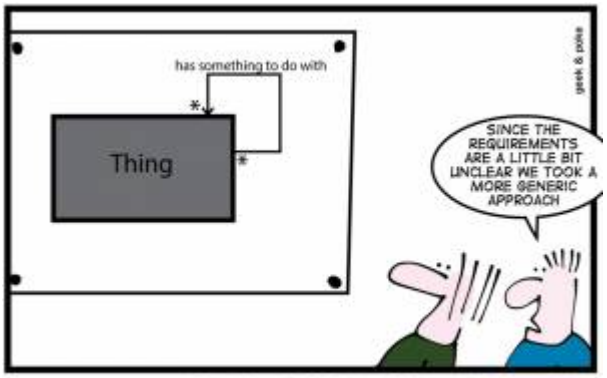
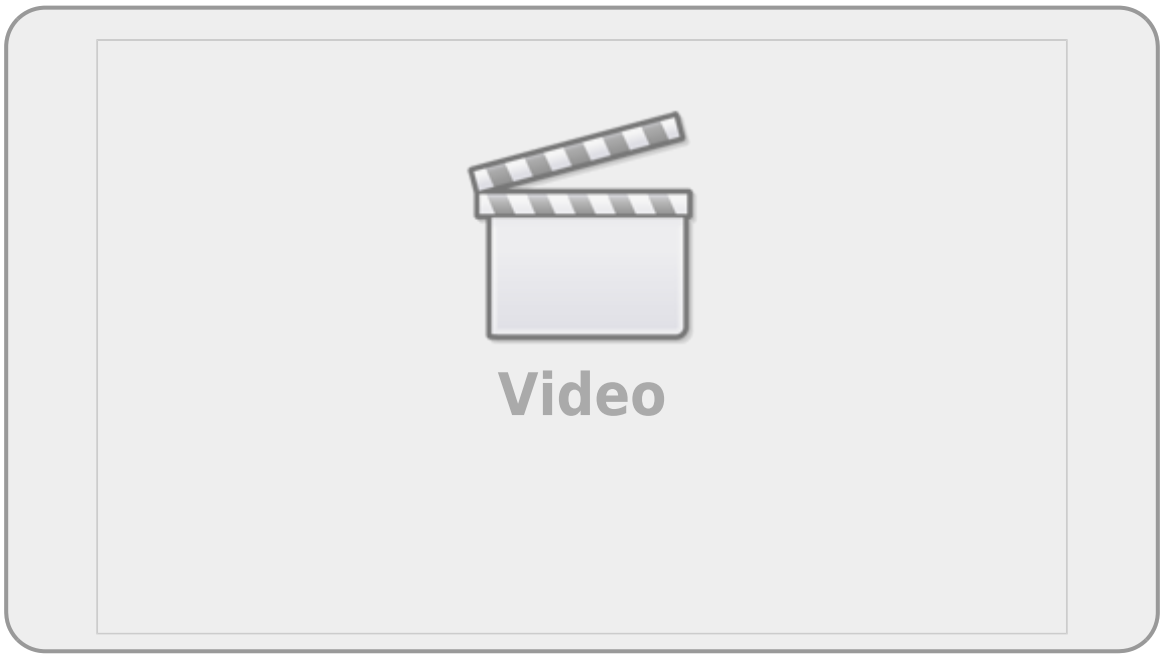
Modelos Lineares Múltiplos I



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.



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

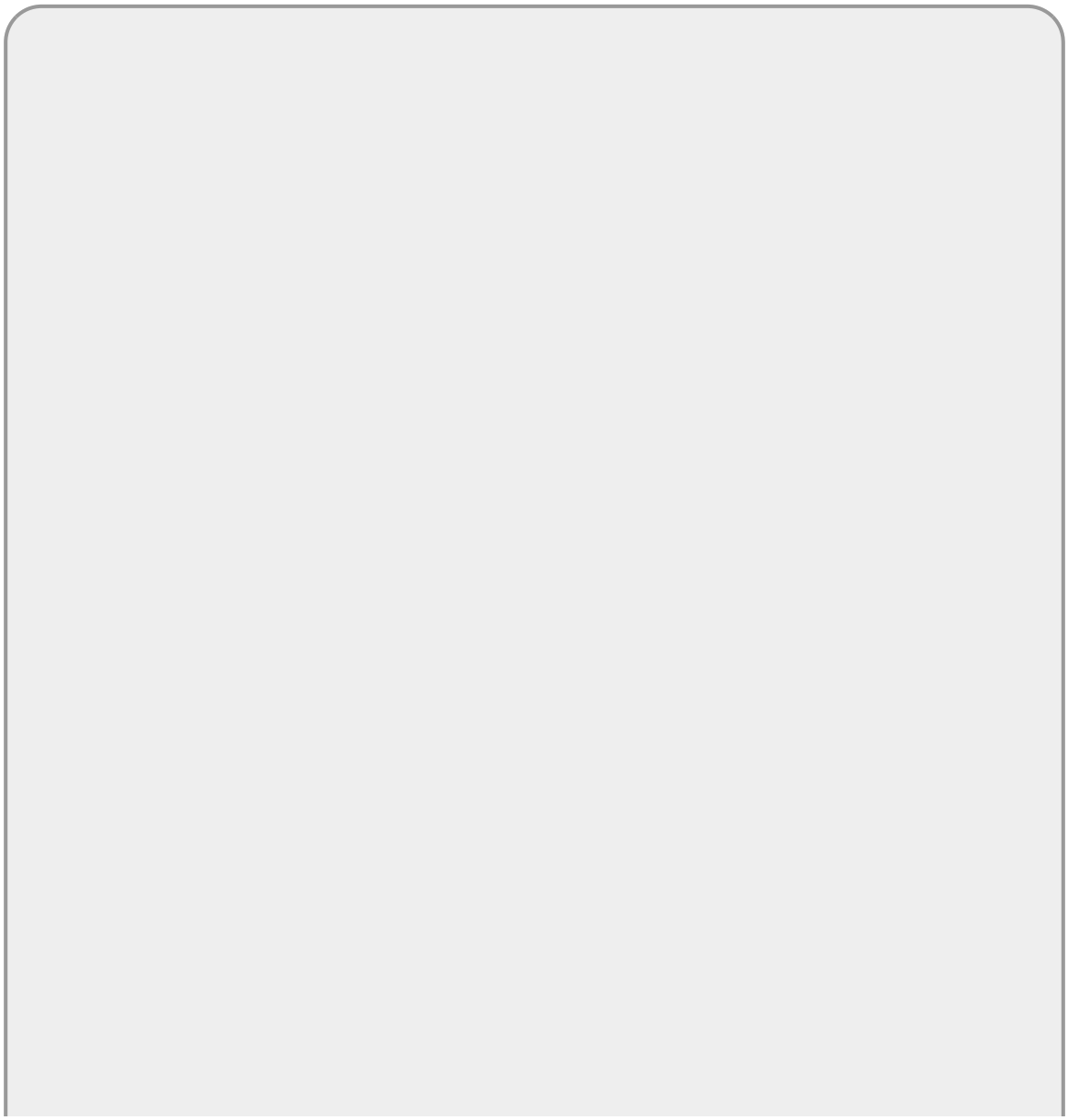
Vimos que existe um efeito do tipo de solo na produção de um cultivar. Uma expectativa plausível é que a adição de adubo também tenha efeito na produtividade. Ou seja, os tipos de solo tem produtividade diferente, assim como o adubo aumenta a produtividade.

