Regression and Clinical prediction models

Session 9
Introducing statistical modeling – Part 4
(Interaction and collinearity)

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

Objetivos

- Continuar a introduzir conceitos e práticas em modelos estatísticos de regressão.
- Introduzir conceitos de interação e colinearidade na modelagem
- Exemplos de método para explorar a interação e a colinearidade.

Interação

- O que é interação?
  - É um fenômeno semelhante com a confusão / confundimento
  - Interação e confusão geralmente são diferenciados pela magnitude ou direção em que o efeito da variável é ajustado.
  - Um preditor pode ter um efeito *"muito"* diferente (em magnitude ou direção), dependendo do valor de um segundo preditor.
  - Alguns chamam a interação como *modificação de efeito*

Interação

- Explorando confusão: *coeficientes simples vs ajustados*
  - *Simples:*

```
reg1<- lm(bwt ~ lwt, data = birthwt)
summary(reg1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2369.624    228.493  10.371   <2e-16 ***
lwt            4.429      1.713   2.585   0.0105 *
```

  - *Ajustados:*

```
reg2 <- lm(bwt ~ lwt + smoke, data = birthwt)
summary(reg2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2501.125    230.836  10.835   <2e-16 ***
lwt            4.237      1.690   2.507   0.0130 *
smoke       -272.081    105.591  -2.577   0.0107 *
```

Interação

- O que é interação?
  - Para haver interação, é necessário que um preditor seja associado com outro preditor, e ambos sejam associados com o desfecho. Um preditor não deve ser intermediário do efeito no desfecho do outro preditor.
  - *Exemplo (fictício):*
    - O risco de adoecer por malária entre os homes é 1.5x o risco das mulheres de adoecer de malará numa região.
    - O risco de adoecer de malária para quem trabalha em campos abertos é de 4x o risco de quem trabalha em ambientes fechados.
    - Quando estratificado pelo ambiente de trabalho o risco dos homes de adoecer é o mesmo que o risco das mulheres.

Interação

- O que é interação?
  - *Exemplo (continuação)*
    - O “mesmo” risco entre homes e mulheres, no ambiente aberto é 6x o risco do ambiente fechado.
    - Nesse exemplo, o efeito observado entre os homes é atribuído pelo ambiente de trabalho já que homens mais frequentemente trabalham em espaços abertos, e mudou simultaneamente o efeito do sexo e do ambiente de trabalho para diferentes direções, amplificando um e anulando o outro.

Interação

- Quando verificar esse comportamento?
  - Esse assunto também será mencionado em outras sessões.
  - Não há regras universais e obrigatórias.
  - Conhecimento teórico prévio do objeto
  - Diferenças da prevalências dos preditores entre as categorias do desfecho pode ser indicador.
  - Diferenças importantes na magnitude do efeito bruto e ajustado no modelo é uma indicação de possível modificação de efeito.
- Para verificar esse comportamento precisamos explorar o "termo de interação", simultaneamente com os preditores
- Para isto, basta multiplicar os preditores:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$$

Interação

- Exemplo 1: Ajustar um modelo de regressão para verificar o efeito conjunto de peso da mãe e tabagismo na previsão do peso da criança
  - Passo 1: criar um variável que indica a interação
  - No R:

```
# Cria uma variável indicando a multiplicação de outras duas
> birthwt$pesomaefumo <- birthwt$lwt * birthwt$smoke
```

Interação

- Passo 2: Calcular os coeficientes da regressão
  - *Sem interação:*

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

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2501.125    230.836  10.835   <2e-16 ***
lwt            4.237      1.690   2.507   0.0130 *
smoke       -272.081    105.591  -2.577   0.0107 *
---
```

  - *Com interação:*

```
> reg2<-lm(bwt~lwt+smoke+pesomaefumo,data=birthwt)
> summary(reg2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2350.578    312.733   7.516 2.35e-12 ***
lwt            5.387      2.335   2.307   0.0222 *
smoke         41.384    451.187   0.092   0.9270
pesomaefumo   -2.422      3.388  -0.715   0.4757
---
```

Interação

- Passo 2: Calcular os coeficientes da regressão
  - *Sem interação:*

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

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2501.125    230.836  10.835   <2e-16 ***
lwt            4.237      1.690   2.507   0.0130 *
smoke       -272.081    105.591  -2.577   0.0107 *
---
```

