	4	7.6%	0.649

3. Restrições em candidatos a preditores (Cap 10)

Exemplo: Predição de mutação no Sistema de reparo de emparelhamentos

- Mais um preditor:
  - Adenoma
- Modelo 5: $\text{Logit(Mutation)} = \text{CRC1} + \text{CRC2} + \text{CRCage} + \text{Adenoma} + \text{AdenomaAge}$
- Modelo 6: $\text{Logit(Mutation)} = \beta_{\text{CRCAdenoma}} \times (\text{CRCage} + \text{AdenomaAge}) + \cdots$

Table 10.4 Performance of alternative modes for the predictive effect of age of diagnosis for CRC and adenoma in the proband,

Model	*df*	R^2	*C*
CRCage + AdenomaAge; adenoma; CRC1+CRC2;	5	8.4%	0.662
CRC.Adenoma.Age; adenoma; CRC1 + CRC2	4	8.4%	0.662

fim

Seção 14
Especificação e Estimação – Parte 2

Marcel de Souza Borges Quintana
marcel.Quintana@ini.fiocruz.br
2019