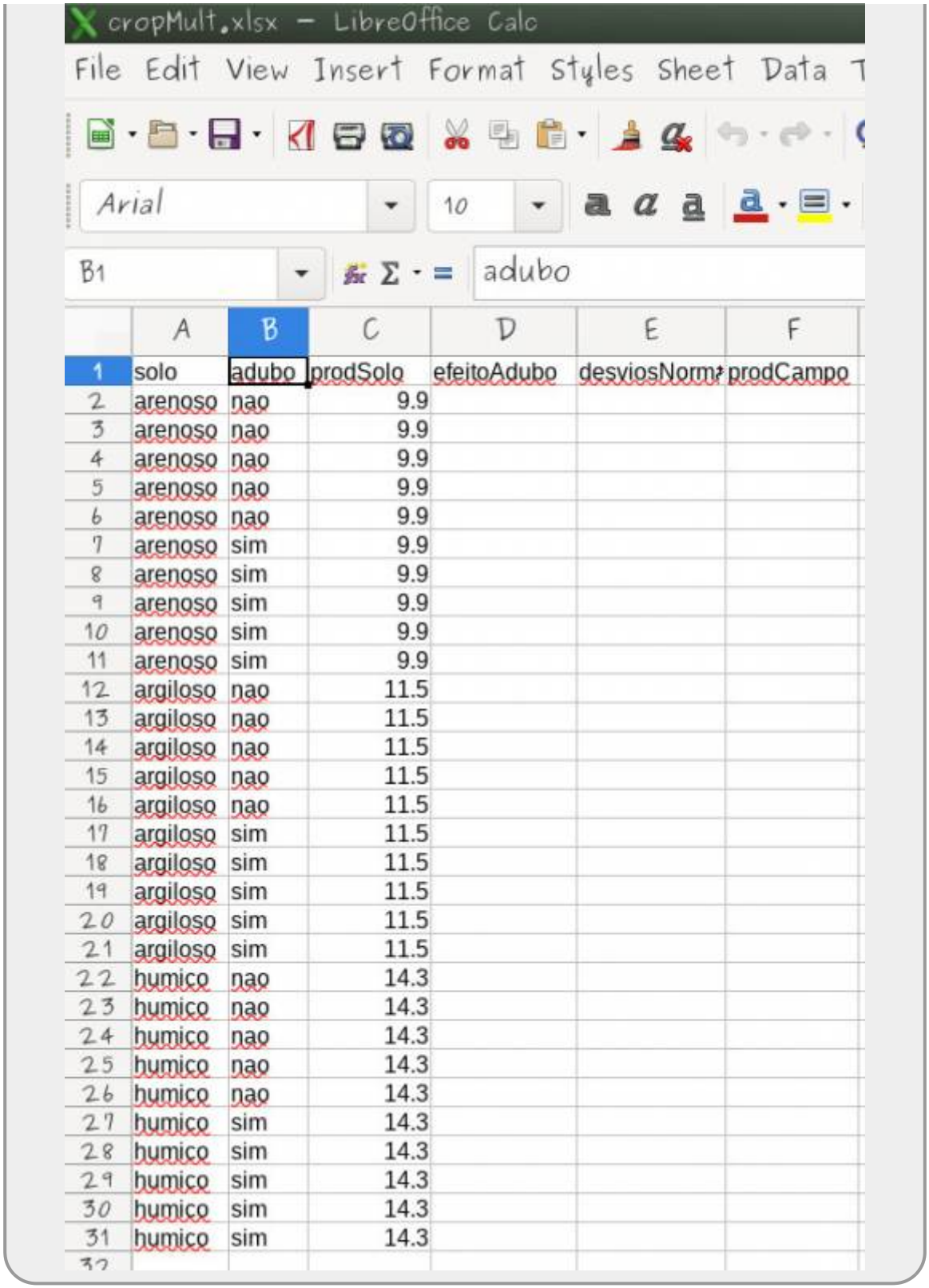
Nos dados originais do exercício de ANOVA a produtividade média nos solos foi de:

- 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:




cropMult.xlsx – LibreOffice Calc

File Edit View Insert Format Styles Sheet Data T

Arial 10

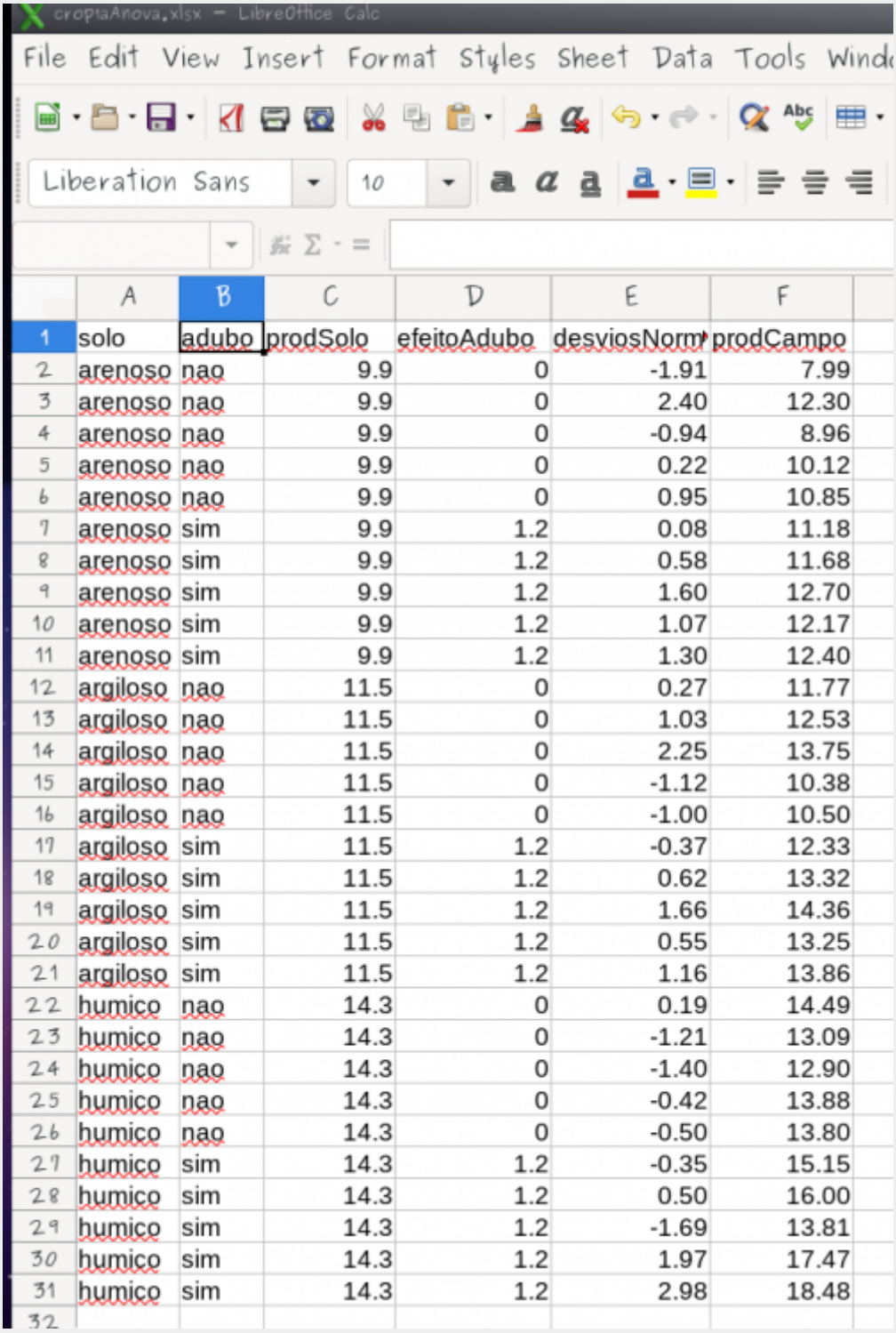
B1 adubo

	A	B	C	D	E	F
1	solo	adubo	prodSolo	efeitoAdubo	desviosNorma	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			

- 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:



	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

- Salve a planilha com o nome `soloAduboAditivo.csv` em formato texto

com campos separados por vírgula;
- Abra os dados no Rcmdr;
- Produza um modelo chamado `mlSolo_Adubo_Aditivo` da seguinte forma:

```
prodCampo ~ solo + adubo;
```

- Avalie o modelo pelo seu sumário e pela tabela de Anova;
- Faça uma interpretação biológica do resultado do modelo.

Modelos Plausíves

O nosso modelo tem duas preditoras e pode ser simplificado. Nesse caso, como temos poucas possibilidades de comparação, podemos comparar os modelos plausíveis, desde que sejam aninhados. 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:

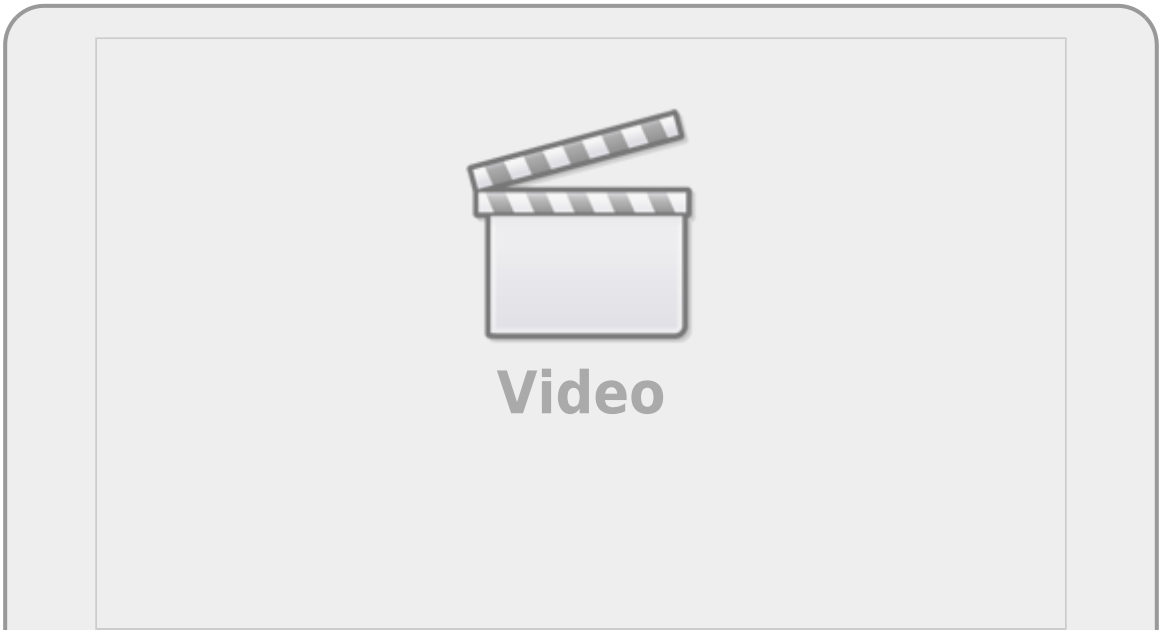
```
prodCampo ~ adubo
```

- **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)]

Interação entre preditoras



Nos modelos acima, desconsideramos um elemento importante que emerge quanto 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 de solo na produtividade agrícola depende da quantidade de adubo orgânico adicionado?* Ou em outras palavras: *O efeito da adubação orgânica depende do tipo de solo?* Note que nestas perguntas o foco não é se há ou não efeito do adubo ou solo, mas se a presença de uma variável afeta o efeito de outra.

- No conjunto de modelos acima, não incluímos o termo da interação. Produza o modelo abaixo incluindo o termo da interação e avalie esse modelo e seus coeficientes.

