Regression and Clinical prediction models

Session 7
Introducing statistical modeling – Part 3
(Multivariable linear regression)

Pedro E A A do Brasil
pedro.brasil@ini.fiocruz.br
2019

Objetivos

- Continuar a introduzir conceitos e práticas em modelos estatísticos de regressão.
- Apresentar e discutir extensão do modelo linear para múltiplos preditores.
- Exemplos de método de seleção de preditores e qualidade de ajuste para modelo múltiplo.

Regressão linear múltipla

- Regressão linear múltipla pretende investigar o efeito conjunto das variáveis preditoras (quantitativas e qualitativas) na variável resposta.
- Verifica a magnitude e direção da associação das variáveis independentes como desfecho
- Determina quais das variáveis independentes são importantes na predição do desfecho
- Permite investigar confundimento e interação entre variáveis.

Regressão linear múltipla

- Regressão linear múltipla é um método para análise de relação linear que envolve mais de duas variáveis.
- Podemos pensar na regressão linear múltipla como uma extensão da regressão linear simples.
- A equação da reta de regressão linear múltipla é escrita de forma similar a regressão linear simples
- A qualidade de ajuste é determinada de forma análoga a regressão linear simples.

Regressão linear múltipla

- Regressão linear simples

$$Y_i = \underbrace{B_0 + B_i X_i}_{\text{determinístico}} + \underbrace{\varepsilon_i}_{\text{aleatório}}$$

- Regressão linear múltipla

$$Y_i = \underbrace{B_0 + B_{1iX1i} + B_{2i}X_{2i} + \ldots + B_{ki}X_{ki}}_{\text{determinístico}} + \underbrace{\varepsilon_i}_{\text{aleatório}}$$

- onde k varia de acordo com o número de variáveis, X_1 representa a variável 1, X_2 representa a variável 2, ..., X_k representa a variável k. O índice i corresponde a variação na i-ésima observação.

Regressão linear múltipla

- Regressão linear simples (ajustada)

$$\underbrace{\hat{Y}_i = b_0 + bixi}_{\text{determinístico}}$$

- Regressão linear múltipla (ajustada)

$$\underbrace{\hat{Y}_i = b_0 + b_{1ix1i} + b_{2i}x_{2i} + \ldots + b_{ki}x_{ki}}_{\text{determinístico}}$$

- onde k varia de acordo com o número de variáveis, X_1 representa a variável 1, X_2 representa a variável 2, ..., X_k representa a variável k. O índice i corresponde a variação na i-ésima observação.

Regressão linear múltipla

- A reta estimada pela regressão linear simples foi representada graficamente por gráfico de dispersão bidimensional
- A regressão linear múltipla possui dimensão k+1, o que dificulta visualização gráfica
- A dimensão refere-se ao número de parâmetros estimados (k) e intercepto (1)

Estimação dos coeficientes

- A estimação dos coeficientes do modelo, utilizam o método dos mínimos quadrados, que minimiza a soma do quadrado dos desvios (resíduos) da reta ajustada pelo modelo.

$$SQR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

- Vai além do propósito do curso demonstrar os cálculos dos coeficientes, pois necessitam familiaridade com de cálculos de matrizes.

Interpretação dos coeficientes

- Interpretando os coeficientes:
  - Variação média na resposta, associada a variação unitária em X_k, mantendo o efeito das demais variáveis num valor fixo (ou controlando o valor das demais variáveis).
  - Efeito parcial de cada variável na resposta.
  - β_0: valor médio quando todas as variáveis independentes são iguais a zero

Interpretação dos coeficientes

- Exemplo

```
> regmult <- lm(bwt ~ lwt + age, data = birthwt)
> summary(regmult)

Call:
lm(formula = bwt ~ lwt + age, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-2233.11  -499.33     9.44   520.48  1897.84

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2214.412    299.311   7.398 4.59e-12 ***
lwt            4.177      1.744   2.395   0.0176 *
age            8.089     10.063   0.804   0.4225
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 719.1 on 186 degrees of freedom
Multiple R-squared:  0.03784,        Adjusted R-squared:  0.02749
F-statistic: 3.657 on 2 and 186 DF,  p-value: 0.02767
```

