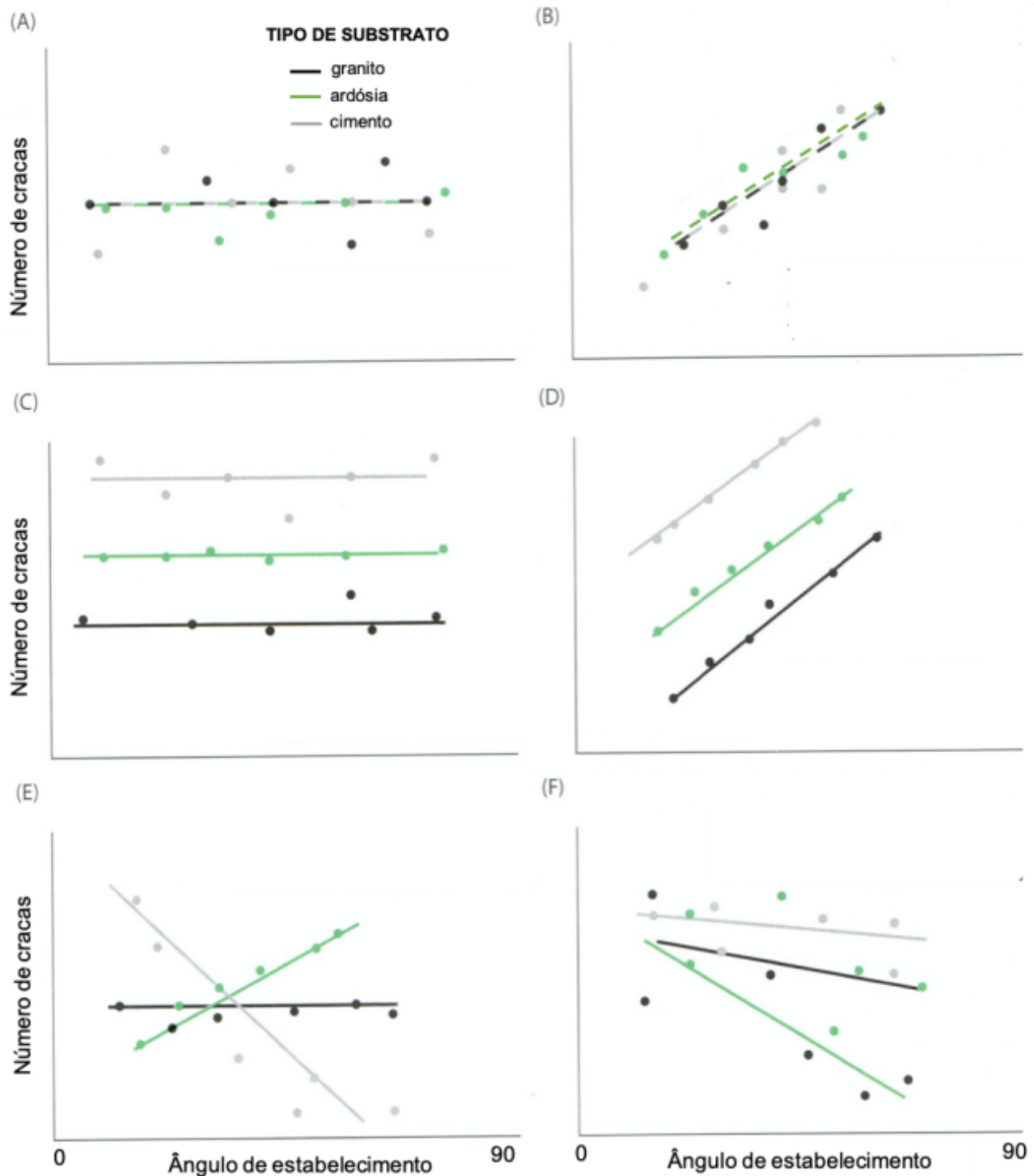


Modelos Lineares Múltiplos II

Entendendo a interação graficamente

Uma das melhores maneiras de entender os resultados de um modelo que envolve mais do que uma variável preditora é visualizar graficamente os possíveis resultados do estudo. Nem sempre isso é possível, quando o modelo é mais complexo. Para exemplificar a interpretação gráfica do modelo vamos aqui supor um estudo experimental realizado para testar o efeito do tipo de substrato (ardósia, granito ou cimento) e do ângulo em que a placa de substrato é colocada (variável contínua de 0-90 graus) sobre o recrutamento de cracas na zona intertidal rochosa (retirado de Gotelli & Ellison 2011). A figura a seguir representa os possíveis resultados quanto à significância dos dois fatores (efeito do substrato e do ângulo) e da interação entre substrato e ângulo. Cada painel (A-F) indica um diferente resultado experimental em relação à significância destes três termos do modelo (substrato, ângulo e substrato:ângulo). Em cada painel, os pontos indicam réplicas em cada um dos níveis de substrato. As linhas representam o melhor ajuste linear para cada tratamento.