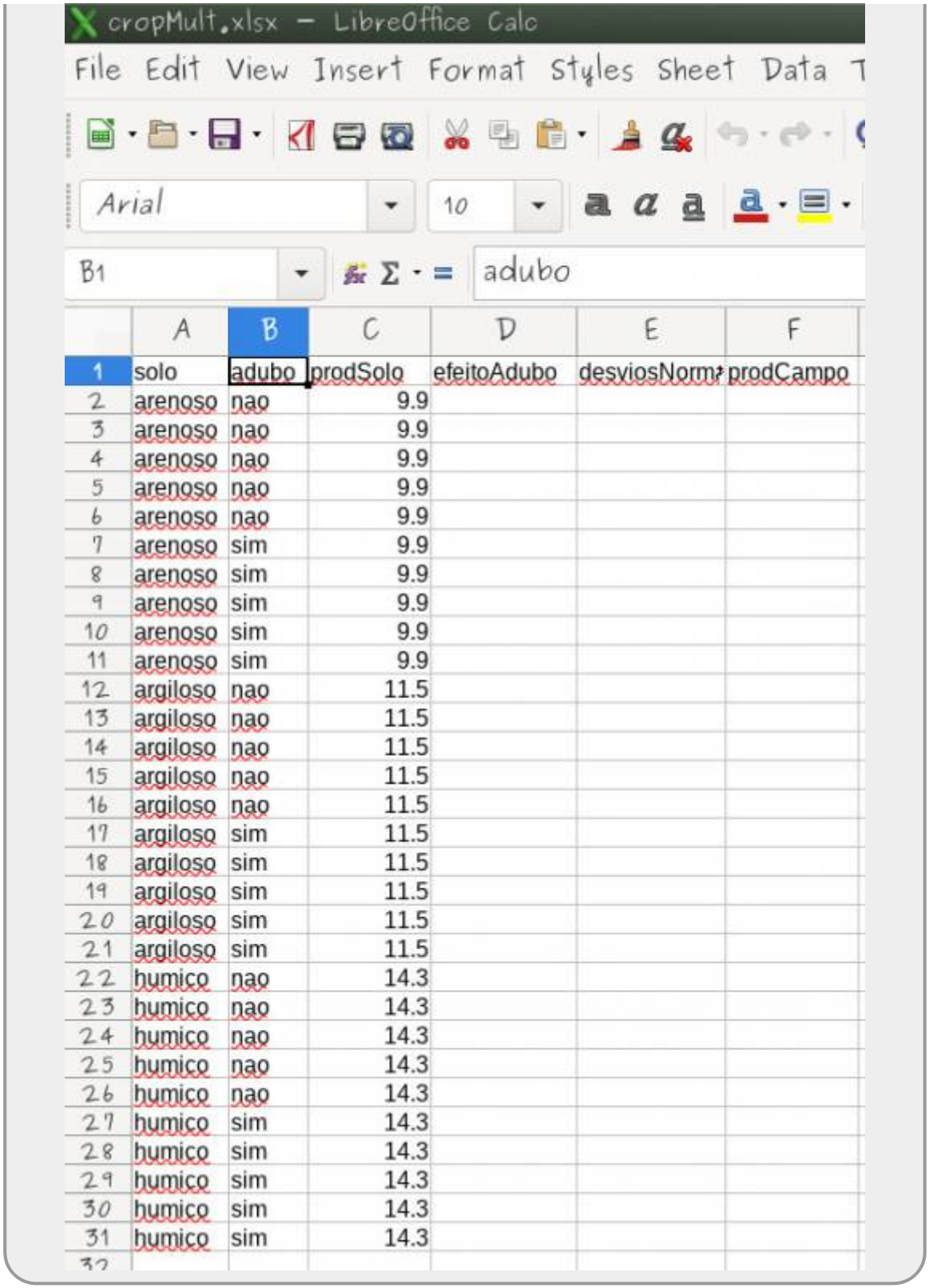
```
prodCampo ~ solo + adubo + solo:adubo
```

Não é esperado encontrar interação entre as preditoras nos dados simulados da maneira como fizemos, ele pode emergir 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:


cropMult.xlsx – LibreOffice Calc

File Edit View Insert Format Styles Sheet Data T

Arial 10

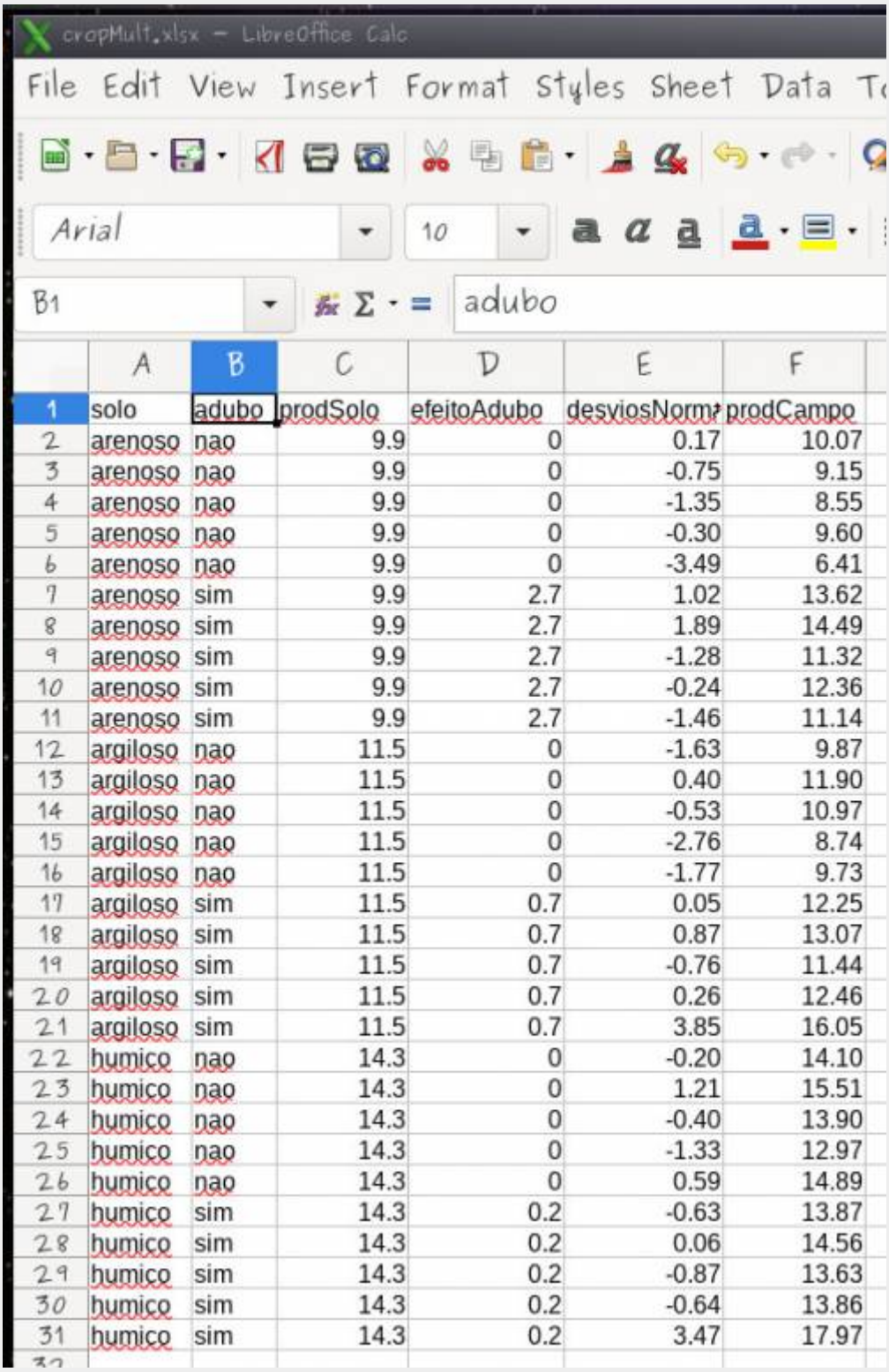
B1 adubo

	A	B	C	D	E	F
1	solo	adubo	prodSolo	efeitoAdubo	desviosNorma	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			

1. 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`
2. O campos da coluna `efeitoAdubo` onde adubo é igual a não devem ser preenchidos com `0`
3. Preencha a célula **E2** da coluna `desvios normal` com a fórmula **=** ***INV.NORM.N(ALEATÓRIO(); 0 ; 1.5)***[7], as atuais utilizam a mesma que o excel.

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:



	A	B	C	D	E	F
1	solo	adubo	prodSolo	efeitoAdubo	desviosNorma	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 o nome `soloAduboInteracao.csv`;
2. Importe os dados para o Rcmdr. **Atenção nomeie os dados na aba de importação com o nome `soloAduboInt`, em alguns casos o Rcmdr não importa se a planilha e os dados importados tiverem o mesmo nome de uma importação anterior**
3. Confira se os dados foram lidos corretamente, inclusive se a decimal é `.`;
4. Produza o modelo cheio `mlSolo_AduboAll` com a seguinte formula:
   - `prodCampo ~ solo + adubo + solo:adubo`
   - interprete o resumo, comparando com o resumo do modelo similar proveniente da planilha de dados anterior

Simplificando Modelos

Durante o curso usaremos o procedimento de simplificar o modelo a partir do modelo cheio. O procedimento consiste em comparar modelos aninhados[8], 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. Quando os modelo comparados são diferentes retemos o mais complexo, pois explica mais variação dos dados [9]. Por outro lado, 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 significativos [10)]
3. ordem de remoção de termos:
    - interações não significativas (primeiro as de maior ordem)
    - termos quadráticos ou não lineares
    - variáveis explicativas não significativas
4. caso faça sentido, agrupe níveis de fatores sem diferença
5. 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
6. Faça o diagnóstico do modelo mínimo adequado
7. Interprete o modelo selecionado

Tomada de decisão

A diferença não é significativa:

- retenha o modelo mais simples
- continue simplificando

A difereça é significativa:

- retenha o modelo complexo
- verifique se existe termo que pode e ainda não foi retirado
- caso não haja nenhum termo que possa ser retirado, este é o modelo MINÍMO ADEQUADO

Interpretando Variáveis Indicadoras (Dummy)

As variáveis indicadoras devem ser interpretadas com cuidado. No exemplo do modelo cheio acima [11], 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:

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

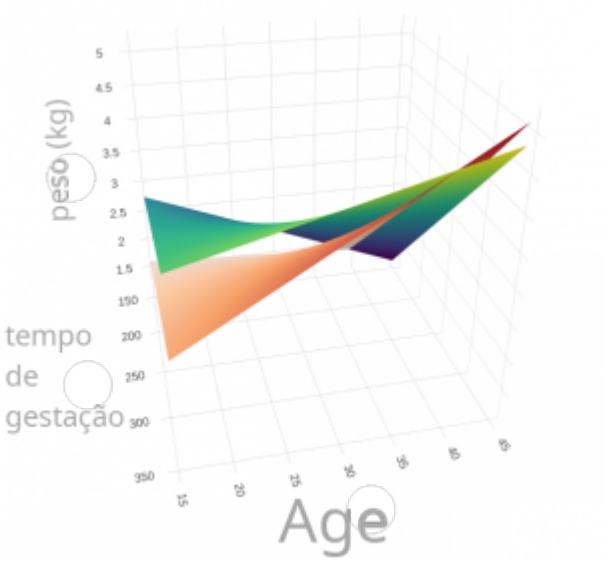
E assim por diante, usando as variáveis indicadoras e os coeficientes estimados para o cálculo do predito pelo modelo.

Procedimento

1. Faça a seleção do modelo mínimo adequado para o conjunto de dados da última planilha, partindo do modelo com a interação, simplificando até o modelo mínimo adequado. Utilize o procedimentos de comparação de modelo pela partição de variância;
2. 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.
3. Preencha a aba `cropIntera` da planilha lmCrop2pred com os resultados do modelo selecionado
4. Na planilha onde os dados foram gerados, calcule, a partir dos coeficientes estimados, os valores preditos pelo modelo para cada uma das observações, coloque esses valores em uma coluna nomeada de `predito`. Veja como calcular os valores preditos no quadro interpretando_variáveis_indicadoras_(dummy)
5. Calcule os resíduos do modelo [12] em uma coluna denominada `residuos`
6. Eleve o valor dos resíduos ao quadrado em uma coluna denominada `resQuad`. A soma destes valores representa a variabilidade não explicada

pelo modelo
7. Calcule a média da variável resposta e calcule a diferença deste valor para todas as observações e eleve ao quadrado e armazene em uma coluna `desvQuadTotal`, a soma destes valores representa a variabilidade total dos dados
8. Calcule o R^2 do modelo, baseado no `resQuad` e no `desvQuadTotal` [13)]

Modelos Lineares Múltiplos: preditoras contínuas e categóricas



Nesse último tópico do bloco vamos resgatar os principais conceitos que emergiram com a generalização do modelo linear, agora com múltiplas preditoras, a partir de um exemplo com mais variáveis preditoras contínuas e categóricas.