Interpretação dos coeficientes

- Como fazer uma estimativa de um valor de peso ao nascimento do bebê dado que se sabe o peso **e** idade da mãe ao início da gestação, por exemplo de 150 libras e 20 anos?

$$\hat{Y}_i = b_0 + b_1 x_1 + b_2 x_2$$

$$\hat{Y}_i = 2214.41 + 4.177\ x\ 150 + 8.089\ x\ 20 = 3002.74$$

  - Coeficiente (inclinação):
    - Se considerarmos o peso da mãe na última menstruação constante, o peso da criança aumenta em 8,089 g a cada variação de 1 ano de vida. Se considerarmos a idade da mãe constante, o peso da criança aumenta em 4,177g a cada variação de 1 libra do peso da última menstruação. P-valor não significativo
    - Observação: O p-valor da variável idade da mãe não é significativo, ela não deveria entrar no modelo de regressão pois não contribui para melhorar as previsões.
  - Intercepto:
    - O intercepto não tem uma interpretação prática. Ele significa que em média os bebês têm 2214,41 g se x1 e x2 são iguais a 0.

Interpretação dos coeficientes

- Variáveis categóricas
  - Nem sempre estamos trabalhando com variáveis explicativas contínuas
  - Às vezes estamos estimando o efeito de variáveis explicativas nominais ou ordinais
  - Pergunta: Como inseri-las no modelo de regressão?
  - Através de variáveis indicadoras de nominadas *dummy*
  - Variáveis *dummy* são variáveis nominais (dicotômicas) que só assumem valores 1 (categoria presente) ou 0 (categoria ausente)

Interpretação dos coeficientes

- Suponha que desejamos saber o efeito de tabagismo no peso da criança ao nascer (ou seja, duas categorias: 1 = tabagista, 0 = não-tabagista)

```
> regtabag2 <- lm(bwt ~ smoke, data = birthwt)
> summary(regtabag2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3055.70      66.93  45.653  < 2e-16 ***
smoke        -283.78     106.97  -2.653  0.00867 **
---
```

$$\hat{Y} = 3055{,}70 - 283{,}78X_1 \text{ onde } \begin{cases} X_1 = 1, \text{ se Tabagista} \\ X_1 = 0, \text{ se Não tabagista} \end{cases}$$

Se Tabagista $\longrightarrow \hat{Y} = 3055{,}70 - 283{,}78(1) = 2771{,}92g$

Se Não-Tabagista $\longrightarrow \hat{Y} = 3055{,}70 - 283{,}78(0) = 3055{,}70g$

Interpretação dos coeficientes

- E se houver mais de duas categorias na variável explicativa?
  - Devemos entrar com (*j-1*) variáveis indicadoras (*dummy*), onde *j* é o número de categorias da variável preditora *X*
  - Cada variável *dummy* indicará uma das *p* categorias da variável explicativa
  - Se entrarmos com variáveis indicadoras no mesmo número de categorias da variável preditora teremos problemas de ajuste do modelo (comparação múltipla)
  - Esse problema decorre da inversão de matrizes necessária ao cálculo dos coeficientes pelo método dos mínimos quadrados

Interpretação dos coeficientes

- E se houver mais de duas categorias na variável explicativa?
  - Suponha que desejemos saber o efeito da raça no peso da criança ao nascer (1 = white, 2 = black, 3 = other)
  - Primeiro devemos transformar a variável explicativa “Raça” em variáveis *dummies* que representem cada uma das três categorias:

```
birthwt$corwhite <- ifelse(birthwt$race == 1, 1, 0)
birthwt$corblack <- ifelse(birthwt$race == 2, 1, 0)
```

