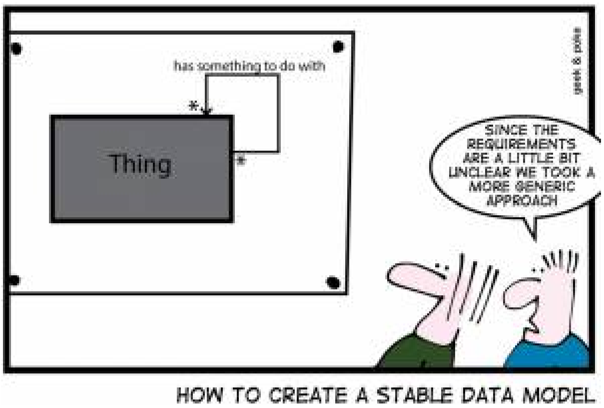
Modelos Lineares Múltiplos I



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

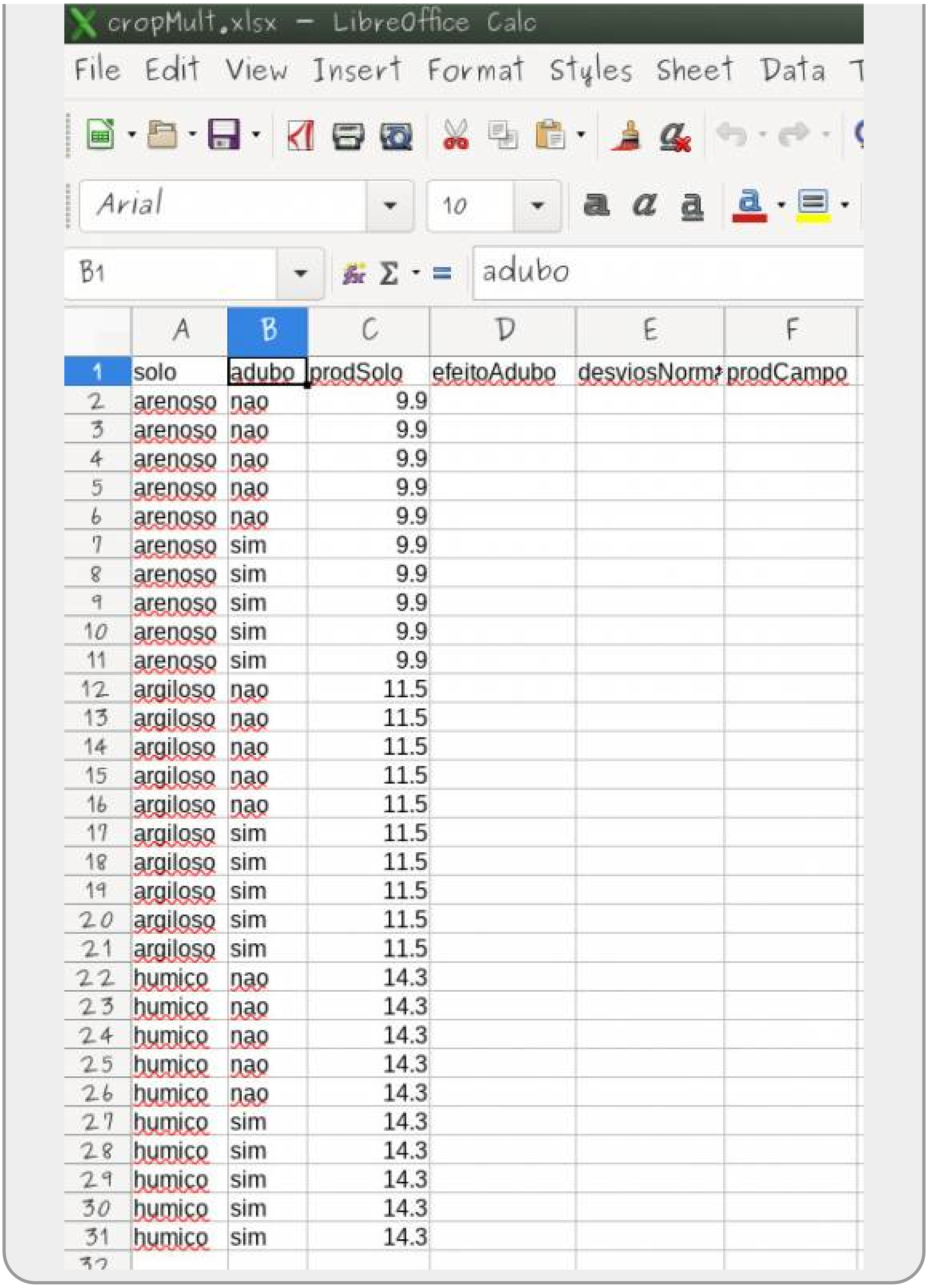
Vimos que existe um efeito do tipo de solo na produção de um cultivar. Uma expectativa plausível é que a adição de adubo também tenha efeito na produtividade. Ou seja, os tipos de solo tem produtividade diferente, assim como o adubo aumenta a produtividade.

