7b. Modelos Lineares Múltiplos

Videoaula do curso **Princípios de Planejamento e Análise de Dados**. Os conceitos abordados são os mesmos desse tutorial, desconsidere referências à disciplina.

Os modelos lineares permitem que sejam incluídas mais do que uma variável preditora, como fizemos até aqui. Nesse tutorial vamos aprender alguns princípios básicos desses modelos mais complexos. Em modelos com mais de uma variável preditora, precisamos tomar a decisão de quais variáveis devemos reter em nosso modelo. É desejável interpretar o modelo mais simples e que contém apenas as variáveis que explicam porções consideráveis da variação na variável resposta.

Formulas Estatísticas

O argumento `formula` da função `lm` funciona de forma diferente das formulas matemáticas e deve-se ter cuidado com a inclusão de termos ou operações dentro dela.

Alguns aspectos básicos do argumento:

- `y ~ x` indica: construa o modelo da variável resposta y como **função estatística linear** de x;
- `y ~ x1 + x2` indica: construa o modelo estatístico de y como **função linear** das variáveis `x1` e `x2` como tendo efeitos aditivos;

Se quisermos utilizar os símbolos matemáticos no sentido matemático usual

dentro de uma fórmula estatística, temos que utilizara a função `I()` para que a operação seja realizada antes da construção do modelo:

- `y ~ I( x1^2 * x2^3 )` indica: modele y como função estatística **da variável** `x1^2 * x2^3`;
- `y ~ I( x1 / x2 )` indica: modele y como função estatística **da variável** `x1/x2`;

No caso de utilizarmos **funções matemáticas** específicas a função `I()` torna-se desnecessária:

- `log(y) ~ log(x)` indica: modele o `log(y)` com função estatística da variável `log(x)`);
- `log(y) ~ log(x1^2 * x2)` indica: modele o **log(y)** com função estatística da variável `log(x1^2 * x2)`;

Modelos Plausíveis

Quando temos uma hipótese onde há mais de uma variável preditora, precisamos avaliar, antes de iniciar as análises, quais modelos são plausíveis e relacionados a que hipótese alternativa. Os modelos só são construídos a partir dessa avaliação. Vamos usar o exemplo da videoaula para exemplificarmos os procedimentos e conceitos relacionados a esse tutorial.

Interação entre preditoras

Videoaula do curso **Princípios de Planejamento e Análise de Dados**. Os conceitos abordados são os mesmos desse tutorial, desconsidere referências à disciplina.

A interação é um elemento muito importante quando temos mais de uma preditora, pois desconsiderá-la pode limitar o entendimento dos processos envolvidos. Um exemplo cotidiano da interação é visto no uso de medicamentos e o alerta da bula sobre interação medicamentosa ou efeitos colaterais para pessoas portadoras de doenças crônicas. Dizemos que um medicamento tem interação com outra substância quando o seu efeito é modificado pela presença de outra substância, como por exemplo a ingestão de álcool junto com muitos medicamentos. Nos modelos, a interação tem uma interpretação similar, a resposta pelo efeito de uma variável preditora se altera com a presença de outra preditora.

Delineamentos Experimentais

Expressão	Significado
`y ~ x`	Modele y como função estatística de x
`y ~ x1 + x2`	inclua as variáveis x1 e x2 como preditoras
`y ~ x1 + x2 + x1:x2`	inclua também a interação de x1 com x2
`y ~ x1 * x2`	mesmo que `y ~ x1 + x2 + x1:x2`
`y ~ (x1 + x2 + x3)^2`	Adiciona acima + `x3 + x1:x3 + x2:x3`
`y ~ (x1 + x2 + x3)^3`	Adiciona acima + `x1:x2:x3`
`y ~ (x1 + x2 + x3)^3 - x1:x2`	Retira o termo `x1:x2` da fórmula acima

Simplificando Modelos

Videoaula do curso **Princípios de Planejamento e Análise de Dados**. Os conceitos abordados são os mesmos desse tutorial, desconsidere referências à disciplina.

Videoaula do curso **Planejamento e Análise de Dados**, os conceitos abordados são os mesmos, desconsidere referências à disciplina.