  - *Com interação:*

Foram alternativa de especificar interação

```
reg3<-lm(bwt~lwt+smoke+lwt*smoke,data=birthwt)
summary(reg3)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2350.578    312.733   7.516 2.35e-12 ***
lwt            5.387      2.335   2.307   0.0222 *
smoke         41.384    451.187   0.092   0.9270
lwt:smoke     -2.422      3.388  -0.715   0.4757
```

Interação

- Passo 3: Interpretar a saída
  - *Com interação:*

```
> reg2<-lm(bwt~lwt+smoke+pesomaefumo,data=birthwt)
> summary(reg2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2350.578    312.733   7.516 2.35e-12 ***
lwt            5.387      2.335   2.307   0.0222 *
smoke         41.384    451.187   0.092   0.9270
pesomaefumo   -2.422      3.388  -0.715   0.4757
---
```

- A interação não é significante (p=0,476) pelo teste t, portanto não deveria entrar no modelo.
- Não há evidência estatística de que o peso da mãe tenha efeito diferente no peso da criança dependendo se a mãe fuma ou não
- Atenção: Os p-valores dos efeitos principais de peso da mãe e tabagismo não devem ser interpretados separadamente

Interação

- Passo 4: Escrever a equação
  - Se a interação fosse significativa...
  - *Com interação:*

```
> reg2<-lm(bwt~lwt+smoke+pesomaefumo,data=birthwt)
> summary(reg2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2350.578    312.733   7.516 2.35e-12 ***
lwt            5.387      2.335   2.307   0.0222 *
smoke         41.384    451.187   0.092   0.9270
pesomaefumo   -2.422      3.388  -0.715   0.4757
---
```

$$\hat{Y} = 2350{,}58 + 5{,}39x_1 + 41{,}38x_2 - 2{,}42x_1x_2$$

$$onde\ x_1 \rightarrow \text{Peso da mãe (em libras)}\ \text{e}\ x_2 \rightarrow \text{Tabagismo}\ \ (1 = \text{Sim},\ 0 = \text{Não})$$

$$Se\ x_2 = 1 \rightarrow \hat{Y} = 2350{,}58 + 5{,}39x_1 + 41{,}38(1) - 2{,}42x_1(1) = 2391{,}96 + 2{,}97x_1$$

$$Se\ x_2 = 0 \rightarrow \hat{Y} = 2350{,}58 + 5{,}39x_1 + 41{,}38(0) - 2{,}42x_1(0) = 2350{,}58 + 5{,}39x_1$$

Interação

- Passo 5: Interpretar as equações
  - Se a interação fosse significativa...

$$\hat{Y} = 2350{,}58 + 5{,}39x_1 + 41{,}38x_2 - 2{,}42x_1x_2$$

$$onde\ x_1 \rightarrow \text{Peso da mãe (em libras)}\ \text{e}\ x_2 \rightarrow \text{Tabagismo}\ \ (1 = \text{Sim},\ 0 = \text{Não})$$

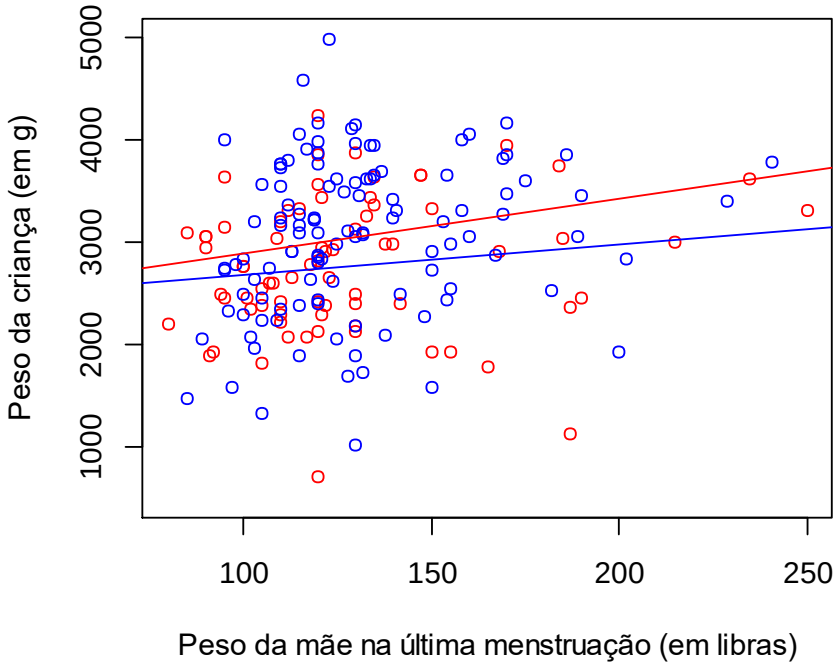
$$Se\ x_2 = 1 \rightarrow \hat{Y} = 2350{,}58 + 5{,}39x_1 + 41{,}38(1) - 2{,}42x_1(1) = 2391{,}96 + 2{,}97x_1$$