Interpretação dos coeficientes

- Efeito da raça no peso da criança ao nascer

```
> regraca <- lm(bwt ~ corwhite + corblack, data = birthwt)
> summary(regraca)

Call:
lm(formula = bwt ~ corwhite + corblack, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-2096.28  -502.72   -12.72   526.28  1887.28

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2805.28      87.29  32.138  < 2e-16 ***
corwhite      297.44     113.74   2.615  0.00965 **
corblack      -85.59     165.09  -0.518  0.60476
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 714.5 on 186 degrees of freedom
Multiple R-squared:  0.05017,         Adjusted R-squared:  0.03996
F-statistic: 4.913 on 2 and 186 DF,  p-value: 0.008336
```

Interpretação dos coeficientes

- Efeito da raça no peso da criança ao nascer

```
> # Formatando raça como fator
> birthwt$race2 <- factor(birthwt$race, levels = c(3,1,2), labels = c("others","white",
"black"))
> regraca2 <- lm(bwt ~ race2, data = birthwt)
> summary(regraca2)

Call:
lm(formula = bwt ~ race2, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-2096.28  -502.72   -12.72   526.28  1887.28

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2805.28      87.29  32.138  < 2e-16 ***
race2white    297.44     113.74   2.615  0.00965 **
race2black    -85.59     165.09  -0.518  0.60476
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 714.5 on 186 degrees of freedom
Multiple R-squared:  0.05017,   Adjusted R-squared:  0.03996
F-statistic: 4.913 on 2 and 186 DF,  p-value: 0.008336
```

Interpretação dos coeficientes

- Exemplo efeito da raça no peso da criança ao nascer
  - Note que no primeiro exemplo não inserimos a *dummy* de “outras”
  - Note que inserir a variável como fator tendo como referência a mesma categoria da *dummy* não inserida retorna o mesmo modelo.
  - Isso porque na construção das matrizes para estimativa dos coeficientes o algoritmo faz operações semelhantes a criação de *dummies* para criar um coeficiente para cada categoria
  - A escolha da categoria de referência é arbitrária: (a) pode ser por uma questão teórica ou de interpretação para o público; (b) escolher a categoria que possui mais observações como referência pode incrementar a precisão do coeficiente.
  - Além disso, note que a *dummy* raça negra não foi significante. Mas se a raça branca for retida no modelo, as demais categorias *dummy* (que não são referência) também devem ser retidas no modelo como seriam caso a análise fosse com a raça como fator.

Interpretação dos coeficientes

- Exemplo efeito da raça no peso da criança ao nascer: fazendo estimativa do peso ao nascer esperado pelo modelo.

$$Opção\ 1:\ \hat{Y} = 2805{,}28 + 297{,}44X_1 - 85{,}59X_2$$

ou

$$Opção\ 2: \hat{Y} = 2805{,}28 + a_j,\ j = 1 = branca, 2 = negra, 3 = outras$$

$$se\ j = 1 \rightarrow \hat{Y} = 2805{,}28 + a_1 = 2805{,}28 + 297{,}44$$

$$se\ j = 2 \rightarrow \hat{Y} = 2805{,}28 + a_2 = 2805{,}28 - 85{,}59$$

$$se\ j = 3 \rightarrow \hat{Y} = 2805{,}28 + a_3 = 2805{,}28 + 0$$

Selecionando preditores

- Como selecionar preditores para o modelo final?
  - Existem diversos métodos de a contribuição parcial, como teste F parcial teste os coeficientes
  - Não existe um método único ou melhor para a seleção das variáveis que irão compor o modelo final
  - Estabelecer um critério único com antecedência e descrevê-lo na seção de métodos do artigo é suficiente
  - A utilização de diferentes estratégias de seleção pode levar a diferentes modelos finais

