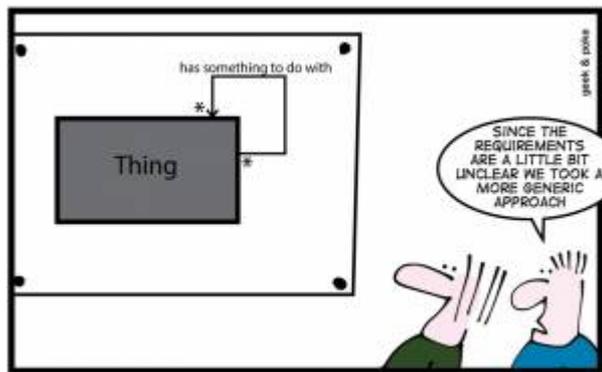


Modelos Lineares Múltiplos I



HOW TO CREATE A STABLE DATA MODEL

Uma extensão do modelo linear simples ¹⁾ são os modelos lineares com mais de uma preditora, aqui definido como modelos múltiplos. Quando temos mais de uma preditora o modelo aumenta em complexidade com mais parâmetros para estimar. Além disso, a estrutura mais complexa do modelo gera desafios para a interpretação e dificulta a avaliação da adequação do modelo aos dados. Uma primeira complexidade está relacionada a como simplificar a estrutura do modelo com a finalidade de facilitar a interpretação e melhorar a estimação dos parâmetros. A tomada de decisão sobre quais variáveis devemos reter em nosso modelo e quais podem ser retiradas, por não terem efeito na variável resposta, pode ser feita utilizando diferentes critérios e técnicas. A seguir apresentamos uma das técnicas utilizadas para essa tomada de decisão e que iremos utilizar ao longo desse curso. Outros critérios ou técnicas podem ser utilizadas com vantagens ou desvantagens em relação ao que utilizaremos. Não é objetivo desse curso se debruçar sobre essas diferentes técnicas.



Video

Duas preditoras categóricas

O primeiro exemplo que iremos trabalhar é baseado nos dados utilizados para exemplificar o [teste de Anova](#). Vamos criar um experimento plausível a partir dele.

Simulando um experimento plausível

Vimos que existe um efeito do tipo de solo na produção de um cultivar. Uma expectativa plausível é

que a adição de adubo também tenha efeito na produtividade. Ou seja, os tipos de solo tem produtividade diferente, assim como o adubo aumenta a produtividade.