Um dos procedimento de simplificar modelos é partir do modelo cheio e ir simplificando, retirando variáveis preditoras que não ajudam na explicação da variabilidade dos dados. O procedimento consiste em comparar modelos aninhados, dois a dois, retendo o que está mais acoplado aos dados. Caso os modelos não seja diferentes no seu poder explicativo, retemos o modelo mais simples, apoiados no princípio da parcimônia.

Princípio da parcimônia (Navalha de Occam)

- número de parâmetros menor possível
- linear é melhor que não-linear
- reter menos pressupostos
- simplificar ao mínimo adequado
- explicações mais simples são preferíveis

Método do modelo cheio ao mínimo adequado

1. ajuste o modelo máximo (cheio)
2. simplifique o modelo:
    - inspecione os coeficientes (`summary`)
    - remova termos não significativos
3. ordem de remoção de termos:
    - interação não significativos (maior ordem)
    - termos quadráticos ou não lineares
    - variáveis explicativas não significativas
    - verifique se a ordem da retirada de termos de mesmo nível de complexidade influencia a retirada ou manutenção dos termos finais.

Tomada de decisão

A diferença não é significativa:

- retenha o modelo mais simples
- continue simplificando

A difereça é significativa:

- retenha o modelo complexo
- este é o modelo MINÍMO ADEQUADO

Já utilizamos esse procedimento no tutorial 7a. Regressão Linear Simples, quando comparamos o modelo linear com preditora e o modelo sem nenhuma variável preditora.

Peso de bebês ao nascer

Vamos analisar o dado de peso dos bebês ao nascer e como isso se relaciona às características da mãe. Esses dados pode ser consultados em https://www.stat.berkeley.edu/users/statlabs/labs.html.

- baixe o arquivo babies.csv no seu diretório de trabalho
- Vamos selecionar o modelo mínimo adequado a partir das variáveis:
  - resposta **bwt** : peso do bebê ao nascer em onças(oz)
  - preditoras:
    - gestation: tempo de gestação (dias)
    - age: idade
    - weight: peso da mãe
    - smoke: 0 não fumante; 1 fumante

Para simplificar nosso tutorial vamos usar apenas as preditoras: tempo de gestação, idade da mãe e se ela é fumante ou não.

```
bebes <- read.table("babies.csv", header= TRUE, as.is = TRUE, sep= "\t")
str(bebes)
mlfull <- lm(bwt ~ gestation + age + smoke
           + gestation:age + gestation:smoke
         + age: smoke + gestation:age:smoke, data = bebes)
summary(mlfull)
```

```
Call:
lm(formula = bwt ~ gestation + age + smoke + gestation:age +
    gestation:smoke + age:smoke + gestation:age:smoke, data = bebes)

Residuals:
    Min      1Q  Median      3Q     Max
-51.433 -10.647   0.156   9.800  50.994

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.843e+02  5.443e+01   3.385 0.000735 ***
gestation               -2.262e-01  1.938e-01  -1.167 0.243542
age                     -6.010e+00  1.942e+00  -3.095 0.002014 **
smokeTRUE               -1.830e+02  8.188e+01  -2.235 0.025635 *
gestation:age            2.177e-02  6.926e-03   3.143 0.001716 **
gestation:smokeTRUE      6.192e-01  2.934e-01   2.110 0.035056 *
age:smokeTRUE            3.967e+00  2.956e+00   1.342 0.179915
gestation:age:smokeTRUE -1.397e-02  1.061e-02  -1.317 0.187994
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.1 on 1166 degrees of freedom
Multiple R-squared:  0.233, Adjusted R-squared:  0.2284
F-statistic:  50.6 on 7 and 1166 DF,  p-value: < 2.2e-16
```

Interação Tripla

Vamos simplificar o modelo, retirando a interação `gestation:age:smoke` que aparenta não ser importante.

```
ml01 <- lm(bwt ~ gestation + age + smoke
           + gestation:age + gestation:smoke
         + age: smoke, data = bebes)
anova(ml01, mlfull)
summary(ml01)
```