Desafios dos modelos com múltiplas preditoras

Além da interação entre as preditoras, tratada no exemplo anterior, a inclusão de múltiplas preditoras eleva a complexidade do modelo e dificulta sua interpretação. Por exemplo, um modelo com três variáveis preditoras pode potencialmente ter quatro termos de interação[14)], além dos termos dos efeitos isolados. Com quatro preditoras já são 11 potenciais termos de interações possíveis. Além do número elevado de termos possíveis, as interações de níveis mais elevados são difíceis de interpretar. Essa é uma das razões para buscarmos a simplificação dos modelos antes de interpretá-los. Um outro ponto importante é nunca incluir termos no modelo que não temos capacidade para interpretar, a menos que seu modelo seja apenas preditivo. Para modelos que se propõem apenas a fazer predições, um termo complexo não impede sua aplicação. Entretanto, se a proposta do modelo é entender os efeitos causais e os processos subjacentes, um termo complexo, que não pode ser interpretado, inviabiliza o modelo. Nesses casos, é importante apenas incluir termos no modelo que são passíveis de serem interpretados.

Uma outra questão que emerge dos modelos com múltiplas preditoras contínuas é a

multicolinearidade. Quando duas variável preditoras apresentam muita sobreposição de variação explicada, ou seja, ambas explicam a mesma porção considerável da variação na resposta, o modelo pode ter problemas na estimativa dos coeficientes. Além disso, modelos diferentes podem ser selecionados na busca do mínimo adequado, dependendo da ordem em que os termos são retirados ou incluídos. Variáveis com muito correlação entre elas tem grande potencial de apresentar multicolinearidade no modelo. Uma boa abordagem para quem está iniciando na modelagem estatística é investigar a correlação entre as preditoras e, caso a correlação seja muito forte, ficar atento com as estimativas do coeficientes e com a ordem de retirada dos termos do modelo ao buscar o modelo mínimo adequado. Se possível, descartar uma das preditoras colineares é uma das estratégias, já que ambas estão explicando a mesma porção da variação e são em grande parte redundantes.

Caso tenha interesse sobre outra técnica de diagnóstico de colinearidade, consulte o roteiro Modelos Lineares Múltiplos III.

Ao final desta seção é desejável que tenha compreendido nos modelos lineares múltiplos:

- a partição da variância do modelo;
- interpretar a `tabela de anova` na comparação de dois modelos;
- entender o procedimento da `anova` para simplificação do modelo;
- interpretar os coeficientes estimados;
- entender quais níveis estão representados no intercepto do modelo;
- compreender os termos de interação;
- compor o predito pelo modelo a partir dos coeficientes;
- interpretar biologicamente o resultado do modelo;
- fazer o diagnóstico do modelo;

Peso de bebês ao nascer

O objetivo dessa pesquisa foi saber quais fatores afetam o tamanho de bebês ao nascer, de modo que fosse possível orientar campanhas de conscientização para evitar o nascimento de bebês com baixo peso, uma vez que isso pode implicar em muitos riscos ao bebê maiores custos devido à permanência no hospital. As variáveis preditoras consideradas para essa pesquisa estão listadas abaixo, mas também havia um interesse genuíno em saber se o efeito de uma variável poderia interferir no efeito das outras.

A descrição destes dados pode ser consultada em https://www.stat.berkeley.edu/users/statlabs/labs.html#babies.

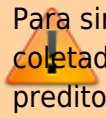
Descrição dos dados

- variável resposta
  - `bwt` : peso do bebê ao nascer em onças(oz)
- preditoras:
  - `gestation`: tempo de gestação (dias)
  - `age`: idade da mãe (anos)
  - `height`: altura da mãe (polegadas)
  - `weight`: peso da mãe (libras)
  - `smoke`: 0 não fumante; 1 fumante

Notem que as preditoras estão relacionadas a características da mãe: dias de gestação, idade, peso, altura e se ela é fumante ou não. Como a variável resposta, peso do bebê ao nascer, foi medida em `onças`, vamos transformar em uma escala de medida que temos mais facilidade para interpretar, multiplicando essa variável por `0.02835` para transformar em kg.

- Abra o arquivo
  babies.csv
  no Rcmdr[15];
- Garanta que os dados foram lidos corretamente;
- Abra a janela para criar uma nova variável no menu `Data > Manage variables in active data set > Compute a new variable`;
- Na caixa `New variable name` nomei a nova variável como `pesoKg`;
- Na caixa `Expression to compute` coloque a expressão: `bwt * 0.02835`;

Agora que temos uma variável que caracteriza o tamanho do bebê ao nascer em uma escala que faz mais sentido (kg), vamos começar a nossa avaliação de quais características da mãe tornam os bebês mais susceptíveis a nascerem com pesos baixo.



Para simplificar nosso exemplo, vamos deixar de lado duas variáveis preditoras que foram coletadas nesse estudo: `parity` e `weight`. Ao final ficamos com quatro variáveis preditoras: `gestation`, `age`, `height` e `smoke`

Um mal começo

Um procedimento de modelagem ineficiente é colocar todas as variáveis preditoras e suas interações e torcer para ter alguma resultado interessante. No caso do peso dos bebês, que é razoavelmente simples com quatro variáveis preditoras sendo uma delas categórica, o modelo resultante é bastantes complexo. Vamos construir esse modelo e verificar o resultado:

No menu `Estatísticas > Ajustes de modelos > Modelo linear…`, construa o modelo `lmFull` com todas as quatro preditoras [16] e suas interações. O modelo resultante tem a seguinte expressão:

```
pesoKg ~ age * gestation * height * smoke
```

Na linguagem R os símbolos de * em expressões de notação de modelos representa tanto a variável isoladamente quanto as interações possíveis. Verifique o resumo do modelo gerado.

Um modelo como o construído acima tem muitos problemas. Nesse caso, foram estimados 16 coeficientes um para cada termo do modelo. Note também que nenhum dos coeficientes estimados é significativo nos testes marginais, ou seja, as incertezas nas estimativas dos coeficientes não nos permite afirmar que nenhum deles seja diferente de zero. Portanto, temos um modelo estatístico complexo, com 16 termos, sendo que todos os termos são multiplicados por coeficientes não distintos de zero!

Complexidade das interações

Vamos calcular o número de interações possíveis em modelo com 4 preditoras:

Combinatória

Combinatória é a operação matemática para calcular de quantas maneiras conseguimos organizar ou combinar um conjuntos de elementos. No nosso caso, temos 5 variáveis e podemos nos perguntar de quantas formas podemos combinar esses elementos em diferentes conjuntos de dois a dois (interação dupla) ou três (interação tripla) e quatro (interação quadrupla)[17].
A expressão matemática para essa operação é:

$$C^{r}_{n} = \frac{n!}{r! (n-r)!} $$

onde: n: número total de elementos r: número de elementos combinados

A expressão ! é a operação fatorial na matemática.