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

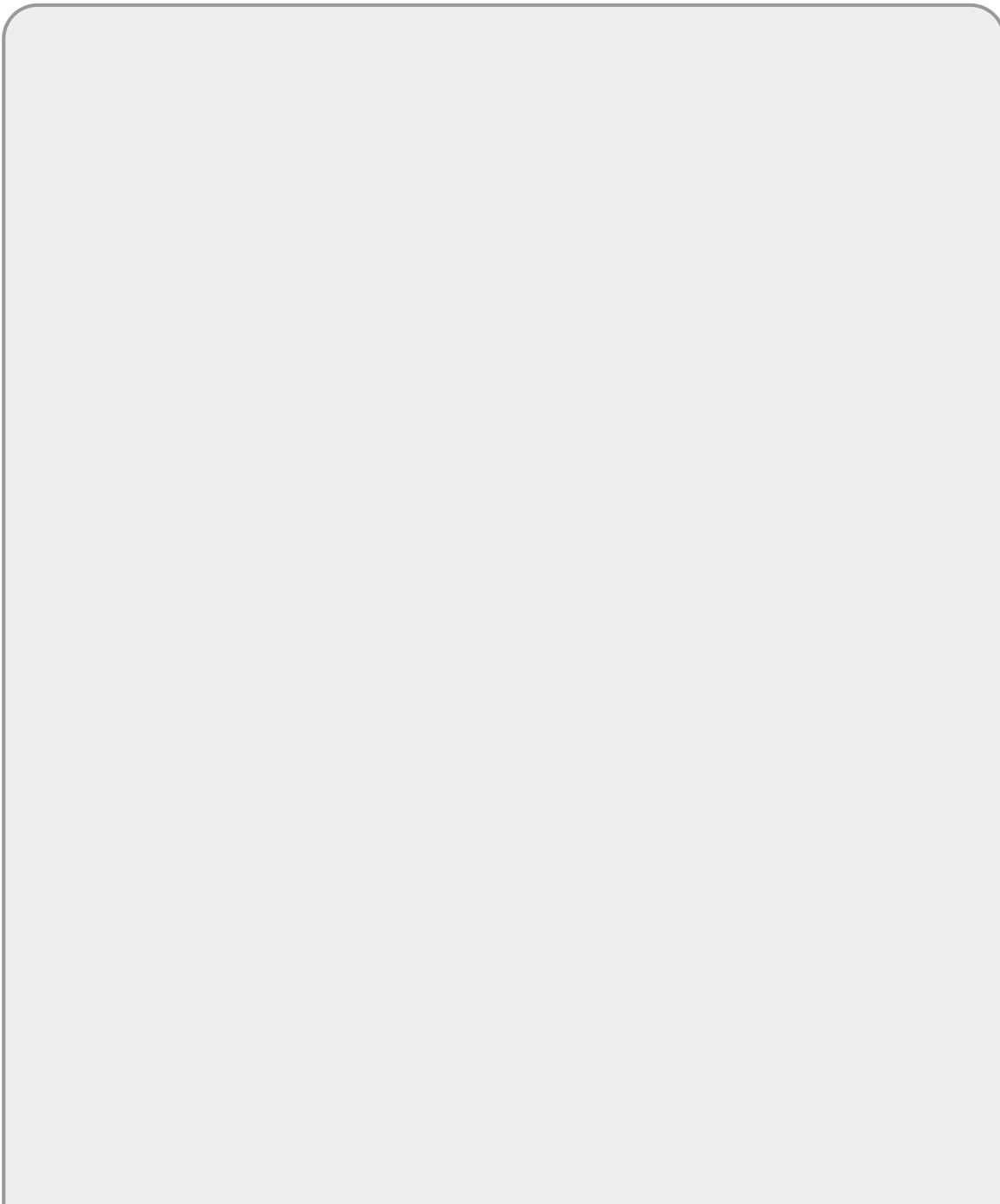
- arenoso: 9.9
- argiloso: 11.5
- humico: 14.3