Tarefa 1

Para cada um dos painéis acima, use o [link](#) para responder sobre a significância do substrato, do ângulo e da interação.

Agora queremos supor sobre o sinal e a significância dos coeficientes do modelo. ATENÇÃO: os coeficientes do modelo não correspondem aos fatores. Os coeficientes são aqueles que aparecem quando pedimos o “summary” do modelo no R. Neste exemplo das cracas, o modelo estatístico estimará 6 coeficientes: intercepto (ardósia); $\beta \cdot \text{cimento}$; $\beta \cdot \text{granito}$; $\beta \cdot \text{ângulo}$; $\beta \cdot \text{ângulo} \cdot \text{cimento}$ e $\beta \cdot \text{ângulo} \cdot \text{granito}$. A seguir uma breve descrição do significado de cada

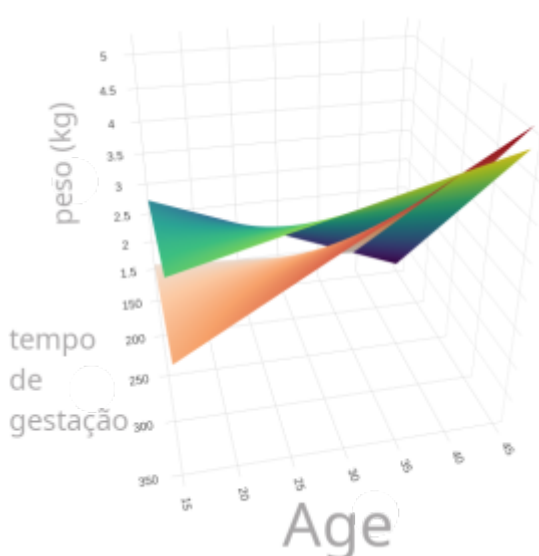
coeficiente:

- **Intercepto (ardosia):** corresponde ao número estimado de cracas sobre ardosia quando o ângulo é igual à zero. Neste caso o teste de significância testa se este valor é diferente de zero.
- **Beta*cimento:** efeito do cimento em comparação ao intercepto (ardosia) quando ângulo igual a zero. Se positivo é porque o número de cracas no cimento é maior do que na ardosia e vice-versa.
- **Beta*granito:** efeito do granito em comparação ao intercepto (ardosia).
- **Beta*ângulo:** efeito do ângulo para o tratamento ardosia (representado no intercepto). Se positivo é porque o número de cracas aumenta com o ângulo sobre o substrato ardósia. Se esse coeficiente for negativo o número de cracas diminui conforme aumenta o ângulo.
- **Beta*ângulo*cimento:** Efeito do ângulo no substrato cimento em comparação ao efeito do ângulo no tratamento ardosia (referência). Caso o valor do coeficiente seja positivo significa que o substrato cimento aumenta a inclinação do efeito do ângulo sobre o estabelecimento de cracas em comparação ao tratamento ardósia (seja deixando-a mais positiva ou menos negativa) e se negativo é porque o substrato cimento reduz a inclinação da relação em comparação à ardósia (seja deixando a relação mais negativa ou menos positiva).
- **Beta*ângulo*granito:** Efeito do ângulo no substrato granito em comparação ao efeito do ângulo no tratamento ardosia. Se positivo é porque o substrato granito aumenta a inclinação da relação em comparação à ardósia (seja deixando-a mais positiva ou menos negativa). Caso o valor do coeficiente seja negativo é porque o substrato granito reduz a inclinação da relação entre ângulo e número de cracas em comparação à ardósia (seja deixando a relação mais negativa ou menos positiva).

Tarefa 2

Para os painéis C e E, use o [link](#) para responder sobre o sinal dos coeficientes do modelo.

Modelos Lineares Múltiplos: várias preditoras



Nesse último tópico de modelos lineares múltiplos 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.

Além da interação entre as preditoras, tratada no roteiro 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 interação¹⁾, 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 previsões, 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.

Colinearidade

Uma outra questão que emerge dos modelos com múltiplas preditoras contínuas é a multicolinearidade. Quando duas variáveis 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 muita 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 dos 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](#).