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.

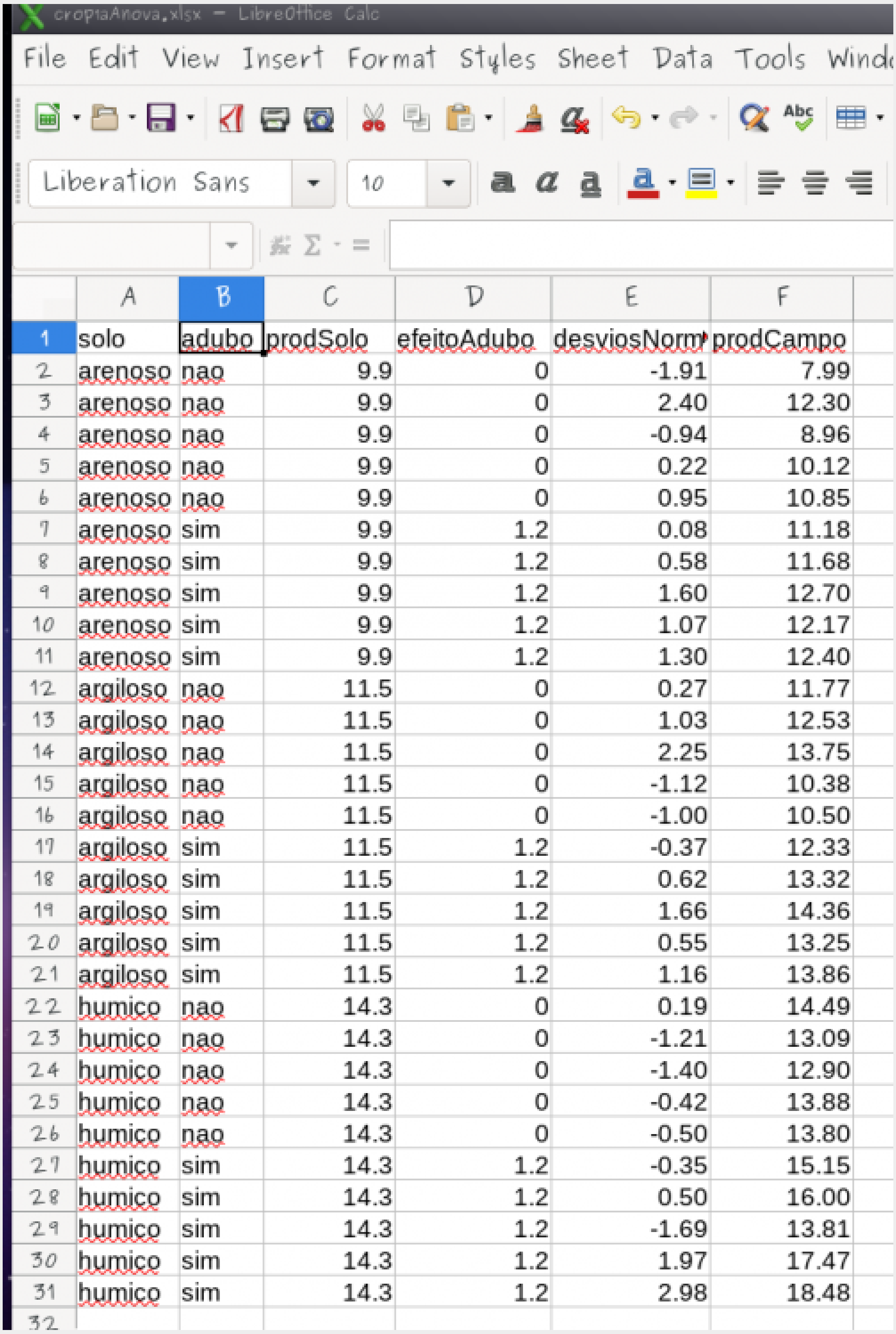
- 1. Abra o arquivo cropMulti}}preservefilenames::cropMult.xlsx em uma planilha eletrônica:



	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			

- 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.

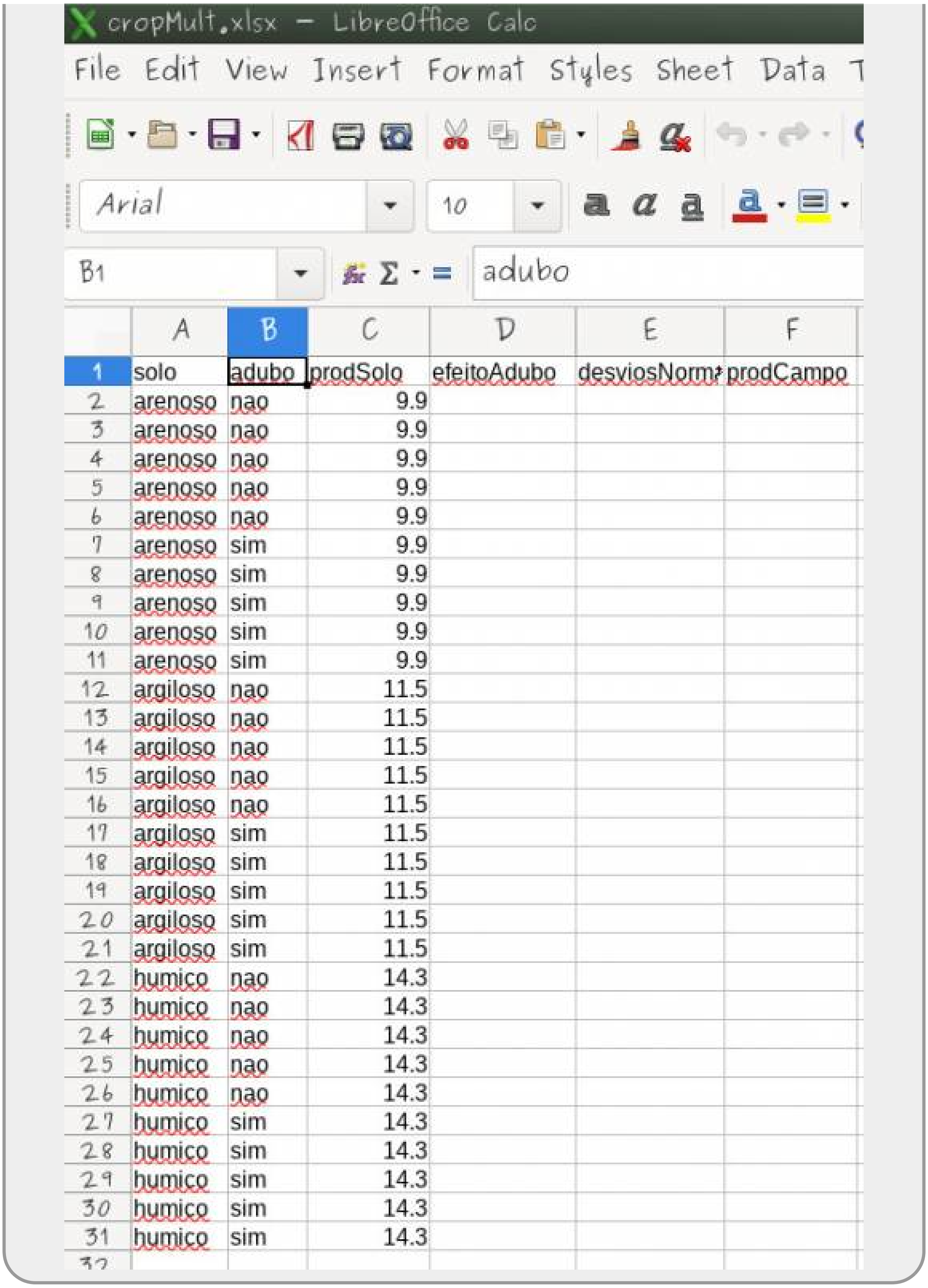
```
prodCampo ~ solo + adubo + solo:adubo
```

