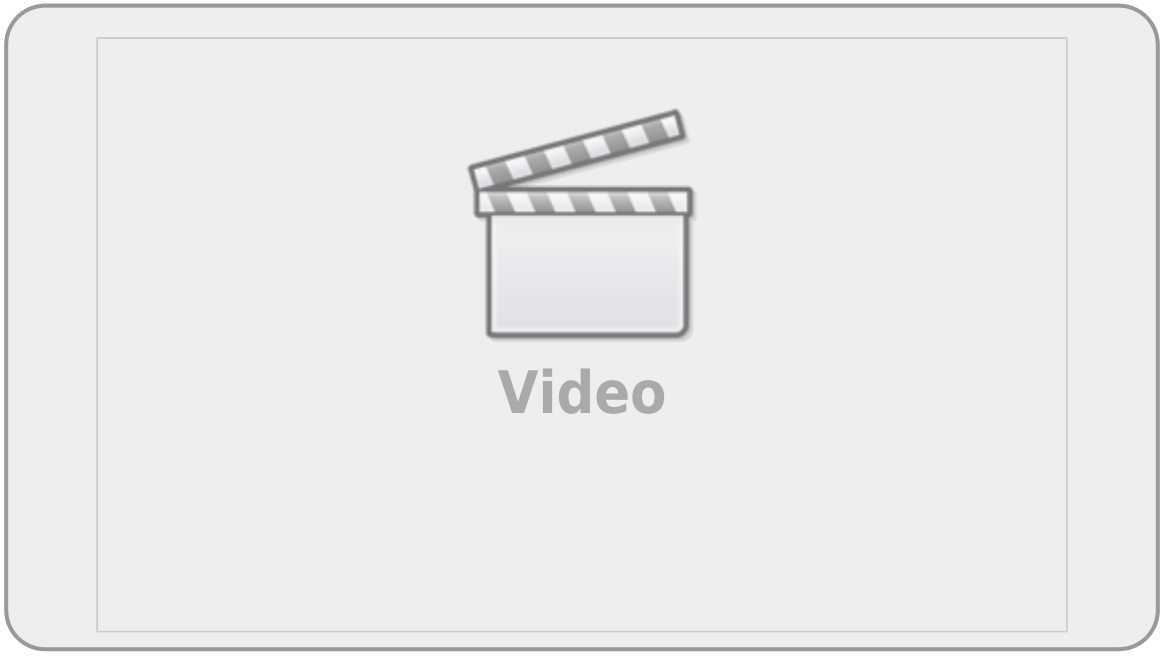
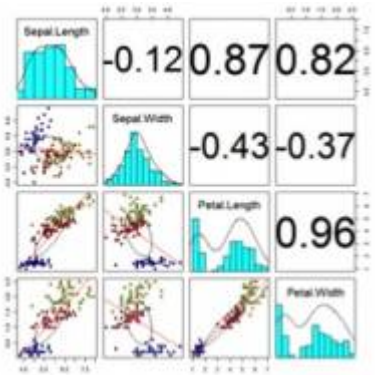
Modelos Lineares Múltiplos III

Interação entre preditoras e Colinearidade



Neste terceiro e último roteiro sobre Modelos Lineares Múltiplos vamos trabalhar com conjuntos de dados com **variáveis preditoras contínuas**. Inicialmente, iremos avaliar uma importante premissa dos modelos associados a esse tipo de dado, aprendendo a identificar colinearidade entre variáveis, entendendo os efeitos sobre a seleção de modelos e interpretando os coeficientes do modelo selecionado. Depois, vamos utilizar um conjunto de dados mais complexo com **variáveis contínuas e categóricas** para reforçar o procedimento de seleção de modelos e posteriormente obter e interpretar, por meio dos resumos dos modelos, os coeficientes dos parâmetros, incluindo as interações. Nessa segunda parte vamos exercitar todos os conceitos importantes de Modelos Lineares Múltiplos e fechar esse módulo da disciplina.