$$Se\ x_2 = 0 \rightarrow \hat{Y} = 2350{,}58 + 5{,}39x_1 + 41{,}38(0) - 2{,}42x_1(0) = 2350{,}58 + 5{,}39x_1$$

  - Note que existem duas retas de regressão com diferentes interceptos e inclinações.
  - Essas duas retas relacionam o peso da criança ao nascer e o peso da mãe na última menstruação (x1)
  - Essas diferentes retas dependem dos resultados de tabagismo.

Interação

- Passo 5: Interpretar as equações
  - Se a interação fosse significativa...



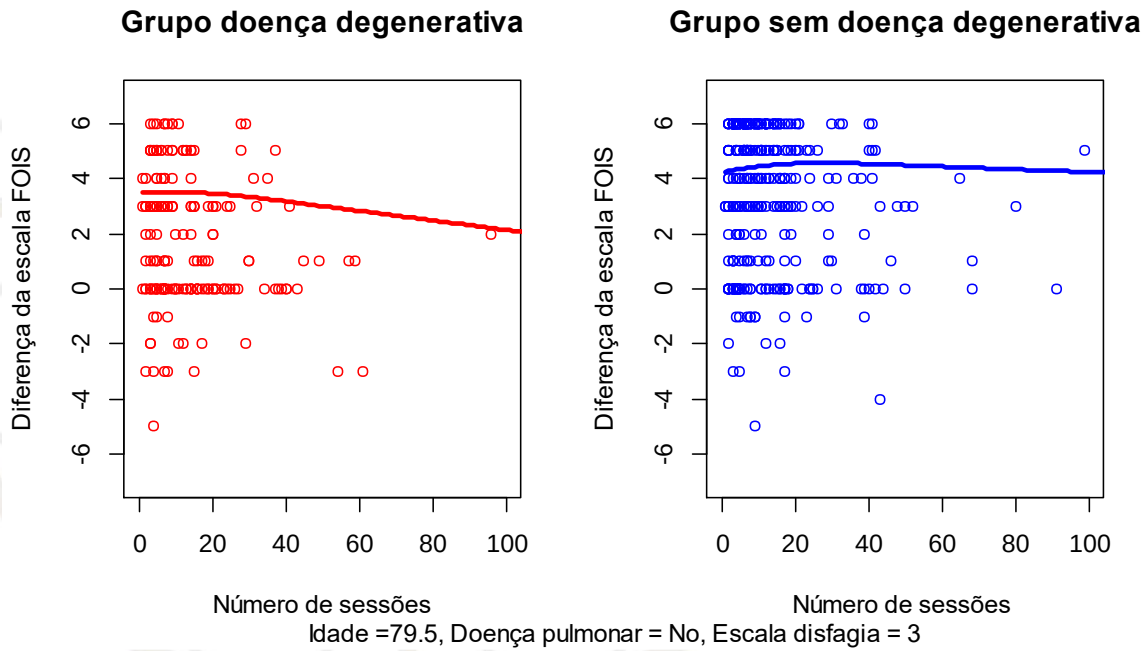
Observa-se os diferentes interceptos e inclinações (apesar das diferenças não contribuir para uma melhor previsão).

Interação

- Passo 5: Interpretar as equações
  - Se a interação fosse significativa (dados do peso ao nascer) ...
  - Embora a inspeção visual evidenciar que os interceptos podem se cruzar antes do valor 0 do peso da mãe, o menor valor observado é 80 lbs.
  - Assim a reta das mães tabagistas sempre será com menores valores de peso ao nascer menor que as mães não tabagistas.
  - O peso médio das crianças entre as não tabagistas sempre será maior que entre as tabagistas.
  - Embora não haja significância a 5% , a interpretação seria as mesmas que acima em novas amostras com significância.

