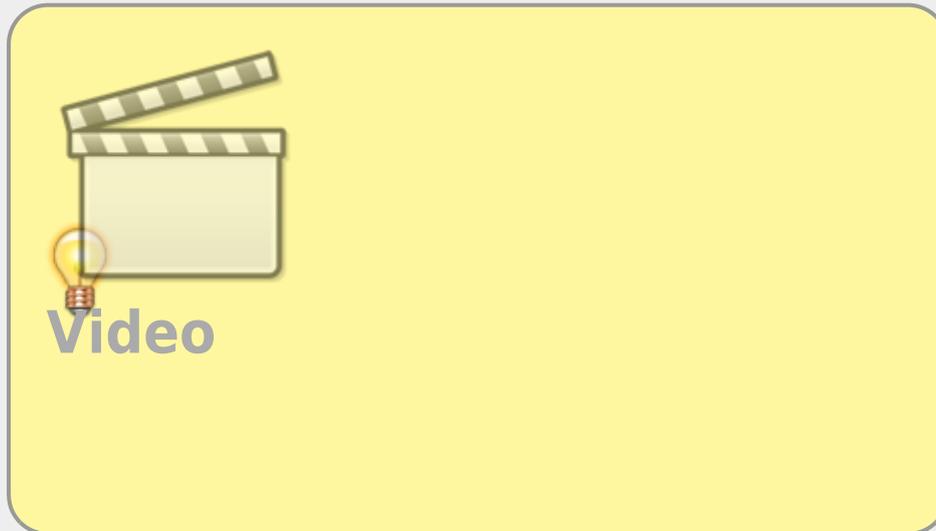


- [Tutorial](#)
- [Exercícios](#)

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.



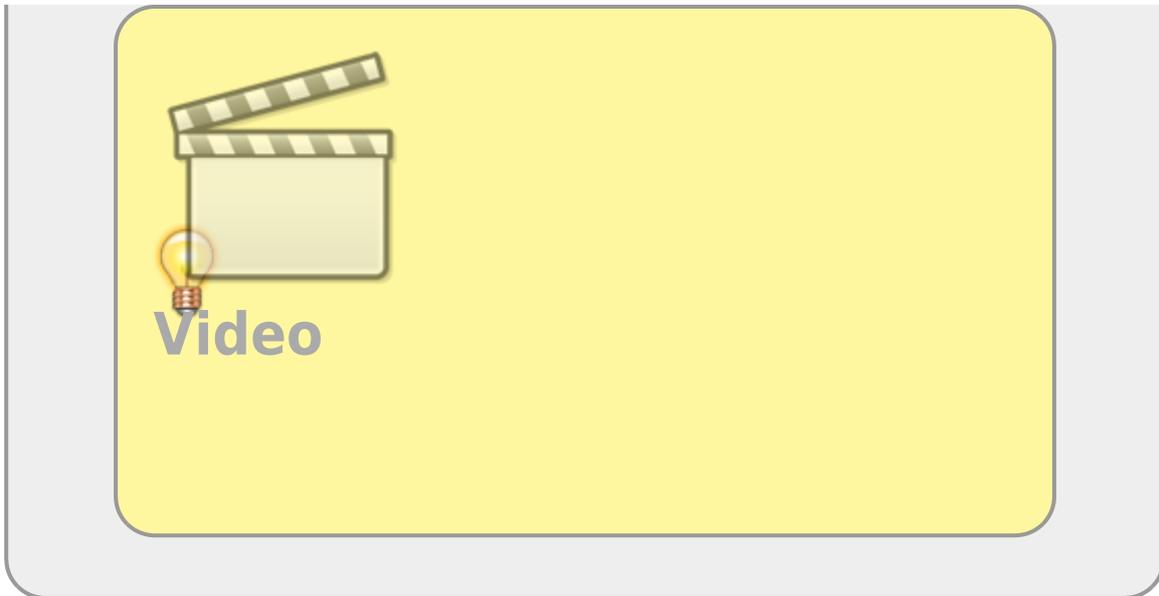
Os modelos lineares permitem que sejam incluídas mais do que uma variável preditora, como fizemos até aqui. Nesse tutorial vamos aprender alguns princípios básicos desses modelos mais complexos. Em modelos com mais de uma variável preditora, precisamos tomar a decisão de quais variáveis devemos reter 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.

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.

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.

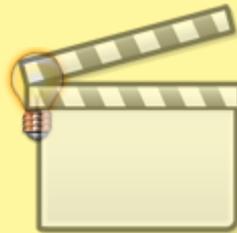
Alguns aspectos básicos do argumento:

- ' $y \sim x$ ' indica: *construa o modelo da variável resposta y como **função estatística linear** de x* ; * ' $y \sim x1 + x2$ ' 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 utilizar a função ' $I()$ ': * ' $y \sim I(x1^2 * x2^3)$ ' indica: modele y como função estatística **da variável** $(x1^2 * x2^3)$; * ' $y \sim I(x1 / x2)$ ' indica: modele y como função estatística **da variável** $(x1/x2)$; No caso de utilizarmos **funções matemáticas** específicas a função ' $I()$ ' torna-se desnecessária: * ' $\log(y) \sim \log(x)$ ' indica: modele o **log(y)** com função estatística da variável **log(x)**; * ' $\log(y) \sim \log(x1^2 * x2)$ ' indica: modele o **log(y)** com

função estatística da variável $\log(x_1^2 * x_2)$;
</WRAP> ===== 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.



Video

Um dos procedimentos 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 sejam 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 diferença 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. =====
===== 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 ¹⁾. `rsplus` `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) </code> 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 </code> ===== Interação Tripla =====`

Vamos simplificar o modelo, retirando a interação `gestation:age:smoke` que aparenta não ser importante. `rsplus` `ml01 ← lm(bwt ~ gestation + age + smoke + gestation:age + gestation:smoke + age: smoke, data = bebes) anova(ml01, mlfull) summary(ml01) </code> ===== 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 a efetividade de outro, principalmente quando compartilham alguma porção de variação explicada.

```
rsplus> ## 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) </code>
```

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. ===== Interpretação do modelo ===== O summary nos fornece as principais informações sobre o modelo mínimo adequado.

```
rsplus> summary(ml02)
confint(ml02) anova(ml02) </code>
```

===== Diagnóstico do modelo =====

```
rsplus> 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)
</code>
```

Os dados desse estudo serão usados também no exercício, porém lá, vamos partir dos dados brutos com mais variáveis

1)

no exercício terão que usar os dados brutos e todas as variáveis

From:
<http://ecor.ib.usp.br/> - ecorR

Permanent link:
http://ecor.ib.usp.br/doku.php?id=02_tutoriais:tutorial7b:start&rev=1601737647

Last update: 2020/10/03 12:07