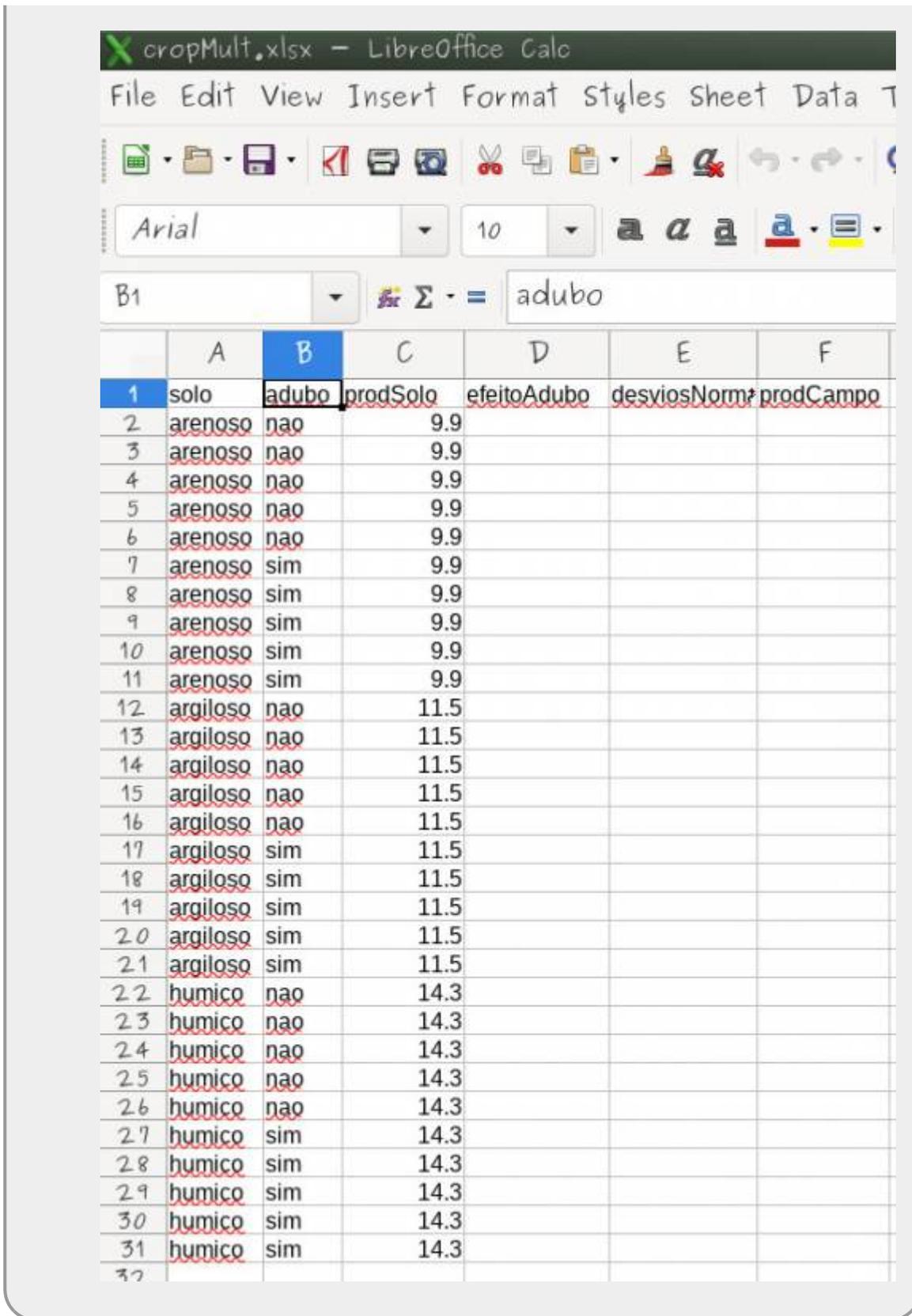
Vamos, a partir dessa informação, criar um experimento onde, além da diferença do solo, metade dos cultivos foram tratados com adubo orgânico.

- 1. Abra o arquivo

`cropMulti}}preservefilenames::cropMult.xlsx`

em uma planilha eletrônica:

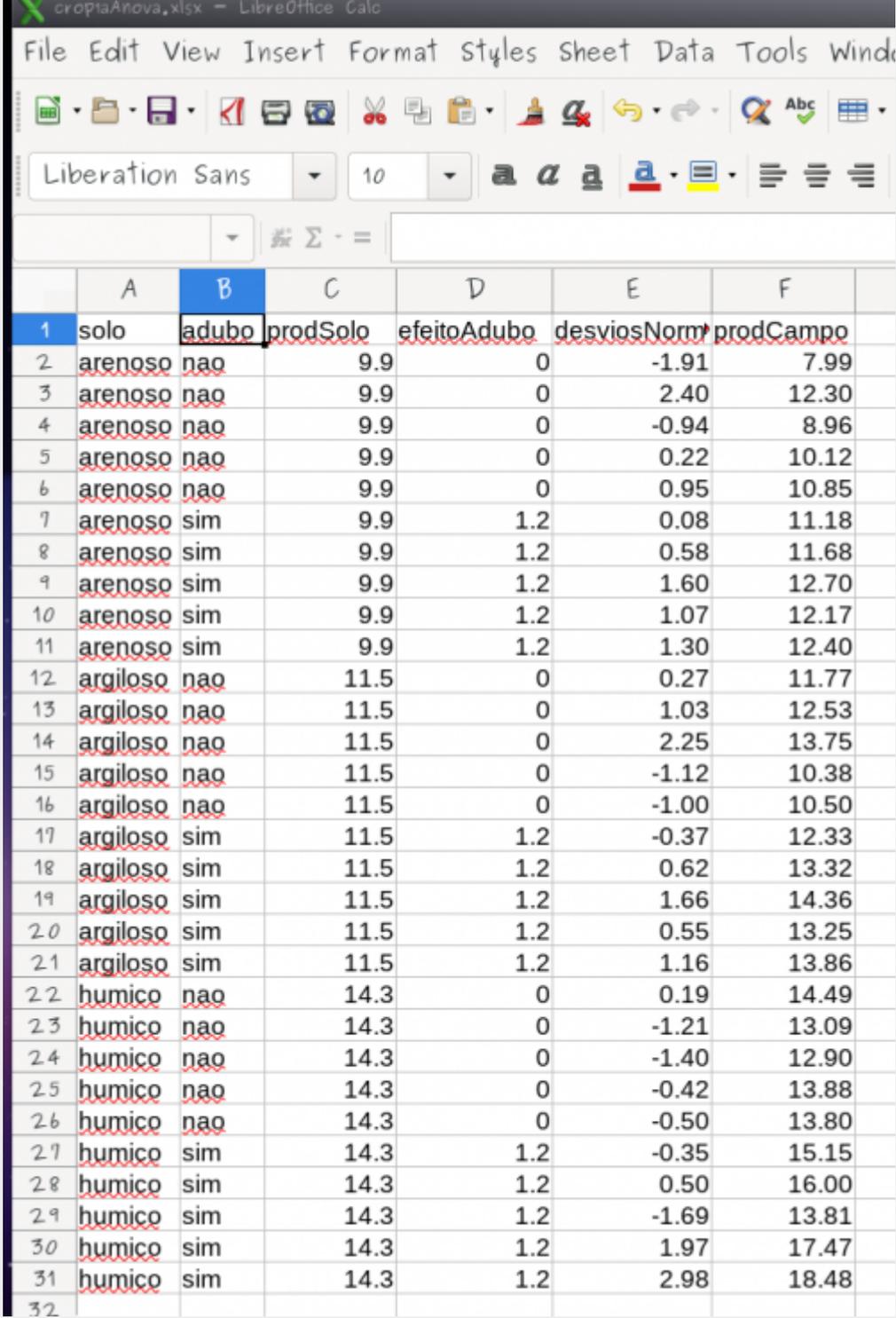




- 2. Preencha a coluna efeitoAdubo com o valor de 1.2 para todas as parcelas adubadas ²⁾ e 0 para aquelas que não foram ³⁾.
- 3. Preencha a célula E2 da coluna desvios normal com a fórmula = **INV.NORM.N(ALEATÓRIO(); 0 ; 1.5)** ⁴⁾.
- 4. Some os valores em uma mesma linha

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

resíduo diferentes:



The screenshot shows a LibreOffice Calc spreadsheet with the following data:

	A	B	C	D	E	F
1	solo	adubo	prodSolo	efeitoAdubo	desviosNorm	prodCampo
2	arenoso	nao	9.9	0	-1.91	7.99
3	arenoso	nao	9.9	0	2.40	12.30
4	arenoso	nao	9.9	0	-0.94	8.96
5	arenoso	nao	9.9	0	0.22	10.12
6	arenoso	nao	9.9	0	0.95	10.85
7	arenoso	sim	9.9	1.2	0.08	11.18
8	arenoso	sim	9.9	1.2	0.58	11.68
9	arenoso	sim	9.9	1.2	1.60	12.70
10	arenoso	sim	9.9	1.2	1.07	12.17
11	arenoso	sim	9.9	1.2	1.30	12.40
12	argiloso	nao	11.5	0	0.27	11.77
13	argiloso	nao	11.5	0	1.03	12.53
14	argiloso	nao	11.5	0	2.25	13.75
15	argiloso	nao	11.5	0	-1.12	10.38
16	argiloso	nao	11.5	0	-1.00	10.50
17	argiloso	sim	11.5	1.2	-0.37	12.33
18	argiloso	sim	11.5	1.2	0.62	13.32
19	argiloso	sim	11.5	1.2	1.66	14.36
20	argiloso	sim	11.5	1.2	0.55	13.25
21	argiloso	sim	11.5	1.2	1.16	13.86
22	humico	nao	14.3	0	0.19	14.49
23	humico	nao	14.3	0	-1.21	13.09
24	humico	nao	14.3	0	-1.40	12.90
25	humico	nao	14.3	0	-0.42	13.88
26	humico	nao	14.3	0	-0.50	13.80
27	humico	sim	14.3	1.2	-0.35	15.15
28	humico	sim	14.3	1.2	0.50	16.00
29	humico	sim	14.3	1.2	-1.69	13.81
30	humico	sim	14.3	1.2	1.97	17.47
31	humico	sim	14.3	1.2	2.98	18.48
32						