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:

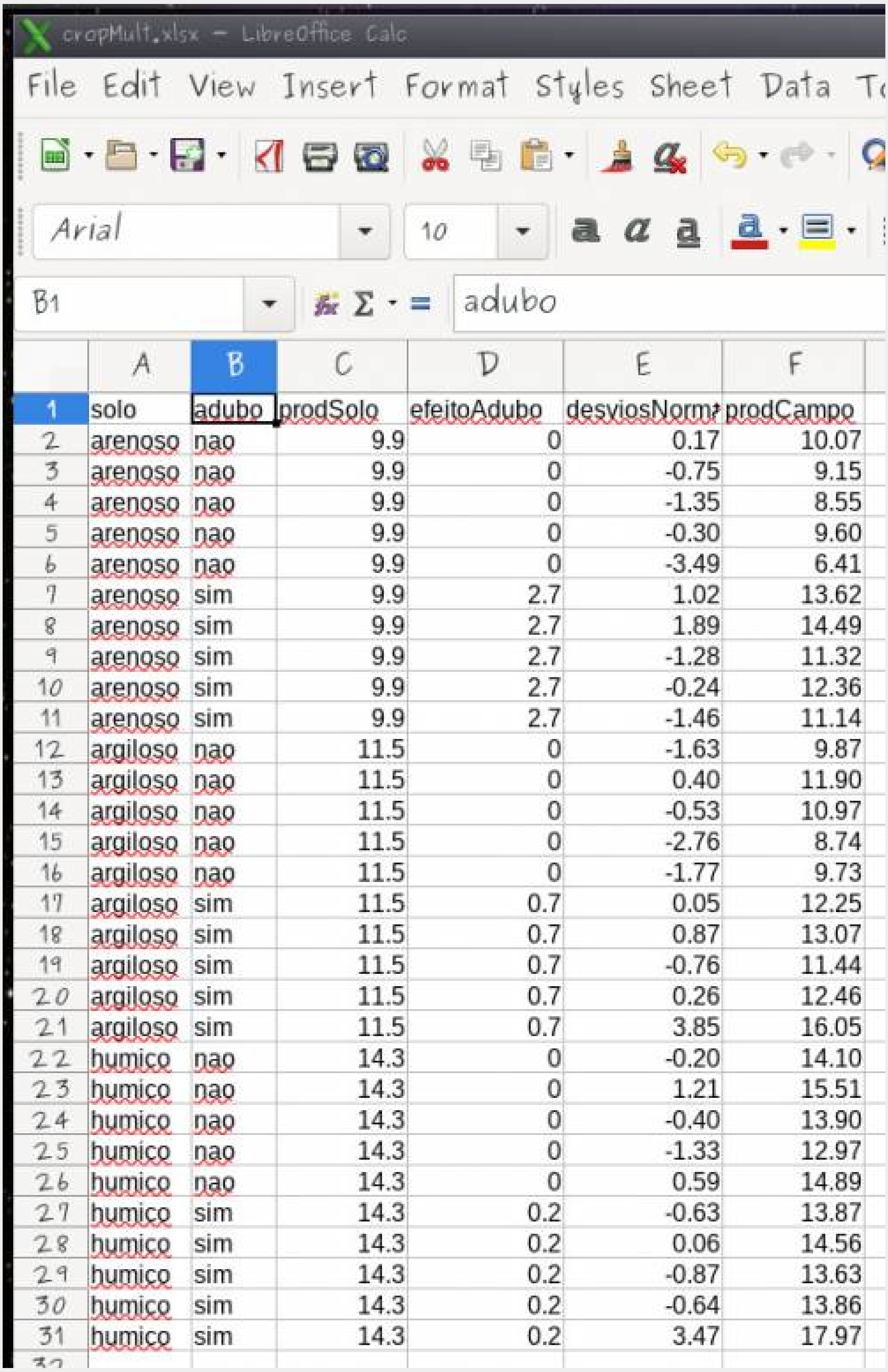


	A	B	C	D	E	F
1	solo	adubo	prodSolo	efeitoAdubo	desviosNorm	prodCampo
2	arenoso	nao	9.9			
3	arenoso	nao	9.9			
4	arenoso	nao	9.9			
5	arenoso	nao	9.9			
6	arenoso	nao	9.9			
7	arenoso	sim	9.9			
8	arenoso	sim	9.9			
9	arenoso	sim	9.9			
10	arenoso	sim	9.9			
11	arenoso	sim	9.9			
12	argiloso	nao	11.5			
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			

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:



	solo	adubo	prodSolo	efeitoAdubo	desviosNorm	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 <u>arg</u>, <u>hum</u> e <u>adubo</u> 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.

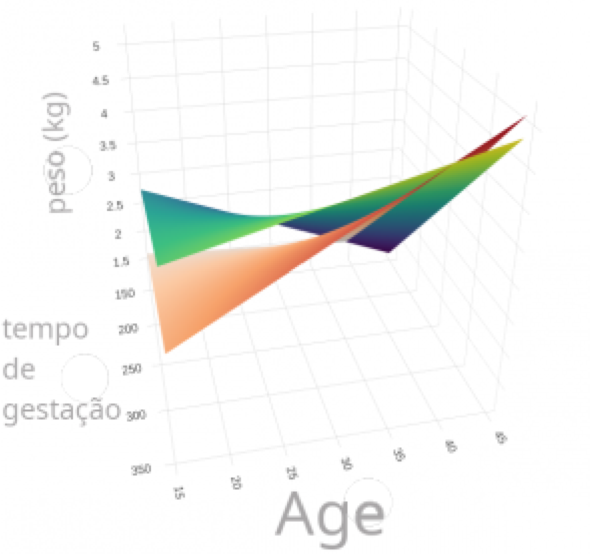
<u>Procedimento</u>

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 que tem duas variáveis preditoras contínuas e duas categórica. Acreditamos que esse exemplo incorpora as complexidades tratadas e ajuda a agrupar os tópicos que devem ficar atentos nos modelos com múltiplas preditoras.

Desafios dos modelos com múltiplas preditoras

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

- compreender 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;
- saber interpretar os gráficos diagnósticos do modelo;
- avaliar a colinearidade entre variáveis no 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.

VIF e as interações

No *Rcmdr* o `VIF` é aplicado ao modelo ativo pelo menu `Models > Numerical diagnostics > Variance-inflation factors`), calculando o valor para todos os termos do modelo, inclusive as interações. Como interações e as variáveis isoladas compartilham parte da variação explicada, a correlação entre eles é esperada. Ou seja, não é possível fazer a avaliação do VIF das variáveis em modelos com interação diretamente. Uma soluçao é fazer modelos sem as interações como fizemos anteriormente. Uma outra forma de contornar esse problema é fazer uma transformação simples nas variáveis contínuas, centralizando a média em `zero`, subtraindo o valor observado da média($x_i - \bar{x}$).