Colinearidade entre variáveis



Uma importante premissa de modelos lineares múltiplos é que as variáveis preditoras sejam

independentes entre si. Entretanto, em estudos observacionais ou exploratórios é relativamente comum que as variáveis preditoras não sejam independentes. Quando duas variáveis preditoras estão correlacionadas e estão explicando a mesma porção da variância da variável resposta estamos diante de um problema de colinearidade. Nos casos mais extremos, a colinearidade pode afetar a significância de algumas variáveis e até mesmo o sinal do efeito.

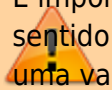
Existem várias formas de lidar com a colinearidade, mas vamos focar nossa atividade em identificar e remover variáveis que estejam inflando as estimativas de variação. Para isso, vamos usar um índice chamado de Variance Inflation Factor (VIF), que é calculado a partir dessa equação:

$$ VIF_{i} = \frac{1}{1 - R_{i}^{2}} $$

Esse R^2 é obtido ajustando um modelo linear múltiplo para analisar a relação entre cada variável preditora, por exemplo $i = X_1$, e todas as outras preditoras no modelo de interesse (X_2, X_3,..., X_n). Fazendo isso para todas as preditoras teremos um `VIF` para cada uma delas. Um alto R^2 significa que grande parte da variação na preditora em questão é compartilhada pelas outras variáveis. Veja no exemplo abaixo um passo-a-passo para calcular o `VIF`.

Quanto maior for o valor do `VIF`, mais os valores de erro padrão dos parâmetros do modelo serão inflados e mais dificilmente um efeito será detectado. Além da imprecisão nas estimativas dos parâmetros colineares, um outro problema que pode emergir é o modelo mínimo adequado ser diferente, dependendo da ordem da simplificação do modelo cheio. Um valor frequentemente usado para definir um limite aceitável de `VIF` é 4,0, acima desse valor as estimativas do modelo podem ficar comprometidas.

Nessa abordagem, após identificar quais são as variáveis com maiores valores de VIF, elas serão removidas sequencialmente. A cada variável retirada verifica-se novamente se os valores de VIF diminuíram ou se ainda precisam ser retiradas outras variáveis colineares.



É importante entender que a escolha de qual variável retirar vai depender também do sentido biológico/ecológico de cada variável. Em alguns casos, pode valer a pena manter uma variável cujo VIF é levemente mais alto, pois o mecanismo de explicação pode ser mais explícito para essa variável.

Vamos ver como isso funciona na prática abaixo.

Biomassa de manguezais e variáveis ambientais

O objetivo dessa pesquisa foi avaliar quais variáveis ambientais predizem melhor a biomassa acima

do solo (Aboveground Biomass-AGB) de manguezais em diferentes locais do mundo[1]. Foram utilizadas 3 variáveis ambientais que são facilmente obtidas em bases de dados mundiais.

1. Baixe o conjunto de dados

mangrove.csv

, importe para o Rcmdr, usando vírgula como separador de campo, e visualize os dados para entender o arquivo.

2. Entenda as variáveis do arquivo:

- variável resposta:
  - **AGB_carbon** = Estoque de Carbono estimado a partir da biomassa acima do solo (AGB) de árvores de manguezais (MgC/ha) no hemisfério sul
- variáveis preditoras:
  - **lat** = latitude (em graus)[2]
  - **temp** = Temperatura média anual (em graus Celsius)
  - **ppt** = Precipitação anual (mm)

3. Inspecione a correlação entre todas as variáveis preditoras contínuas:

Para fazer isso no *Rcmdr*, você tem duas opções:

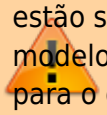
- Uma opção é avaliar numericamente as correlações. Para isso, entre em **Statistics→ Summary → Correlation Matrix**, selecione todas as variáveis preditoras e clique em OK. Você verá os valores de correlação de todos os pares de variáveis. Observando os valores mais altos de correlação, você já pode ter uma ideia se existem variáveis com potencial para apresentar colinearidade.
- Outra opção é avaliar graficamente as correlações entre as variáveis. Para isso, entre em **Graphics → Dispersion Matrix** e selecione todas as variáveis preditoras contínuas. Na aba **Options** selecione “Minimum Square Line” e clique em OK. Na figura que foi gerada, você poderá avaliar quais pares de variáveis parecem ter uma maior correlação entre elas.