Procedimentos



- 1. Salve a planilha e abra os dados no Rcmdr.
- 2. Produza um modelo chamado mlSolo_Adubo da seguinte forma:

prodCampo ~ solo + adubo



- 3. Avalie o modelo pelo seu sumário e pela tabela de Anova

Modelos Plausíveis

O nosso modelo tem duas preditoras e pode ser simplificado. Nesse caso, como temos poucas possibilidades de comparação, podemos comparar todas os modelos plausíveis. O que produzimos acima tem o efeito de solo e de adubo, podemos pensar em mais algumas possibilidades de modelo:

- **mlSolo** só com o efeito do solo:

prodCampo ~ solo

- **mlAdubo** só com efeito do adubo:

prodCampo ~ adubo

- **mlNull** sem efeito de solo ou adubo:

prodCampo ~ 1

O valor 1 na última formula indica que o modelo não tem nenhuma variável preditora ⁵⁾



caixa de sugestão

Interação entre preditoras



Video

Na comparação acima, desconsideramos um elemento importante que emerge quando temos mais de uma preditora, a possibilidade de uma variável preditora interferir no efeito de outra, efeito esse chamado de interação. 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. Muitas vezes a interação pode ser o efeito de interesse do estudo, como na pergunta: *O efeito do tamanho do fragmento na diversidade de plantas dependendo do seu grau de isolamento?*



- Na comparação de modelos acima, não incluímos o termo da interação. Produza o modelo abaixo incluindo o termo da interação e compare esse modelo com o selecionado no procedimento acima.

prodCampo ~ solo + adubo + solo:adubo

Não é esperado encontrar interação entre as preditoras nos dados simulados da maneira como fizemos, ele pode emergir apenas por acaso, apenas porque temos uma variável aleatória ⁶⁾. Da maneira como simulamos os dados temos duas preditoras que tem efeitos aditivos onde não há interação. Uma outra forma de dizer isso é que o efeito do adubo não interfere no efeito do solo, ou que esses efeitos são independentes. A interpretação biológica nesse caso também pode ser feita independentemente.