Interações Dupla

Continuamos a simplificação, retirando as interações duplas uma a uma para avaliar quais delas devem ser mantidas. Os testes parciais das variáveis no `summary` nos dá uma indicação de quais devem ser mantidas, mas uma boa prática é fazer o processo completo, já que um elemento no modelo pode mudar o efetividade de outro, principalmente quando compartilham alguma porção de variação explicada.

```
## sem age:smoke
ml02 <- lm(bwt ~ gestation + age + smoke
           + gestation:age + gestation:smoke, data = bebes)
```

```
anova(ml01, ml02)
## sem gestation:smoke
ml03 <- lm(bwt ~ gestation + age + smoke
           + gestation:age  + age:smoke, data = bebes)
anova(ml01, ml03)
## sem gestation:age
ml04 <- lm(bwt ~ gestation + age + smoke
           + gestation:smoke + age: smoke, data = bebes)
anova(ml01, ml04)
```

A única interação dupla que não parece fazer diferença quando retiramos do modelo é a `age:smoke`, as outras explicam uma porção razoável da variação dos dados. Poderíamos continuar simplificando para garantir que não retemos nenhum termo que não é relevante para explicar o peso do bebê ao nascer. Entretanto, a menos que se tenha um bom motivo [1)], **não retiramos os termos das variáveis isoladas quando ela está em algum termo de interação**.

Interpretação do modelo

O `summary` nos fornece as principais informações sobre o modelo mínimo adequado.

```
summary(ml02)
```

```
Call:
lm(formula = bwt ~ gestation + age + smoke + gestation:age +
    gestation:smoke, data = bebes)

Residuals:
    Min      1Q  Median      3Q     Max
-51.978 -10.769   0.108  10.027  50.599

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        135.598062  41.406657   3.275 0.001088 **
gestation           -0.055381   0.147986  -0.374 0.708301
age                 -4.248772   1.458653  -2.913 0.003650 **
smokeTRUE          -75.235972  17.213833  -4.371 1.35e-05 ***
gestation:age        0.015584   0.005224   2.983 0.002911 **
gestation:smokeTRUE  0.239947   0.061676   3.890 0.000106 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.1 on 1168 degrees of freedom
Multiple R-squared:  0.2317, Adjusted R-squared:  0.2284
F-statistic: 70.45 on 5 and 1168 DF,  p-value: < 2.2e-16
```

Uma interpretação importante é com relação a variável `smoke`. Onde foi parar o nível `smokeFALSE`? Como é uma variável categórica de dois níveis, `smoke`foi transformada em variáveis indicadoras e um dos níveis deslocado para o intercepto. O que está representado no intercepto? É a estimativa do modelo para uma mulher que não é fumante com tempo de gestação `zero` e idade `zero`. O que não

faz sentido biológico nenhum.

O intervalo de confiança dos coeficientes é retornado pela função `confint`:

```
(coefml02 <- coef(ml02))
confint(ml02)
```

Interpretação da tabela de Anova em Modelos Multiplos

A função anova aplicada a um único modelo com múltiplas preditoras, nos fornece a comparação de múltiplos modelos na ordem em que as variáveis foram colocadas na fórmula. Vamos interpretar a tabela de anova do nosso modelo:

```
anova(ml02)
```

```
Analysis of Variance Table

Response: bwt
                  Df Sum Sq Mean Sq  F value    Pr(>F)
gestation          1  65450   65450 252.4963 < 2.2e-16 ***
age                1    939     939   3.6241 0.0571933 .
smoke              1  19024   19024  73.3941 < 2.2e-16 ***
gestation:age      1   1964    1964   7.5776 0.0060012 **
gestation:smoke    1   3923    3923  15.1354 0.0001057 ***
Residuals       1168 302757     259
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A segunda linha nos diz que o modelo com `gestação` ao adicionar `age` não explica muita variação a mais. Na terceira linha a comparação é entre os modelos `bwt ~ gestation + age` com o modelo `bwt ~ gestation + age + smoke` a quarta é a comparação deste último com `bwt ~ gestation + age + smoke + gestation:age` e assim por diante, sempre comparando o modelo com tedos os termos anteriores e o que inclui todos os termos anteriores mais o termo que está na linha da tabela. Portanto, se colocarmos termos em outra ordem, as comparações serão outras.