Selecionando preditores

- Como selecionar preditores para o modelo com o teste F?
  - Teste F geral (ANOVA):
    - Avalia a contribuição conjunta de todas as variáveis independentes
  - Teste F parcial (ANOVA tipo 1 e 2):
    - Avalia a contribuição parcial de uma variável independente
    - Teste para adição de uma variável independente
  - Teste F parcial múltiplo:
    - Teste para adição de um grupo de variáveis
    - Avalia a contribuição de um grupo de variáveis independente

Selecionando preditores

- Como entender o teste F geral da ANOVA
  - Testa se existe ao menos um coeficiente significativo no modelo (diferente de zero).

$$H_0: B_1 = 0$$
$$H_1: B_1 \neq 0$$

Linear simples

$$H_0: B_1 = B_2 = B_3 = B_p$$
$$H_1: B_1 \neq 0 \; ou \; B_2 \neq 0 \; ou \; B_3 \neq 0 \; ou \; Bk \neq 0$$

Linear múltiplo

- Estatística-teste (F de Snedecor): $F = \frac{QMM}{QMR}$
  - Se p-valor < α. Rejeito H_0, existe ao menos um coeficiente diferente de zero (ou significativo) no modelo.
  - Caso deseje olhar a tabela F, o valor crítico é determinado por $>F_{k,n-k-1,1-\alpha}$.

Selecionando preditores

- Como entender o teste F geral da ANOVA

Fonte de Variação	Soma dos quadrados	gl	Quadrado médio	F
Modelo	$SQM = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2$	k-1	QMM= SQM/k-1	QMM/QMR
Resíduo	$SQR = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$	n-k	QMR= SQR/(n-k)	
Total	$SQT = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2$	n-1	QMT= SQT/n-1	

Selecionando preditores

- Como entender o teste F geral da ANOVA
  - Soma dos Quadrados Total (SQT):
    - Variabilidade total em Y, antes de considerar o efeito conjunto das variáveis
  - Soma dos Quadrados dos resíduos (SQR):
    - Total da variação em Y não explicada pelas variáveis independentes do modelo
  - Soma dos Quadrados do modelo (SQM):
    - Parte da variabilidade de Y explicada pelas variáveis independentes ajustadas no modelo de regressão

Selecionando preditores

- Como entender o teste F geral da ANOVA

```
> regmult <- lm(bwt ~ lwt + age, data = birthwt)
> summary(regmult)

Call:
lm(formula = bwt ~ lwt + age, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-2233.11  -499.33     9.44   520.48  1897.84

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2214.412    299.311   7.398 4.59e-12 ***
lwt            4.177      1.744   2.395   0.0176 *
age            8.089     10.063   0.804   0.4225
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 719.1 on 186 degrees of freedom
Multiple R-squared:  0.03784,       Adjusted R-squared:  0.02749


F-statistic: 3.657 on 2 and 186 DF,  p-value: 0.02767
```

$\sqrt{QMR}$

Selecionando preditores

- Como entender o teste F geral da ANOVA
  - p-value: 0.02767 < 0.05 → Rejeito H_0. Existem evidências estatísticas suficientes ao nível de 5% para afirmar que existe pelo menos um coeficiente (idade ou peso da mãe) diferente de 0.
  - ou
  - A contribuição conjunta das variáveis idade e peso da mãe aumenta significantemente a explicação da variabilidade do peso ao nascer quando comparado ao modelo nulo (sem preditores).

Selecionando preditores

- Se o teste F global (ANOVA) aponta significância, então o melhor modelo é o ajustado?
  - Não, o teste apenas indica que existe pelo menos uma variável que explica de forma importante a variabilidade dos dados em Y (desfecho).
  - Pode ser que uma das variáveis incluídas não tenha uma contribuição importante.
  - Sempre buscar o modelo mais parcimonioso (complexidade vs desempenho vs aplicabilidade).
  - Soluções: observar a contribuição de cada variável por testes de hipóteses, F parcial ou teste t.