Para calcular o número de combinações possíveis de grupos de 2 e 3 elementos para um total de 4, temos:

$$C^{2}_{4} = \frac{4!}{2! (4-2)!} = 6$$ $$C^{3}_{4} = \frac{4!}{3! (4-3)!} = 4$$ $$C^{4}_{4} = \frac{4!}{4! (4-4)!} = 1$$

A lista de interações triplas e quadruplas tem essas expressões:

- `gestation:age:weight:smokeTRUE`
- `age:gestation:height`
- `age:gestation:smoke`
- `age:height:smoke`
- `gestation:height:smoke`

As interações duplas:

- `age:gestation`
- `age:height`
- `age:smoke`
- `gestation:height`
- `gestation:smoke`
- `height:smoke`

Além dessas, temos os termos isolados para cada variável e o intercepto do modelo, totalizando os 16 termos do modelo `lmFull`.

Se tivéssemos mais uma preditora contínua em nosso modelo o número de interações possíveis subiria para 26 e o total de termos do modelo para 31.

Note que mesmo com poucas variáveis as possibilidade de interações e grande para decidirmos o que será incluído no modelo. Esse processo não é trivial e deve estar embasado no conhecimento prévio do sistema e na teoria para definir aquilo que faz sentido e pode ser interpretado, caso seja mantido no modelo.

Vamos agora comparar esse modelo cheio com o modelo sem nenhuma preditora.

Comparação com o modelo nulo

- Produza o modelo nulo chamado `lmNull` com a formula:
  - `pesoKg ~ 1`
- compare com o modelo cheio produzido acima `lmFull` utilizando a função:
  - `anova(lmNull, lmFull)`

Nosso modelo `lmFull`, apesar de não ter nenhum termo com coeficientes significativamente diferentes de zero, explica uma porção razoável e “significativa” da variação dos dados. Por volta de

25% da variação dos dados é explicada pelo modelo, por outro lado, não temos nenhum termo do modelo que valha a pena interpretar. Um poder de predição razoável mas com nenhum poder de interpretação dos processos subjacentes.

Faça uma tentativa de interpretar o que o termo mais complexo deste modelo, a interação que aparece na última linha dos coeficientes estimados, está informando:

`age:gestation:height:smoke[T.TRUE]`

O que significa uma interação entre as características da mãe: idade, dias de gestação, altura e que é fumante? Normalmente, conseguimos interpretar interações de segundo nível (levando em conta duas variáveis), algumas poucas vezes há sentido em interpretar a interação de três preditoras. Então o primeiro passo é retirar qualquer termo de interação de ordem mais alta.

Definindo as interações

Para modelos com várias variáveis preditoras é sempre complicado definir quais interações devem ser contempladas. E como vimos acima, a estimativa e a interpretação dos coeficientes do modelo ficam comprometidos. Se a intenção do modelo é entender os processos causais da variação na variável resposta, é importante iniciarmos a seleção do modelo mínimo adequado com um modelo cheio que contemple apenas os termos que são importantes para o processo em questão.

Modelo Cheio

Depois de definir quais são os termos que queremos incluir no nosso modelo (variáveis simples e interações), podemos iniciar o procedimento de modelagem seguindo algum protocolo para chegar ao modelo mínimo adequado. No nosso caso, iremos partir do modelo cheio, simplificar até o mínimo adequado. Como não temos experiência prévia do sistema e não temos “muita” experiência sobre tamanhos de bebês ao nascer, “consultamos vários especialistas na área” e eles chegaram à conclusão que os termos que deveriam ser contemplados são:

`pesoKg ~ gestation + age + weight + smoke + gestation:age + gestation:smoke + age:weight + age:smoke + weight:smoke + gestation:age:smoke`

Seleção do mínimo adequado

- No menu `Estatísticas > Ajustes de modelos > Modelo linear…`, construa o modelo com a seguinte expressão:

```
bwt ~ gestation + age + weight + smoke +
          gestation:age + gestation:smoke +
          age:weight + age:smoke + weight:smoke +
          gestation:age:smoke
```

- simplifique esse modelo até o mínimo adequado;

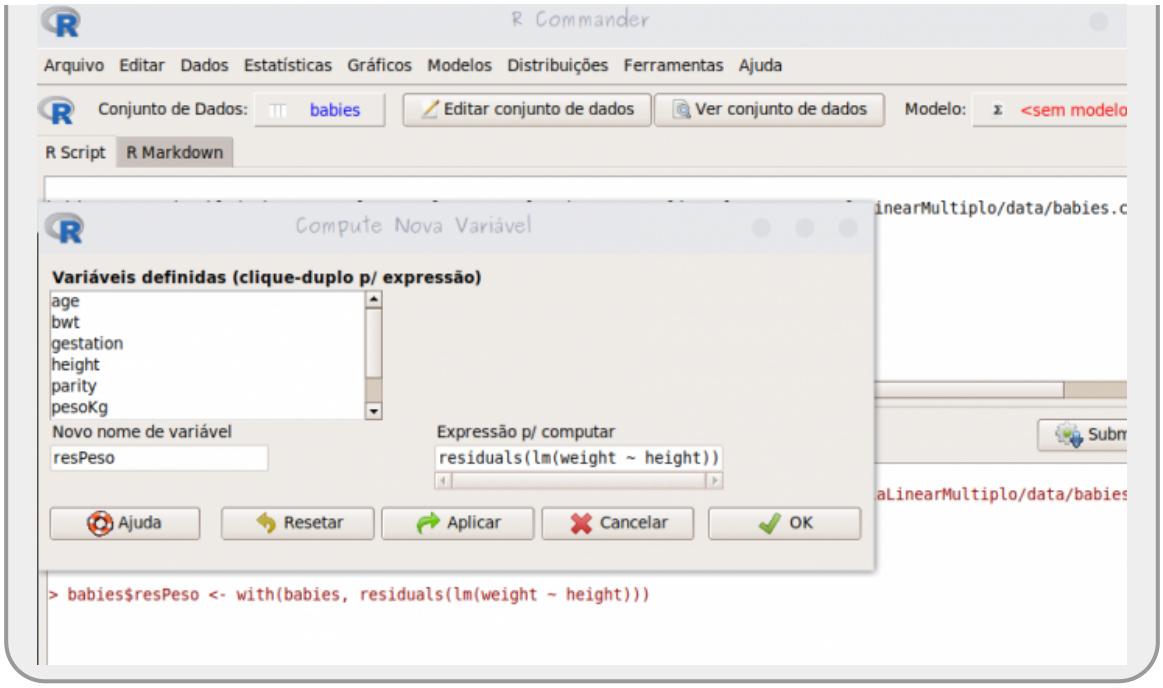
Durante o processo de simplificação, quando nos defrontamos com vários termos de mesma ordem não significativos, um bom procedimento é retirar um deles de cada vez e, mesmo que o termo não seja retido no modelo, retorna-lo ao modelo antes de retirar o outro. Caso a ordem de retirada não torne nenhum dos termos significativos, ambos podem ser retirados. No caso de serem mais do que dois termos de mesma ordem, é importante também testar a retirada de dois a dois termos depois do procedimento de retirada de um a um não ter tornado nenhum termo significativo. Isso garante que a ordem de retirada não define o termo que será retido no modelo. Lembre-se que um termo "não significativo" em um modelo mais complexo pode se tornar "significativo" em um modelo mais simples.