Com essa transformação o valor `0` passa a representar a média e os valores positivos o aumento em relação a média e negativos a diminuição, na mesma unidade de escala da variável original. A **centralização** das variáveis contínuas é uma transformação corriqueira pois não dificulta a interpretação e ao contrário, evita muitos problemas analíticos e de interpretação. Entre as vantagens da centralização está a possibilidade de interpretar o `VIF` diretamente no modelo selecionado e incorporar uma interpretação biológica para o valor do intercepto, onde muitas vezes não existia.

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 maiores custos e muitos riscos ao bebê devido à permanência no hospital. Três variáveis preditoras (explicadas abaixo) foram consideradas relevantes para essa pesquisa, mas também havia um interesse genuíno em saber se alguma das variáveis poderia interferir no efeito das outras. Como a variável resposta, peso do bebê ao nascer, foi medida em `onças` vamos primeiro transformar em uma escala de medida que temos mais facilidade para interpretar, multiplando essa variável por `0.02835` para transformar em kg.

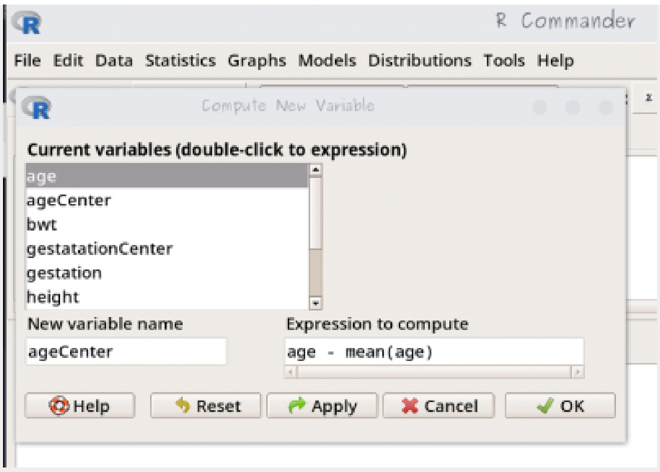
- Abra o arquivo

  babies.csv

  no Rcmdr, usando tabulação(Tabs) como separador de campo
- 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`;
- Ajuste um modelo contendo apenas as variáveis indicadas abaixo e todas as interações entre elas:
- variável resposta: `pesoKg` = peso do bebê (medido em kg)
- preditoras:
  - `gestation` = tempo de gestação (dias)
  - `age` = idade da mãe
  - `smoke`: FALSE mãe não fumante; TRUE mãe fumante

- Selecione o modelo mínimo plausível pelo método de simplificação para mínimo adequado (ver roteiro I de MLM)

- Calcule o VIF do modelo selecionado pelo menu `Models > Numerical diagnostics > Variantion Inflation Factor`
- Guarde o resultado dos VIF destes modelos;
- Crie uma nova variável pelo menu : `Data > Manage variable in active data set > Computer new variable`;
- Na janela que se abre coloque em `New variabel name` o nome `ageCenter` e em `Expression to compute` inclua a expressão `age - mean(age)`;



- Faça o mesmo para uma nova variável com o nome `gestationCenter` usando a expressão `gestation - mean(gestatation)`;
- Construa o modelo selecionado utilizando estas novas variáveis contínuas centralizadas em substituição às originais;
- Refaça o calculos dos VIFs para esse novo modelo com as variáveis selecionadas. Guarde o resultado.
- Para o modelo final selecionado, com as variáveis preditoras contínuas centralizadas:
  - avalie os gráficos diagnósticos;
  - faça a avaliação da colinearidade entre os termos do modelo;
  - identifique qual(is) nível(is) está(ão) representado(s) no intercepto;
  - interprete cada um dos parâmetros do modelo, incluindo interações, se houver;
- A partir dos resultados do modelo proponha uma campanha para evitar que bebês nasçam com baixo peso.

Retorne à lista de desafios dos modelos com múltiplas preditoras do início desta seção e avalie se todos os pontos foram compreendidos.

Exercício

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