Interação

- Passo 5: Interpretar as equações
  - Outro exemplo de interação, desta vez com significância.
  - Pacientes com dificuldade de deglutição são avaliados pela escala FOIS no início e no final da internação
  - São ofertadas sessões de fonoaudiologia durante a internação



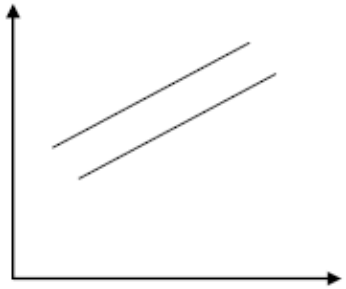
Aqui há uma sofisticação de um efeito não linear. Observa-se os diferentes interceptos e inclinações ao longo do número de sessões.

Interação

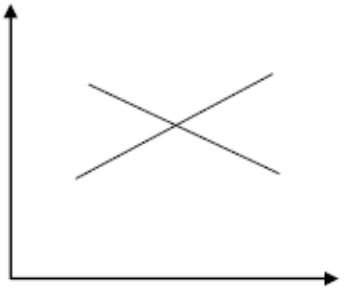
- Passo 5: Interpretar as equações
  - Embora a inspeção visual evidenciar que os interceptos são próximos de 4 na diferença final – inicial da escala FOIS, em média o grupo de doença *Não degenerativa* tem valores iniciais mais elevados.
  - Assim entende-se que o prognóstico pela escala FOIS para pacientes com doença degenerativa de partida é sempre pior, já que a diferença entre o final e inicial em média é menor.
  - O efeito não linear torna a interpretação um pouco sofisticada.
  - Nesse caso, não se observa qualquer benefício do número de sessões entre os pacientes com condições degenerativas, enquanto que nos com não degenerativa há um acréscimo inicial e um decréscimo depois de aproximadamente 40 sessões voltando progressivamente ao patamar inicial.

Interação

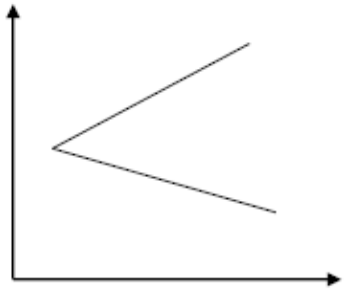
– Possíveis padrões na interpretação de interação



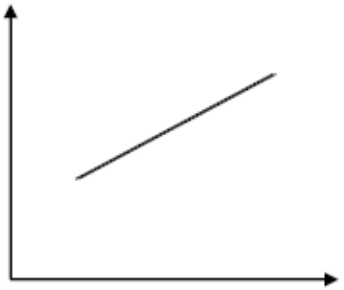
Retas paralelas (diferentes interceptos e mesma inclinação)



Retas não- paralelas (diferentes interceptos e inclinações)



Retas não- paralelas (mesmo intercepto e diferentes inclinações)



Retas coincidentes (mesmo intercepto e inclinação)

Interação

- Passo 5: Interpretar exploração de interação
  - Se desejar, podem ser realizados testes para verificar se a inclinação ou intercepto das retas tem diferença significante
  - A estratégia mais simples é realizar dois modelos de regressão (no exemplo anterior: um para os fumantes e outro para os não fumantes) e comparar as estimativas do intercepto e inclinação.
  - Podem ser realizados também testes F parciais específicos:
    - *Coincidência*: compare um modelo com interação com um modelos em interações em a variável responsável pela inclinação.
    - Exemplo:

```
reg1<-lm(bwt~lwt,data=birthwt)
reg2<-lm(bwt~lwt+smoke+lwt*smoke,data=birthwt)
anova(reg1,reg2)
```

$$H_0 : \beta_2, \beta_3 = 0 \mid \beta_1$$
$$H_1 : \beta_2, \beta_3 \neq 0 \mid \beta_1$$