- interprete o resultado do modelo mínimo adequado com relação aos termos selecionados;
- qual a predição do modelo selecionado do peso do bebê de uma mãe de 30 anos, que teve uma gestação de 280 dias e peso de 125 pounds
- interprete o resultado biologicamente.

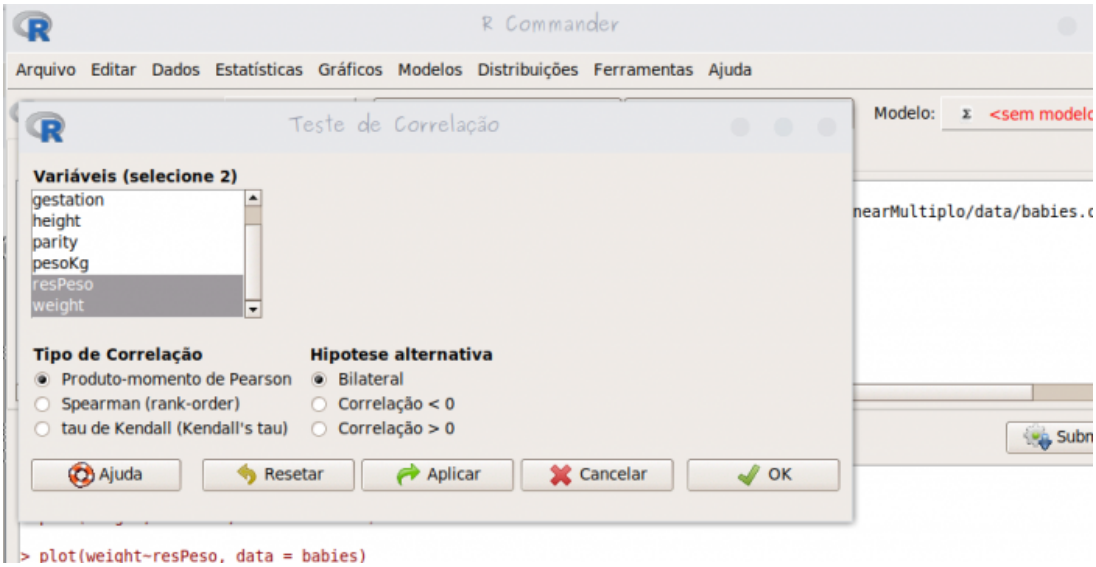
Tamanho e sobrepeso da mãe

Um "outro especialista" ao analisar o modelo mínimo adequado selecionado acima afirmou que deveria incluir a variável `height` para controlar o peso do bebe ser maior apenas porque a mãe era grande. Além disso sugeriu incluir uma nova variável que indicasse a quanto a gestante tem de sobrepeso ou subpeso. Indicou ainda que uma forma de criar essa variável seria utilizar os resíduos de uma regressão simples da variável peso como respotas e a variável altura como preditora. Todos os "outros especialistas" concordaram que essa era uma boa sugestão, já que essa nova variável representa o quanto a gestante tem mais peso ou menos peso do que esperado para uma gestante com a sua altura.

- inclua a variável `height` no modelo mínimo selecionado anteriormente e verifique se deve se essa variável deve ser retida no modelo, com a comparação dos modelos pela partição da variação;
- crie uma nova variável `resPeso` com a expressão `residuals(lm(weight ~height));`



- construa um modelo que inclua a variável `resPeso` como o primeiro termo no modelo mínimo adequado selecionado no procedimento anterior;
- cronstrua um segundo modelo que inclua `resPeso` agora como o último termo no modelo mínimo adequado selecionado nos passos anteriores;
- compare os resumos de ambos os modelos e anote as diferenças encontradas;
- no menu `Estatísticas` > `Resumos` > `Teste de correlação` selecione as variáveis `resPeso` e `weight`, anote o valor da correlação entre essas variáveis;



- a partir da avaliação da correlação acima tome a decisão de reter ambas, uma ou nenhuma das duas variáveis do passo acima, justifique sua decisão e construa o modelo resultante;
- simplifique o modelo do item anterior para o mínimo adequado, caso necessário;
- faça uma interpretação biológica do modelo final.

Formulário de resposta

Responda o o formulário MLM III incluindo arquivos de resultados e figuras quando solicitado.

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)
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 significativo, confira os cálculos e mantenha o resultado como está, esse resultado emerge com baixa frequência, simplesmente por acaso.
7)
Essa expressão retorna valores associados a uma distribuição normal com média 0 e desvio padrão 1.5. Para versões antigas do libreoffice a função pode ser `= NORM.INV(RAND(), 0, 1.5)`
8)
o modelo mais simples está contido no mais complexo
9)
Este é um atributo associado aos modelos aninhados: aquele que tem mais variáveis ou parâmetros só pode explicar mais ou a mesma quantidade de variação do mais simples, já que todos os parâmetros do modelo mais simples estão contidos no mais complexo
10)
um de cada vez
11)
aquele que inclui a interação entre `solo` e `adubo`
12)
diferença entre observado e o predito pelo modelo
13)
O R^2 é a razão entre (`desQuadTotal` - `resQuad`) sobre a `desvQuadTotal`. Ou seja, quanto da variação dos dados é explicada pelo modelo em relação ao total de variação dos dados
14)
três interações duplas e uma tripla
15)

os campos neste arquivo são separados por tabulação
16)

`gestation`, `age`, `height` e `smoke`
17)

Essa operação é chamada de combinatória simples, pois a ordem dos elementos nas combinações não importa