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 está associado a muitos riscos ao bebê e maiores custos devido ao prolongamento

da 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 transformá-la em uma escala de medida que temos mais facilidade para interpretar, multiplicando essa variável por 0.02835 para transformá-la em kg. Vamos fazer o mesmo para o peso da mãe que foi medido em libras e deve ser multiplicado por 0.453592 para ser transformado em kg.

- Abra o arquivo

babies.csv

no Rcmdr²⁾;

- 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 nomeie a nova variável como bebeKg;
- Na caixa Expression to compute coloque a expressão: `bwt * 0.02835`;
- Em seguida crie a variável maeKg multiplicando weight por 0.453592.

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 height. Ao final ficamos com quatro variáveis preditoras: gestation, age, maeKg 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 ³⁾ e suas interações. O modelo resultante tem a seguinte expressão:

```
bebeKg ~ age * gestacion * maeKg * 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 4 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) ⁴⁾. 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:

$$\begin{aligned} C^2_4 &= \frac{4!}{2! (4-2)!} = 6 \\ C^3_4 &= \frac{4!}{3! (4-3)!} = 4 \\ C^4_4 &= \frac{4!}{4! (4-4)!} = 1 \end{aligned}$$

A lista de interações triplas e quádruplas tem essas expressões:

- `gestation:age:maeKg:smokeTRUE`
- `age:gestation:maeKg`
- `age:gestation:smoke`
- `age:maeKg:smoke`
- `gestation:maeKg:smoke`

As interações duplas:

- `age:gestation`
- `age:maeKg`
- `age:smoke`
- `gestation:maeKg`
- `gestation:smoke`
- `maeKg: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 a possibilidade de interações é 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:
 - `bebeKg ~ 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:maeKg: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:

bebeKg ~ gestation + age + maeKg + smoke + gestation:age + gestation:smoke + age:maeKg + age:smoke + maeKg:smoke + gestation:age:smoke

Seleção do mínimo adequado

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