- Observe a saída. Rejeita-se H0. As retas não são coincidentes

Interação

- Passo 5: Interpretar exploração de interação
  - *Paralelismo*: Compare um modelo com interação com um modelo sem interação.
  - Exemplo:

$$H_0 : \beta_3 = 0 \mid \beta_1, \beta_2$$
$$H_1 : \beta_3 \neq 0 \mid \beta_1, \beta_2$$

```
> reg3 <- lm(bwt ~ lwt  + smoke, data = birthwt)
> reg2 <- lm(bwt ~ lwt + smoke + lwt*smoke, data = birthwt)
> anova(reg3,reg2)
Analysis of Variance Table

Model 1: bwt ~ lwt + smoke
Model 2: bwt ~ lwt + smoke + lwt * smoke
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1    186 93194298
2    185 92937722  1    256576 0.5107 0.4757
```

Equivale ao teste t do modelo com intercepto

Colinearidade

- O que é colinearidade?
  - Ocorre quando um preditor é muito correlacionado com outro preditor o que torna muito difícil ou impossível distinguir seus efeitos independentes. (multi-colinearidade)
  - Geralmente isso acontece com preditores que medem um mesmo conceito de forma diferente.
    - *Exemplo: hemoglobina e hematócrito representando anemia como preditores de tempo de permanência hospitalar*

Colinearidade

- O que é colinearidade?
  - Espera-se que tenham uma alta associação entre si por representar o mesmo constructo, ou seja, informam a mesma coisa para o modelo
  - Quando possível estimar o efeito dos preditores simultaneamente, observa-se uma mudança importante na magnitude dos efeitos (comparado com o efeito bruto)
  - Não faz sentido inserir dois preditores que representam a mesma coisa e são colineares
  - Pode ser considerado um caso particular extremo de interação ou confusão.

Colinearidade

- Explorando a colinearidade:
  - Não é necessário uma correlação perfeita para haver problemas de estimação no modelo
  - Usualmente o efeito dos preditores colineares mudarão muito quando outros preditores são incluídos ou excluídos do modelo
  - Erros padrão dos coeficientes tenderão a ser grandes quando comparados ao efeito univariado.
  - Em geral intervalos de confiança de preditores colineares são muito maiores que os demais preditores.
  - Preditores com efeitos conhecidos poderão ter seus efeitos anulados ou não significantes

Colinearidade

- Explorando a colinearidade:
  - *Tolerânicia*: quando regredir um preditor colinear pelo outro, o R2 da regressão será próximo de 1. Tolerância é definida como 1-R, e alguns definem o limite (arbitrário) de 0,10.
  - É comum que a tolerância seja expressa como *variance inflation factor* (VIF), que é 1/tolerância. Assim tolerância de 0,10 ou menos, tornam-se **VIF de 10 ou mais**.
  - Variáveis transformadas por polinômios, *splines* ou por somas de transformações como quadrado e cubo, para capturar efeitos não lineares podem apresentar VIF elevados sem prejuízo do modelo.

Colinearidade

- Explorando a colinearidade:
  - **Extendendo o conceito de VIF**: em modelos de múltiplos preditores, múltiplas correlações são observadas, uma para cada par de coeficientes (preditores categóricos).
  - A colinearidade pode ser multideminsional.
  - **GVIF**s (generalized VIF) facilita a comparação entre as dimensões.
  - **GVIF^(1/(2*Df))**, onde Df é o número de coeficientes avaliados, reduz GVIF a uma medida linear e será idêntico ao VIF em um modelo simples (com um ou dois preditores)
  - A interpretação de **GVIF^(1/(2*Df))** e **VIF** é a mesma.