Selecionando preditores

- Como selecionar preditores para o modelo final com o F parcial?
  - Teste F parcial
    - Verifica a contribuição parcial de cada variável (explicativa/preditora) na explicação da variável dependente (resposta) Y
    - Testa se a inclusão de uma variável aumenta de forma importante a explicação do desfecho em relação ao modelo sem essa determinada variável
    - Para isso compara dois modelos de regressão ajustada (um com a inclusão da variável e um sem a variável)

Selecionando preditores

- Como analisar se a inclusão da variável tabagismo melhora de forma significativa a explicação do modelo com idade da mãe (anos) e peso da mãe (libras)?
  - São comparados dois modelos para a avaliação da contribuição parcial.

$$H_0 : \beta_3 = 0 \mid \beta_1, \beta_2$$

$$H_1 : \beta_3 \neq 0 \mid \beta_1, \beta_2$$

ou

$$H_0 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$H_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Selecionando preditores

```
> mod1 <- lm(bwt ~ age + lwt, data = birthwt)
> mod2 <- lm(bwt ~ age + lwt + smoke, data = birthwt)
> anova(mod1, mod2)
Analysis of Variance Table

Model 1: bwt ~ age + lwt
Model 2: bwt ~ age + lwt + smoke
  Res.Df      RSS Df Sum of Sq      F  Pr(>F)
1    186 96186834
2    185 92933286  1   3253548 6.4768 0.01175 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value: 0.01175 < 0,05 → Rejeito H_0. Existe evidência estatística suficiente ao nível de 5% para rejeitar que o coeficiente de tabagismo é zero quando ajustado para idade e peso da mãe.

Selecionando preditores

- Teste F parcial

$$F(smoke \mid age, lwt) = \frac{[SQM_{(completo)} - SQM_{(reduzido)}]/1}{QMR(completo)}$$

ou

$$F(smoke \mid age, lwt) = \frac{[SQR_{(reduzido)} - SQR_{(completo)}]/1}{QMR(completo)}$$

$$F(smoke \mid age, lwt) = \frac{(96186834 - 92933286)/1}{92933286/185} = \frac{3253548}{502342,09} \cong 6,48$$

$$valor\ crítico : F(1,185; \alpha = 0,05) \cong 3,89$$

```
> qf(0.95,1,185)
[1] 3.892216
```

ou pelo p-valor:

```
>1-pf(6.4768,1,185)
[1] 0.01174561
```

Selecionando preditores

- Selecionando preditores com ANOVA.
  - Como vimos, precisaríamos realizar inúmeros testes F parciais para verificar a contribuição parcial de cada variável na regressão
  - Há métodos padronizados que realizam todos os testes F parciais de uma única vez
  - Existem inúmeros testes alternativos inseridos nos softwares estatísticos.
  - Os mais conhecidos são:
    - Anova tipo II-Parcialmente sequencial
    - Anova tipo I- Sequencial
    - Anova tipo III-Marginal
  - Diferentes estratégias de seleção podem resultar em diferentes conjuntos de preditores no modelo final.

Selecionando preditores

- Selecionando preditores com F parcial automatizado.
  - Vantagens:
    - Para evitar a realização de diversos modelos na avaliação das contribuições parciais
    - Um único comando no software realiza todas as possíveis contribuições parciais
  - Como proceder:
    - Realize o modelo múltiplo desejado com todas as variáveis pré-selecionadas para a modelagem
    - Utilize o comando “Anova” do pacote “car” (para o tipo II)
    - Cada p-valor lista do refere-se a um teste F parcial realizado para testar a inclusão daquela variável
    - Ou seja, compara o modelo sem e com a variável desejada

Selecionando preditores

```
> anova(mod2) # TIPO I
Analysis of Variance Table

Response: bwt
           Df   Sum Sq Mean Sq F value  Pr(>F)
age         1   815483  815483  1.6234 0.20422
lwt         1  2967339 2967339  5.9070 0.01604 *
smoke       1  3253548 3253548  6.4768 0.01175 *
Residuals 185 92933286  502342
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\neq F(age|1)$

$\neq F(smoke|age)$

$\neq F(lwt|age,smoke)$

p-value: 0.01175 < 0,05 → Rejeito H0. Há evidências estatísticas suficientes ao nível de 5% para afirmar que a inclusão da variável tabagismo melhora significantemente a explicação do peso da criança. O modelo completo é melhor que o modelo só com idade e peso da mãe.