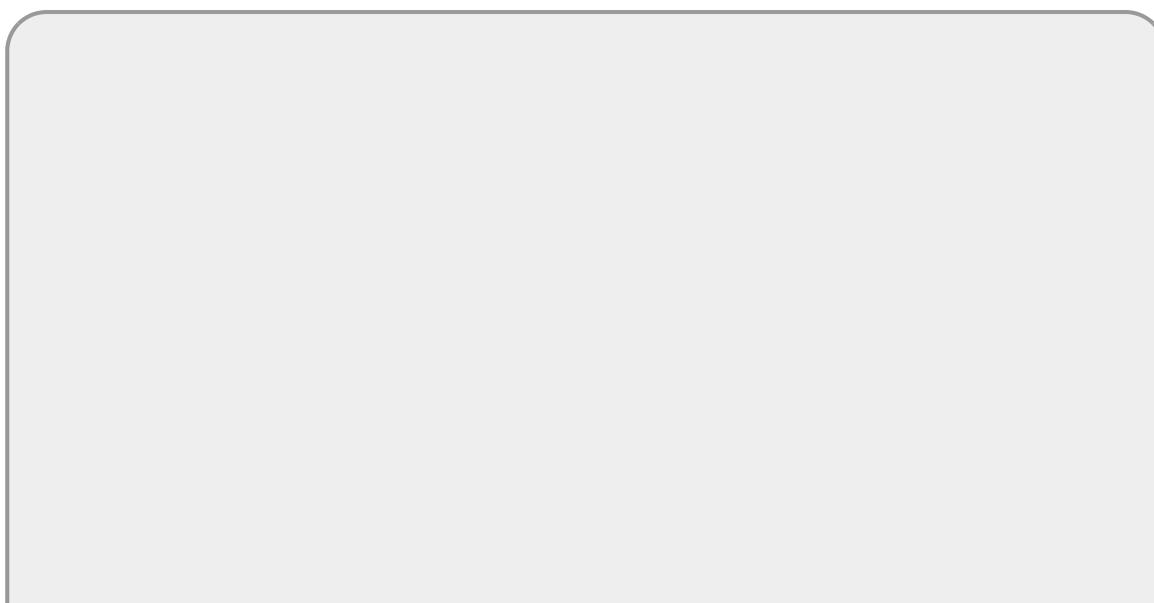
Simulando dados com interação

Seguindo a mesma abordagem anterior, vamos produzir dados simulando a interação entre as variáveis solo e adubo. Para isso precisamos produzir dados em que o efeito do adubo depende do tipo de solo.

1. Abra o arquivo

cropMulti}}preservefilenames::cropMult.xlsx

em uma planilha eletrônica:



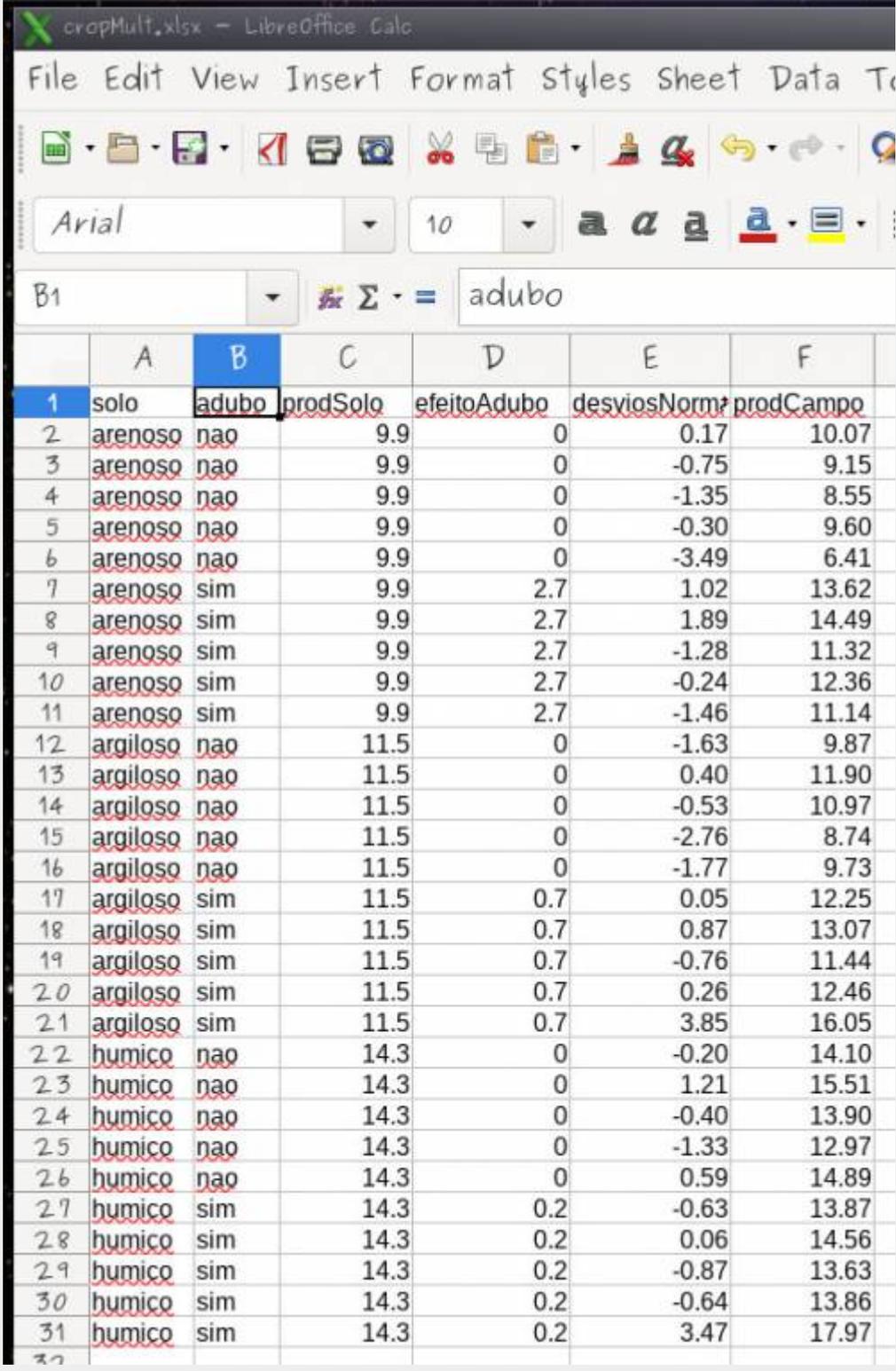
The screenshot shows a LibreOffice Calc spreadsheet with the following data:

	A	B	C	D	E	F
1	solo	adubo	prodSolo	efeitoAdubo	desviosNorm	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	9.9			
8	arenoso	sim	9.9			
9	arenoso	sim	9.9			
10	arenoso	sim	9.9			
11	arenoso	sim	9.9			
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	11.5			
18	argiloso	sim	11.5			
19	argiloso	sim	11.5			
20	argiloso	sim	11.5			
21	argiloso	sim	11.5			
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	14.3			
28	humico	sim	14.3			
29	humico	sim	14.3			
30	humico	sim	14.3			
31	humico	sim	14.3			
32						

- Preencha a coluna efeitoAdubo com os valores:
 - 2.7 para arenoso com adubo igual a sim
 - 0.7 para argiloso com adubo igual a sim
 - 0.2 para humico com adubo igual a sim
- O campos da coluna efeitoAdubo onde adubo é igual a não devem ser preenchidos com 0
- Preencha a célula **E2** da coluna desvios normal com a fórmula = **INV.NORM.N(ALEATÓRIO(); 0 ; 1.5)⁷⁾**.

4. Some na coluna prodCampo os valores prodSolo + efeitoAdubo + desviosNormal

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



The screenshot shows a spreadsheet with the following data:

	A	B	C	D	E	F
1	solo	adubo	prodSolo	efeitoAdubo	desviosNormal	prodCampo
2	arenoso	nao	9.9	0	0.17	10.07
3	arenoso	nao	9.9	0	-0.75	9.15
4	arenoso	nao	9.9	0	-1.35	8.55
5	arenoso	nao	9.9	0	-0.30	9.60
6	arenoso	nao	9.9	0	-3.49	6.41
7	arenoso	sim	9.9	2.7	1.02	13.62
8	arenoso	sim	9.9	2.7	1.89	14.49
9	arenoso	sim	9.9	2.7	-1.28	11.32
10	arenoso	sim	9.9	2.7	-0.24	12.36
11	arenoso	sim	9.9	2.7	-1.46	11.14
12	argiloso	nao	11.5	0	-1.63	9.87
13	argiloso	nao	11.5	0	0.40	11.90
14	argiloso	nao	11.5	0	-0.53	10.97
15	argiloso	nao	11.5	0	-2.76	8.74
16	argiloso	nao	11.5	0	-1.77	9.73
17	argiloso	sim	11.5	0.7	0.05	12.25
18	argiloso	sim	11.5	0.7	0.87	13.07
19	argiloso	sim	11.5	0.7	-0.76	11.44
20	argiloso	sim	11.5	0.7	0.26	12.46
21	argiloso	sim	11.5	0.7	3.85	16.05
22	humico	nao	14.3	0	-0.20	14.10
23	humico	nao	14.3	0	1.21	15.51
24	humico	nao	14.3	0	-0.40	13.90
25	humico	nao	14.3	0	-1.33	12.97
26	humico	nao	14.3	0	0.59	14.89
27	humico	sim	14.3	0.2	-0.63	13.87
28	humico	sim	14.3	0.2	0.06	14.56
29	humico	sim	14.3	0.2	-0.87	13.63
30	humico	sim	14.3	0.2	-0.64	13.86
31	humico	sim	14.3	0.2	3.47	17.97



