

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, reterendo 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:
 - interações não significativas (primeiro as de maior ordem)
 - termos quadráticos ou não lineares
 - variáveis explicativas não significativas
 - agrupe níveis de fatores sem diferença
 - ANCOVA: intercepto não significativos $\rightarrow 0$

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

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.

- 1. Abra a planilha

colheita}}preservefilenames::crop2Anova.csv

, no excel:



Untitled 1 - LibreOffice Calc

File Edit View Insert Format Styles Sheet Data Tools Window

DejaVu Sans 10

A1 Σ = solo

	A	B	C	D	E	F
1	solo	adubo	prodMedia	residuo	prodCampo	
2	arenoso	nao	9.9			
3	arenoso	nao	9.9			
4	arenoso	nao	9.9			
5	arenoso	nao	9.9			
6	arenoso	nao	9.9			
7	arenoso	sim				
8	arenoso	sim				
9	arenoso	sim				
10	arenoso	sim				
11	arenoso	sim				
12	argiloso	nao	11.5			
13	argiloso	nao	11.5			
14	argiloso	nao	11.5			
15	argiloso	nao	11.5			
16	argiloso	nao	11.5			
17	argiloso	sim				
18	argiloso	sim				
19	argiloso	sim				
20	argiloso	sim				
21	argiloso	sim				
22	humico	nao	14.3			
23	humico	nao	14.3			
24	humico	nao	14.3			
25	humico	nao	14.3			
26	humico	nao	14.3			
27	humico	sim				
28	humico	sim				
29	humico	sim				

Sheet1

Find Find All Formatted Display

Sheet 1 of 1 Default English (USA)

- 2. Preencha as médias de produtividade no tratamento com adubo com o seguintes valores:
 - arenoso: + 2.7
 - argiloso: + 0.7
 - humico: + 0.2
- 3. Preencha a célula **D2** da coluna resíduo com a fórmula = **INV.NORM.N(ALEATÓRIO(); 0 ; 1)**¹⁾;

Ao final sua planilha deve estar preenchida como a que segue, apenas com os valores da coluna resíduo diferentes:

crop2AnovaFim.csv - LibreOffice Calc

File Edit View Insert Format Styles Sheet Data

DejaVu Sans 10

A1 Σ = solo

	A	B	C	D	E
1	solo	adubo	prodMedia	residuo	prodCampo
2	arenoso	nao	9.9	1.35	11.25
3	arenoso	nao	9.9	-1.81	8.09
4	arenoso	nao	9.9	-0.88	9.02
5	arenoso	nao	9.9	0.22	10.12
6	arenoso	nao	9.9	1.28	11.18
7	arenoso	sim	12.6	-1.11	11.49
8	arenoso	sim	12.6	-1.51	11.09
9	arenoso	sim	12.6	-0.14	12.46
10	arenoso	sim	12.6	0.49	13.09
11	arenoso	sim	12.6	-0.22	12.38
12	argiloso	nao	11.5	1.35	12.85
13	argiloso	nao	11.5	0.12	11.62
14	argiloso	nao	11.5	-0.03	11.47
15	argiloso	nao	11.5	-1.17	10.33
16	argiloso	nao	11.5	0.17	11.67
17	argiloso	sim	11.9	1.23	13.13
18	argiloso	sim	11.9	0.73	12.63
19	argiloso	sim	11.9	0.2	12.1
20	argiloso	sim	11.9	0.08	11.98
21	argiloso	sim	11.9	-1.82	10.08
22	humico	nao	14.3	0.09	14.39
23	humico	nao	14.3	1.1	15.4
24	humico	nao	14.3	-2.01	12.29
25	humico	nao	14.3	0.98	15.28
26	humico	nao	14.3	-1.06	13.24
27	humico	sim	14.5	-0.28	14.22
28	humico	sim	14.5	-1.19	13.31
29	humico	sim	14.5	1.29	15.79

crop2AnovaFim

ATIVIDADE

- 1. Abra a planilha preenchida no Rcmdr
- 2. Monte o modelo linear com as preditoras e sua interação (**solo**, **adubo** e **solo:adubo**) para a resposta **prodCampo**
- 3. Interprete os valores dos coeficientes do modelo (summary) e calcule, a partir deles, os valores preditos pelo modelo para cada um dos tratamentos
- 4. Compare os preditos com relação aos valores utilizados para gerar os dados de **prodCampo**
- 5. Preencha a [planilha google](#) com os resultados do modelo

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.

Interação entre preditoras: contínua e categórica

1. abra o arquivo

davis.csv

no R ou no Rcmdr

2. monte o modelo para peso com relação às variáveis preditoras: altura, sexo e suas interação.
3. a partir do modelo mais cheio, simplifique o modelo até o mínimo adequado
4. apresente o resultado em um gráfico

Peso de bebês ao nascer

- abra o arquivo `babies.csv`
no R ou Rcmdr
- faça um modelo plausível com as variáveis:
 - resposta **bwt** : peso do bebê ao nascer
 - preditoras:
 - gestation: tempo de gestação (dias)
 - age: idade
 - weight: peso
 - smoke: 0 não fumante; 1 fumante
 - interações: até a 3 ordem para aquelas que são plausíveis e passíveis de interpretação biológica
 - selecione o modelo mínimo plausível pelo método de simplificação para mínimo adequado
 - interprete o resultado

Exercício Extra

Um artigo recente (Fisher, R. & Ai C. 2018)¹⁾ sobre métodos de regressões múltiplas, apresenta dados sobre peixes de recifes de corais. Entre as questões apresentadas pelos autores originais do trabalho estava se a biomassa de diferentes guildas de peixes em zonas protegidas ou não de recifes. Aqui vamos usar apenas a guilda de peixes que se alimentam de plânctons e apenas parte das preditoras. O método desenvolvido no artigo é uma forma de automatizar a seleção de preditoras em modelos com muitas variáveis potenciais de influenciar a resposta.

- faça uma análise exploratória das variáveis do dados `plankivore.csv`
e suas relações;
- monte o modelo mínimo adequado, partindo de todas as variáveis preditoras e suas interações;
- verifique se não há nenhuma variável com relação não linear com a biomassa, e se houver use mais uma variável representada pelo quadrado dessa variável;
- ao final faça o diagnóstico do modelo e veja se o resíduo cumprem com as premissas do modelo linear;
- caso diagnostique problema no modelo, transforme a variável resposta usando o logaritmo natural;
- refaça a seleção e o diagnóstico com a resposta na escala log;
- interprete o resultado.

Fisher, R.; Ai C. L. et al. 2018. A simple function for full-subsets multiple regression in ecology with R. *Ecology and Evolution* 8: 6104-6113}}preservefilenames::Fisher_et_al-2018-Ecology_and_Evolution.pdf

¹⁾

Essa expressão retorna valores associados a uma distribuição normal com média 0 e desvio padrão 1

2)

note que temos dois homônimos aqui, o mais famoso estatístico e um critério de seleção de modelos juntos, como autores!! Os outros nomes foram omitidos propositalmente.

From:

<http://labtrop.ib.usp.br/> - **Laboratório de Ecologia de Florestas Tropicais**

Permanent link:

<http://labtrop.ib.usp.br/doku.php?id=cursos:planeco2019:roteiro:09-lm02>



Last update: **2019/12/11 12:31**