Selecionando preditores

```
> library(car)
Carregando pacotes exigidos: carData
> Anova(mod2)
Anova Table (Type II tests)

Response: bwt
            Sum Sq  Df F value  Pr(>F)
age         261012   1  0.5196 0.47193      F(age|lwt,smoke)
lwt        2739002   1  5.4525 0.02061 *    F(lwt|age,smoke)
smoke      3253548   1  6.4768 0.01175 *    F(smoke|age,lwt)
Residuals 92933286 185
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value: 0.01175 < 0,05 → Rejeito H0. Há evidências estatísticas suficientes ao nível de 5% para afirmar que a inclusão da variável tabagismo melhora significantemente a explicação do peso da criança. O modelo completo é melhor que o modelo só com idade e peso da mãe.

Selecionando preditores

- Seleção de preditores/modelo pelo AIC (*Akaike's information criteria*)
  - Medida alternativa de qualidade de ajuste baseado na verossimilhança do modelo
    - Verossimilhança é a probabilidade que uma população com parâmetros especificados foi examinado dada uma amostra.
  - Quanto menor a estatística (pode ser negativa) → melhor o modelo
  - Possui a vantagem de não ser influenciada pelo tamanho amostral
  - Pode ser utilizada para comparação de modelos, mesmo não aninhados

Selecionando preditores

- Seleção de preditores/modelo pelo AIC (*Akaike's information criteria*)

$$AIC = -2\ln(likelihood) \times 2k$$

  - K é o número de coeficientes a serem estimados
  - *likelihood* é uma função que resulta dos parâmetros (e.g. média e desvio padrão) da distribuição em que o modelo é assumido. Geralmente toma a forma geral de L(parametros | dados), isto é, "*likelihood* de ter esses parâmetros, uma vez que esses são os dados que representam a população.
  - Seguindo os exemplos do peso ao nascimento

```
> reg <- lm(formula = bwt ~ lwt + factor(race) + age + smoke, data = birthwt)
AIC(reg)
[1] 3010.782
```

Selecionando preditores

- Seleção de preditores/modelo pelo AIC
  - O método automático "stepwise" testa de forma automática como a inclusão e/ou exclusão de variáveis melhoram o ajuste do modelo
  - *Quanto menor AIC => melhor modelo*
    - Método do modelo cheio para o vazio (*backwards*)
      - Se a exclusão da variável diminui o AIC → Variável deve ser removida
      - Se a exclusão da variável aumenta o não modifica o AIC → variável deve ser retida
    - Método do vazio para o cheio (*forward*)
      - Se a inclusão da variável diminui o AIC → Variável deve ser retida
      - Se a inclusão da variável aumenta o AIC → variável deve ser removida

Selecionando preditores

```
> step(reg, direction = "both")
Start:  AIC=2472.42
bwt ~ lwt + factor(race) + age + smoke

               Df Sum of Sq      RSS    AIC
- age           1     18305 85162590 2470.5
<none>                     85144285 2472.4
- lwt           1   2464351 87608636 2475.8
- smoke         1   6291918 91436202 2483.9
- factor(race)  2   7789001 92933286 2485.0

Step:  AIC=2470.46
bwt ~ lwt + factor(race) + smoke

               Df Sum of Sq      RSS    AIC
<none>                     85162590 2470.5
- lwt           1   2468766 87631356 2473.9
- smoke         1   6281818 91444408 2481.9
- factor(race)  2   8031708 93194298 2483.5

Call:
lm(formula = bwt ~ lwt + factor(race) + smoke, data = birthwt)