```
bebeKg ~ gestation + age + maeKg + smoke +  
          gestation:age + gestation:smoke +  
          age:maeKg + age:smoke + maeKg:smoke +  
          gestation:age:smoke
```

- simplifique esse modelo até o mínimo adequado;

Durante o processo de simplificação, quando nos deparamos 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

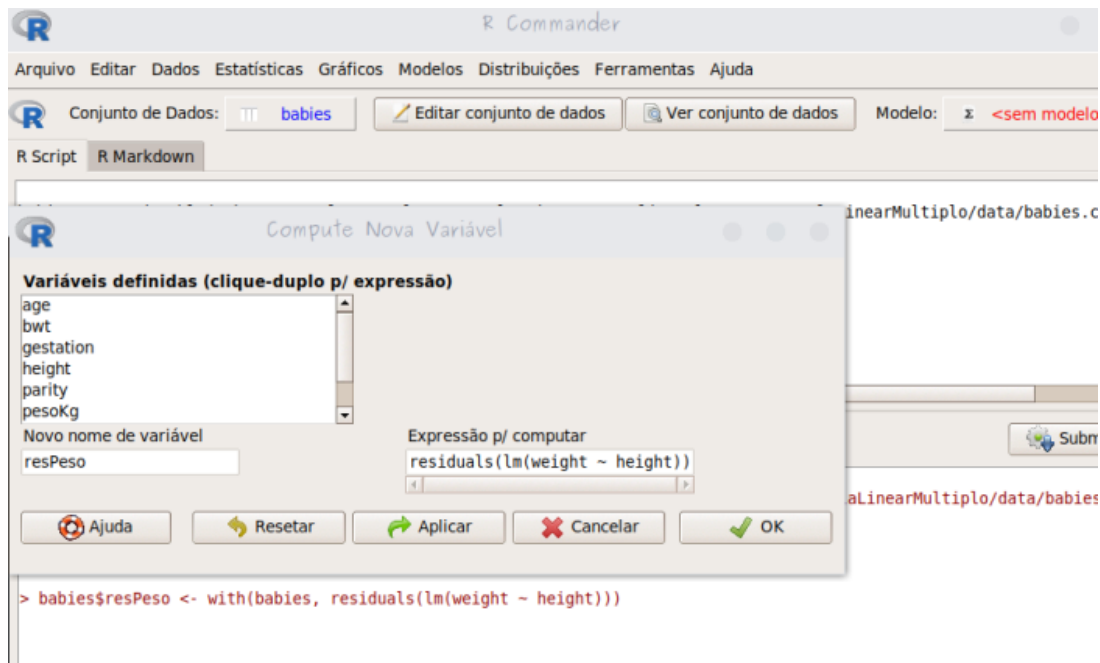
tornar “significativo” em um modelo mais simples.

- qual a predição do modelo selecionado do peso do bebê das seguintes mães:
 - 26 anos, 280 dias de gestação, peso 55 kg e não fumante
 - 26 anos, 280 dias de gestação, peso 55 kg e fumante
 - 26 anos, 210 dias de gestação, peso 55 kg e não fumante
 - 26 anos, 210 dias de gestação, peso 55 kg e fumante
 - 26 anos, 280 dias de gestação, peso 70 kg e não fumante
 - 26 anos, 280 dias de gestação, peso 70 kg e fumante
 - 26 anos, 210 dias de gestação, peso 70 kg e não fumante
 - 26 anos, 210 dias de gestação, peso 70 kg e fumante
- faça as mesmas predições acima, agora para uma mãe com 35 anos;
- coloque as características da mãe e as predições do modelo para os casos acima em um planilha ⁵⁾;
- interprete o resultado do modelo biologicamente.

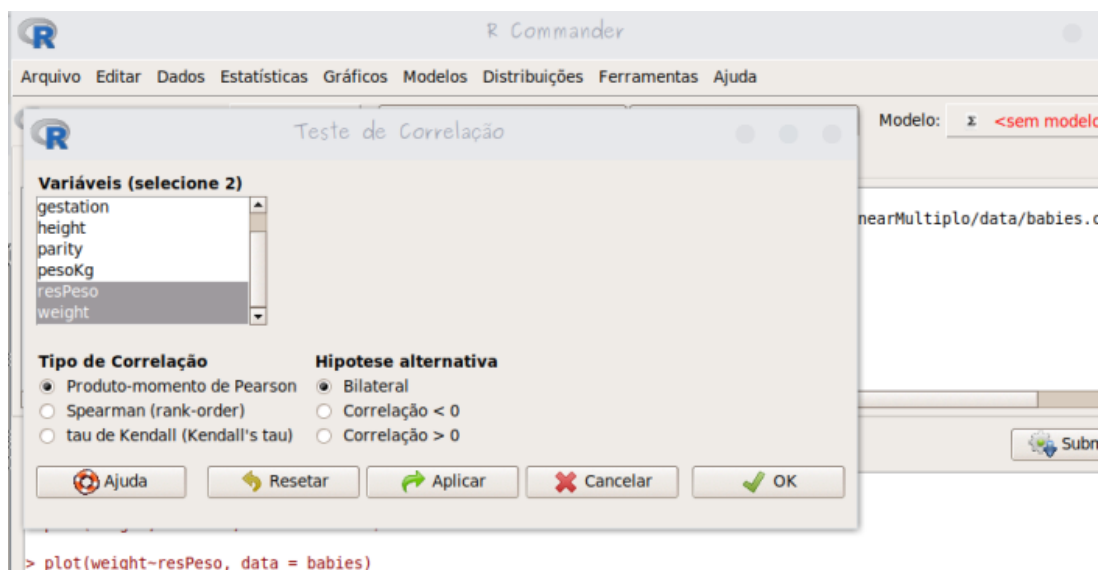
Tamanho e sobrepeso da mãe

Um “outro especialista”, ao analisar o modelo mínimo adequado selecionado acima, afirmou que a variável `height` deveria ser incluída para controlar o peso do bebê ser maior apenas porque a mãe era grande. Além disso, sugeriu incluir uma nova variável que indicasse a quanto a mãe tem de sobrepeso ou subpeso antes de engravidar. Indicou ainda que uma forma de criar essa variável seria utilizar os resíduos de uma regressão simples da variável peso como resposta e a variável altura como preditora. Todos os “outros especialistas” concordaram que eram boas sugestões, já que a nova variável representa o quanto a mãe tem mais peso ou menos peso do que esperado para uma mãe com a mesma altura.

- inclua a variável `height` no modelo mínimo selecionado anteriormente e verifique se esta 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;
- construa 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 **maeKg**, 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.

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;

O que preciso entregar

Link formulário

1)

três interações duplas e uma tripla

2)

os campos neste arquivo são separados por tabulação

3)

gestation, age, maeKg e smoke

4)

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

5)

essa planilha pode ser feita no R como um `data.frame`, exportando depois como csv

From:

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

Permanent link:

<http://labtrop.ib.usp.br/doku.php?id=cursos:planeco:roteiro:09-lm02a>

Last update: **2025/05/29 19:02**