Procedimentos

1. Salve a planilha com **um nome diferente** para não sobrescrever a planilha de dados usada anteriormente e importe os dados para o Rcmdr. **Atenção nomeie os dados na aba de importação com um nome diferente dos dados importados anteriormente, em alguns casos o Rcmdr não importa se a planilha e os dados importados tiverem o mesmo nome de uma importação anterior**
2. Produza o modelo cheio `mlSolo_AduboAll` com a seguinte fórmula:
 - `prodCampo ~ solo + adubo + solo:adubo`
3. Faça a seleção do modelo mínimo adequado utilizando o procedimentos de comparação de modelo pela partição de variância
4.  Avalie o modelo selecionado pelo sumário e pela tabela de Anova. Reconheça os valores utilizados para gerar os dados a partir das estimativas do modelo.
5. Preencha a aba `cropInt2020` da planilha [lmCrop2pred](#) com os resultados do modelo selecionado
6. Na planilha onde os dados foram gerados, calcule, a partir dos coeficientes estimados, os valores preditos pelo modelo para cada um dos tratamentos, coloque esses valores em uma coluna nomeada de `predito`, veja como calcular os valores preditos no quadro [interpretando_variáveis_indicadoras_\(dummy\)](#) abaixo
7. Calcule os resíduos do modelo ⁸⁾ em uma coluna denominada `residuos`
8. Eleve o valor dos resíduos ao quadrado em uma coluna denominada `resQuad`
9. Some o a coluna em `resQuad` para calcular a soma quadrática do modelo
10. Calcule o R^2 do modelo, baseado no `resQuad` ⁹⁾

Interpretando Variáveis Indicadoras (Dummy)

As variáveis indicadoras devem ser interpretadas com cuidado. No exemplo do modelo cheio acima ¹⁰⁾, 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.

Simplificando Modelos



Video

Durante o curso usaremos o procedimento de simplificar o modelo a partir do modelo cheio. O procedimento consiste em comparar modelos aninhados¹¹⁾, dois a dois, retendo o que está mais acoplado aos dados. Para comparar os modelos utilizaremos o procedimento da partição da variância baseado na tabela de anova. O modelo que explica mais variação dos dados é retido e quando os modelos não são diferentes no seu poder explicativo, retemos o modelo mais simples, apoiados no princípio da parcimônia. Para tomar a decisão se os modelos são iguais ou diferentes utilizamos a estatística F da tabela de anova.

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 significativos¹²⁾
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 → 0
4. verifique se a ordem de remoção não interfere na seleção do modelo
- retorne ao modelo cheio
 - retire as variáveis que não foram retidas no outro procedimento em outra ordem
 - confirme que o modelo mínimo adequado é o mesmo

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

1. Construa os modelos plausíveis com esses dados
2. Faça a comparação do modelo cheio, simplificando pela partição da variação
3. Retenha o modelo mínimo adequado para interpretar

O que preciso entregar



- 1. As estimativas dos modelos devem ter sido incluídas nas planilhas quando foram solicitados ao longo do roteiro
- 2. Preencha as perguntas do quadro abaixo ou pelo [link do formulário](#)

1)

modelo linear com apenas uma preditora

2)

coluna adubo igual a sim

3)

coluna adubo igual a não

4) 7)

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

1.5. Para libreoffice use = NORM.INV(RAND(), 0, 1.5)

5)

o valor 1 indica que a resposta é predita apenas pela sua própria média

6)

se o termo da interação foi significativa, confira os cálculos e mantenha o resultado como está, esse resultado emerge com baixa frequência

8)

diferença entre observado e o predito pelo modelo

9)

O R^2 é a razão entre o resQuad sobre a soma quadrática total dos dados, definida como os desvio dos valores observado em relação à média dos valores elevado ao quadrado em seguida somados

10)

aquele que inclui a interação entre solo e adubo

11)

o modelo mais simples está contido no mais complexo

12)

um de cada vez

From:

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

Permanent link:

<http://labtrop.ib.usp.br/doku.php?id=cursos:planeco2020:roteiro:09-lm02>



Last update: **2021/03/01 15:59**