Coefficients:
  (Intercept)            lwt  factor(race)2  factor(race)3          smoke
     2799.285          3.938       -503.914       -394.979       -399.772
```

Akaike do modelo sem "age"
Akaike do modelo com as 4
Akaike do modelo sem "lwt"
Akaike do modelo sem "smoke'
Akaike do modelo sem "race"

Akaike do modelo com os 3 preditores
Akaike do modelo sem lwt
Akaike do modelo sem smoke
Akaike do modelo sem race

Selecionando preditores

- O método automático é realizado em passos (“step”)
- Em cada passo, o modelo testa qual a contribuição de cada variável se for excluída (-) ou incluída (+) no modelo
- Após o primeiro passo, exclui ou inclui variáveis de acordo com o passo anterior e ajusta um novo modelo para teste
- Em cima desse novo modelo de teste, realiza um novo “step” testando novamente a inclusão e/ou exclusão devariáveis
- São realizados diversos “steps” até chegar um momento onde a exclusão / inclusão de variáveis não seja mais possível (AIC não reduza)

Selecionando preditores

- Seleção de preditores
  - Métodos de seleção abordados na sessão
    - teste-t da variável,
    - F parcial,
      - ANOVA tipo 2
      - ANOVA tipo 1
    - Minimização do AIC (stepwise)
  - Na análise de exemplo todos excluem idade e chegam ao mesmo modelo
  - Nem sempre isso acontece
  - A escolha entre um dos métodos deve ser única e pessoal
  - Testar por diversos métodos até eleger o que chega à melhor conclusão pode ser um caminho.
    - Isso não é regra (cuidado para não torturar os dados)

Qualidade do ajuste

- Verificando qualidade do ajuste
  - R^2 ajustado
    - Varia entre 0 e 1 (quanto maior melhor o modelo)
    - É o percentual da variação total do desfecho que o modelo explica
    - Um grande problema é que a estatística R2 geralmente é incrementada com a inclusão de uma nova variável
    - Isto não significa que o ajuste melhorou significantemente, pois pode melhorar em apenas poucos pontos percentuais (exemplo: 1%, 2%...)
    - O R^2 ajustado emprega uma ponderação, que penaliza a estatística de qualidade de ajuste R2 pelo número de variáveis explicativas.
    - A interpretação é a mesma do R2: (%) de variância explicada pelo modelo de regressão

$$R^2 ajustado = 1 - \left(\frac{n-1}{n-2} \right) (1 - R^2)$$

Qualidade do ajuste

- Verificando qualidade do ajuste
  - Exemplo do peso do bebê ao nascimento
    - Somente com o peso da mãe como preditor

```
Residual standard error: 718.4 on 187 degrees of freedom
Multiple R-squared:  0.0345,      Adjusted R-squared:  0.02933
F-statistic: 6.681 on 1 and 187 DF,  p-value: 0.0105
```

    - Peso da mãe e idade como preditores

```
Residual standard error: 719.1 on 186 degrees of freedom
Multiple R-squared:  0.03784,     Adjusted R-squared:  0.02749
F-statistic: 3.657 on 2 and 186 DF,  p-value: 0.02767
```

Qualidade do ajuste

- Interpretação do exemplo acima.
  - Comparação dos valores de R^2, observa-se que o modelo com a variável idade tem um maior poder de explicação (3,8%).
  - Na comparação dos valores de R^2 ajustado, o modelo com idade e peso da mãe explica 2,7% da variabilidade dos dados, enquanto que sem a variável idade a explicação era de 2,9%
  - Ou seja, o R^2 mostra que o modelo com idade da mãe é o melhor, enquanto que o R^2 ajustado indica o modelo sem a variável idade da mãe

Brincar de ajustar modelo linear

- Ajuste um modelo linear de forma interativa.
  - https://danielrivera1.shinyapps.io/Regression2/

fim

Session 7

Introducing statistical modeling – Part 3 (Multivariable linear regression)

Pedro E A A do Brasil
pedro.brasil@ini.fiocruz.br

2018