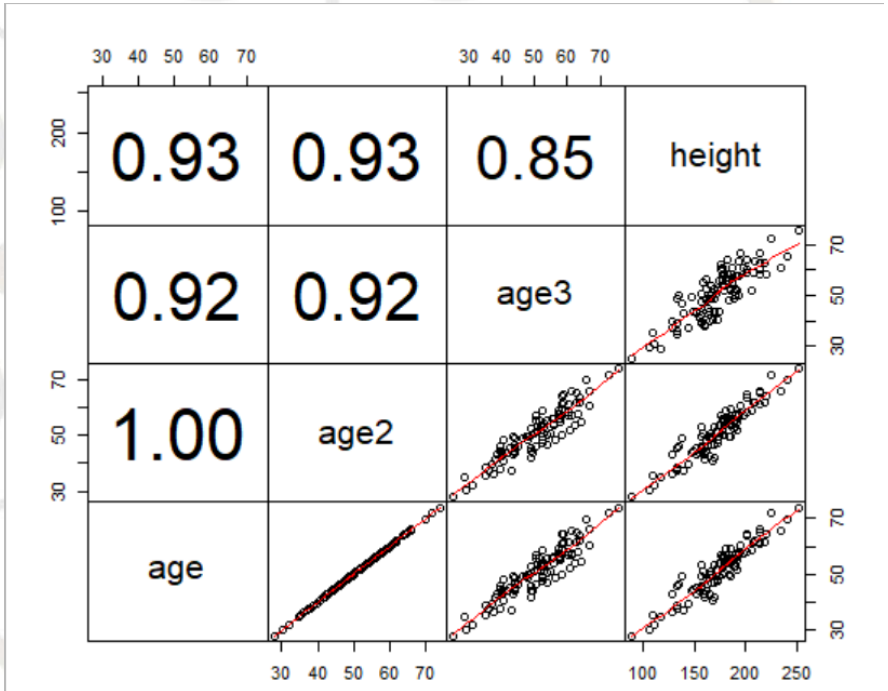
Colinearidade

- Explorando a colinearidade:
  - Excluir variáveis colineares é boa prática, mas sem conhecimento teórico do objeto de estudo pode ser difícil decidir quais devem ser excluídas ou retidas.
  - Exemplos a seguir:
    - Dados fictícios...
      - age = gerado aleatoriamente da normal com média = 50 e sd = 10
      - age2 = copia exata de age
      - age3 = copia de age adicionando erro com sd = +4
      - height = 20 + 3*age + adicionado erro sd = 10
  - **Antes da modelagem:**
    - **Contínuas - Correlação de linear Pearson**
    - Categóricas – raramente explorada, difícil interpretação

Colinearidade

- Explorando a colinearidade:
  - Antes da modelagem

```
pairs(height.dat, lower.panel = panel.smooth, upper.panel = panel.cor, gap = 0, row1attop = FALSE)
```



No painel superior os índices de correlação de Pearson de cada par de variáveis. No painel inferior uma visualização de pontos da correlação de cada para variáveis.
Correlação alta entre todas as variáveis e perfeita entre age e age2.

Colinearidade

- Explorando a colinearidade: *depois da modelagem*
  - *Um preditor:*

```
reg4 <- lm(height ~ age, data = height.dat)
summary(reg4)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.2181     6.0316   3.186  0.00193 **
age           3.0211     0.1163  25.978  < 2e-16 ***
```

  - *Preditor c/ erro:*

```
reg6 <- lm(height ~ age3, data = height.dat)
summary(reg6)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.3778     8.1444   5.449 3.78e-07 ***
age3          2.5362     0.1571  16.148  < 2e-16 ***
```

Colinearidade

- Explorando a colinearidade: *depois da modelagem*
  - *Colinear:*

```
reg7 <- lm(height ~ age + age2, data = height.dat)
summary(reg7)
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.1582     5.8445   2.251   0.0266 *
age          3.1167     0.1129  27.599   <2e-16 ***
age2             NA         NA      NA       NA
```

  - *Colinear c/ erro:*

```
reg8 <- lm(height ~ age + age3, data = height.dat)
summary(reg8)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.5516     5.8285   2.153   0.0338 *
age           3.4801     0.2771  12.560   <2e-16 ***
age3         -0.3570     0.2488  -1.435   0.1545
```

Colinearidade

- Explorando a colinearidade: *depois da modelagem*
  - *VIF:*

```
> rms::vif(reg8)
     age     age3
6.085739 6.085739

> rms::vif(reg7)
 age age2
  NA   NA
```

  - O coeficiente de age2 em reg7 não é estimado, o R entende automaticamente que há colinearidade e desconsidera age2
  - O coeficiente de age em reg4 é o mesmo que em reg7
  - O coeficiente de age3 em reg8 é invertido em relação a reg6.
  - Os erros padrão em reg8 são aproximadamente o dobro das mesmas variáveis em reg4 e reg6
  - VIF de age e age2 nem é estimado, e de age e age3 é alto mas não chega ao limite de 10.

Brincar de ajustar modelo linear

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

fim

Session 9
Introducing statistical modeling – Part 4
(Interaction and collinearity)

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