Modelos Lineares Múltiplos

Simplificando Modelos

Durante o curso usaremos o procedimento de simplificar o modelo a partir do modelo cheio. 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
    - agrupe níveis de fatores sem diferença
    - ANCOVA: intercepto não significativos → 0

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

Interação entre preditoras

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.

Simulando um experimento plausível

Vimos que existe um efeito do tipo de solo na produção de um cultivar no exemplo de ANOVA. Uma expectativa plausível é que a adição de adubo também tenha efeito na produtividade e modifique o efeito do solo. Esse é nosso próximo exemplo. Para ele vamos usar uma simulação de dados similar ao que fizemos no modelo linear simples.

Nos dados originais do exercício de ANOVA a produtividade média nos solos foi de:

- arenoso: 9.9
- argiloso: 11.5
- humico: 14.3

Vamos, a partir dessa informação, criar um experimento onde, além da diferença do solo, metade dos cultivos foram tratados com adubo orgânico.

</WRAP>

- 1. Criamos vetores para representar as variáveis solo e adubo.
- 2. Para cada observação incluímos o efeito médio de produtividade de cada solo (10 réplicas para cada solo)
- 3. Associamos um valor de efeito do tratamento adubo, como:
    - arenoso: + 2.7
    - argiloso: + 0.7
    - humico: + 0.2
- 4. Em seguida somamos um efeito aleatório na resposta para criar um data frame com as variáveis preditoras e resposta.

```
solo <- rep(c("are", "arg", "hum"), each=10)
adubo <- rep(rep(c(FALSE, TRUE), each=5), 3)
meansolo <- rep(c(9.9, 11.5, 14.3), each=10)
efeitoadubo <- rep(c(0, 2.7, 0, 0.7, 0, 0.2), each=5)
residuo <- rnorm(30, 0, 1)
dadosolo <- data.frame(solo, adubo, prod = meansolo + efeitoadubo + residuo)
str(dadosolo)
```

Confira se o objeto `dadosolo` foi organizado corretamente

Gráfico dos dados

Agora um gráfico simples. Busque entender todos os argumentos das funções abaixo.

```
par( mar=c(4,4,2,2),   cex.lab=1.5, cex.axis=1.2, las=1, bty="n")
boxplot(prod ~  adubo + solo, data = dadosolo, ann= FALSE, xaxt= "n",
outline= FALSE, col = rep(c(rgb(0,0,0, 0.1),rgb(0,0,0, 0.5)), 3)  )
mtext(c("arenoso", "argiloso", "húmico"), side = 1, at= c(1.5, 3.5, 5.5),
line = 1, cex = 2)
legend("bottomright", legend= c("sem", "com"),title = "Adubo", bty= "n", pch
= 15, cex = 1.5,col = c(rgb(0,0,0, 0.1),rgb(0,0,0, 0.5)))
```

Modelo Cheio

Abaixo construímos o modelo cheio com as variáveis adubo e tipo de solo.

```
soloFull <- lm(prod ~ adubo + solo + solo:adubo, data = dadosolo)
summary(soloFull)
```

Modelo sem Interação

A primeira simplificação possível é retirar o efeito da interação entre as preditoras e comparar com o modelo cheio.

```
solo01 <- lm(prod ~ adubo + solo , data = dadosolo)
anova(solo01, soloFull)
```

O resultado nos indica que o modelo cheio é o modelo mínimo adequado. Ou seja, explica uma porção considerável da variação dos dados a mais que o modelo mais simples, sem a interação entre tipo de solo e adubo. Para completar, vamos fazer a comparação com o modelo nulo. Essa comparação pode ser feito de duas maneiras: (1) construindo o modelo nulo e comparando por anova, ou (2) interpretando a tabela de anova do modelo mínimo adequado.

```
solo00 <- lm(prod ~ 1 , data = dadosolo)
anova(solo00, soloFull)
anova(soloFull)
```

Diagnóstico

O passo final é investigar se o modelo cumpre com as premissas do modelo linear.

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

Não poderia ser mais comportado. Isso significa que criamos os dados corretamente!! Agora é a parte mais difícil e interessante de qualquer análise de dados, a interpretação biológica suscita do resultado!

Cabeçalho

Interpretando Variáveis Indicadoras (Dummy)

As variáveis indicadoras devem ser interpretadas com cuidado. No exemplo acima, o modelo pode ser descrito da seguinte forma:

$$ y_{tr} = \alpha + \beta_1 * arg + \beta_2 * hum + \beta_3 * adubo + \beta_4 * arg * adubo + \beta_5 * hum * adubo $$

As variáveis arg, hum e adubo são dummy ou indicadoras, representadas por 1 quando presente e 0 quando ausentes. α, β_i representam as estimativas do modelo e estão relacionados, nesse caso, ao efeito de cada tratamento.

Para calcular o valor predito para o tratamento no solo arenoso com adubo, temos:

$$ y_{arenAdubo} = \alpha + \beta_3 * adubo $$

Isso em decorrência do tratamento **arenoso sem adubo** estar representado pelo intercepto (α) do modelo.

Para o tratamento de solo **argiloso com adubo** o predito é:

$$ y_{argAdubo} = \alpha + \beta_1 * arg + \beta_3 * adubo + \beta_4 * arg * adubo $$

E assim por diante, usando as variáveis indicadoras e os coeficientes estimados para o cálculo do predito pelo modelo.

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 exercício 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)
```

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.

Interpretação do modelo

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

```
summary(ml02)
confint(ml02)
anova(ml02)
```