```
ml02b <- lm(bwt ~ age + smoke + gestation + gestation:smoke
           + gestation:age , data = bebes)
anova(ml02b, ml02)
anova(ml02b)
```

```
Analysis of Variance Table

Response: bwt
                  Df Sum Sq Mean Sq  F value    Pr(>F)
age                1    287     287   1.1068 0.2929867
smoke              1  23757   23757  91.6509 < 2.2e-16 ***
gestation          1  61370   61370 236.7568 < 2.2e-16 ***
smoke:gestation    1   3580    3580  13.8130 0.0002115 ***
```

```
age:gestation      1   2307    2307   8.9001 0.0029108 **
Residuals       1168 302757     259
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Só para entendermos o que está apresentado nessa anova, vamos comparar os modelos:

- 1. bwt ~ age + smoke + gestation
- 2. bwt ~ age + smoke + gestation + smoke:gestation

```
ml05 <-lm(bwt ~ age + smoke + gestation, data = bebes)
ml06 <-lm(bwt ~ age + smoke + gestation + gestation:smoke, data = bebes)
anova(ml05, ml06)
```

Pode haver pequenas variações nos valores por conta arredondamentos. O importante aqui é que um termo pode ser significativo ou não dependendo da ordem que for colocado, principalmente se há alguma colinearidade entre as variáveis incluídas. Ou seja, o termo que é colocado antes explica a variação que o termo que vem depois poderia explicar também!

Diagnóstico do modelo

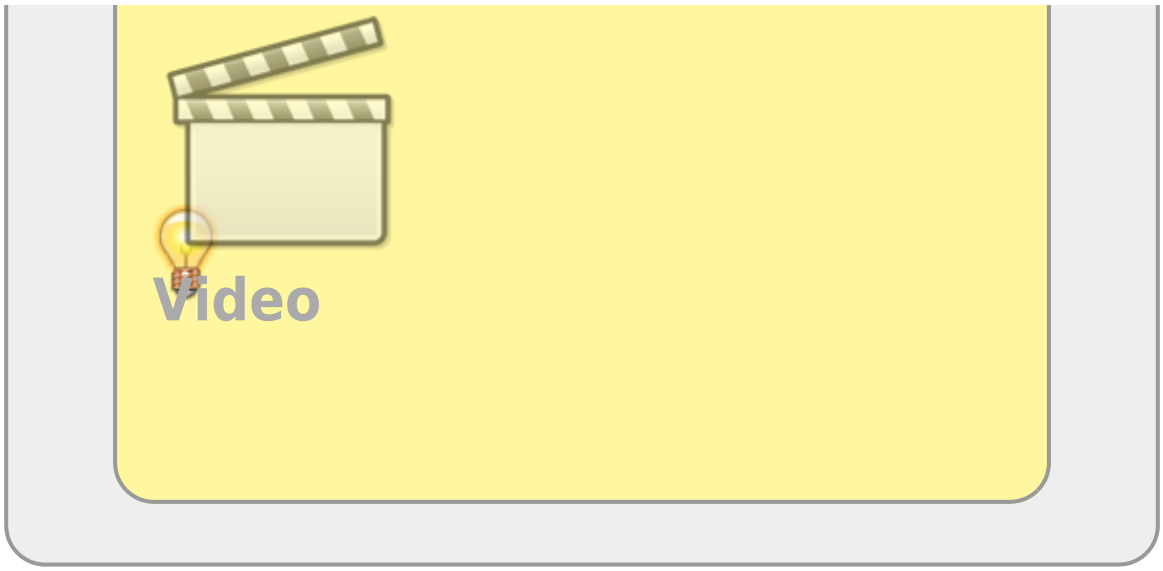
O diagnóstico das premissas do modelo é importante, para mais informações veja o tutorial da disciplina Princípios de Planejamento e Análise de Dados sobre o assunto. O basico pode ser interpretado nos gráficos que são feitos por padrão se usamos a função `plot` no objeto de modelo:

```
par(mfrow = c(2,2), mar=c(4,4,2,2), cex.lab=1.2,
    cex.axis=1.2, las=1,  bty="n")
plot(ml02)
```

Estando tudo certo com nosso modelo podemos passar para outras fases como preparar gráficos e interpretar os resultados.

Videoaula Síncrona

Aula síncrona gravada pelo Google Meet em 05 de outubro de 2020



1)

desenhos experimentais aninhados podem incluir a variável aninhada apenas na interação