Esse procedimento de analisar a correlação entre todas as nossas variáveis preditoras contínuas deveria ser sempre realizado antes de fazermos nossas análises.

4. Ajuste um modelo, relacionando AGB-carbon com todas as variáveis preditoras, mas ainda sem incluir as interações. Nomeie esse modelo como “carbon1”. No *summary* do modelo, repare nos efeitos e na significância de cada um dos parâmetros.

5. Calcule os VIFs[3] para as variáveis incluídas no modelo

Para isso, entre em **Models → Numerical diagnostics → Variance-inflation factors**. O primeiro resultado apresentado é uma linha com os valores de VIF para cada parâmetro do modelo. O segundo resultado apresentado é uma matriz de correlação das estimativas dos parâmetros. Note que os valores são diferentes das correlações feitas diretamente para as variáveis (item 3, acima).



Importante: Como o valor de VIF de cada parâmetro depende de quais outros parâmetros estão sendo incluídos no modelo, só é possível calcular os VIFs depois de ter ajustado um modelo. Ao usar o *Rcmdr*, fique sempre atento(a) se o modelo ativo é realmente o modelo para o qual você quer calcular os VIFs.

PAUSA OPCIONAL caso você queira aprender a calcular manualmente os valores de VIF:

Em primeiro lugar, reveja a equação de cálculo de VIF apresentada acima.

Agora, vamos calcular manualmente o valor de VIF para a variável preditora **lat** e comparar com o valor obtido acima no *Rcmdr*. Para isso, precisamos calcular o R^2 da relação entre essa variável preditora e todas as outras preditoras que estavam no modelo completo, sem as interações (carbon1). Para isso, vamos criar um novo modelo no qual a variável para a qual estamos interessados em calcular o VIF (**lat**) passará agora a ser a variável resposta desse novo modelo que criaremos.

Entre em **Statistics → Fit models → Linear model**. Coloque **lat** como variável resposta na caixa da esquerda da equação e coloque as outras 2 variáveis preditoras (**temp** + **ppt**) na caixa da direita da equação. Defina o nome desse modelo como “viflat”. No *summary* do modelo será apresentado o valor de R^2 Múltiplo (*Multiple R-square*). Utilize esse valor na equação de cálculo de VIF e veja se o resultado é igual ao valor de VIF calculado pelo R Commander para a variável **lat** a partir do modelo “carbon1” feito acima. Deveria ser. Se não foi, peça ajuda a alguém da equipe.

Repita o mesmo procedimento para outra variável de sua escolha. Você pode fazer isso para todas as variáveis do modelo, se quiser.

Continuando nossa análise sobre o estoque de Carbono em manguezais:

6. Após analisar os valores dos VIFs do modelo “carbon1”, se houver alguma variável com valor maior que 4, remova a variável com o maior VIF e ajuste um novo modelo. Coloque “carbon2” como nome desse modelo. Olhe para o *summary* desse modelo e para as variáveis que permaneceram nele. Cheque os valores dos coeficientes e a significância de cada variável em relação ao modelo “carbon1”. **Houve alguma alteração? Alguma variável deixou de ser significativa? Alguma variável passou a ser significativa? O sinal do efeito mudou?**

7. Calcule os VIFs das variáveis do modelo “carbon2” usando o caminho **Models → Numerical diagnostics → Variance-inflation factors** e veja se ainda tem alguma variável com VIF maior que 4.

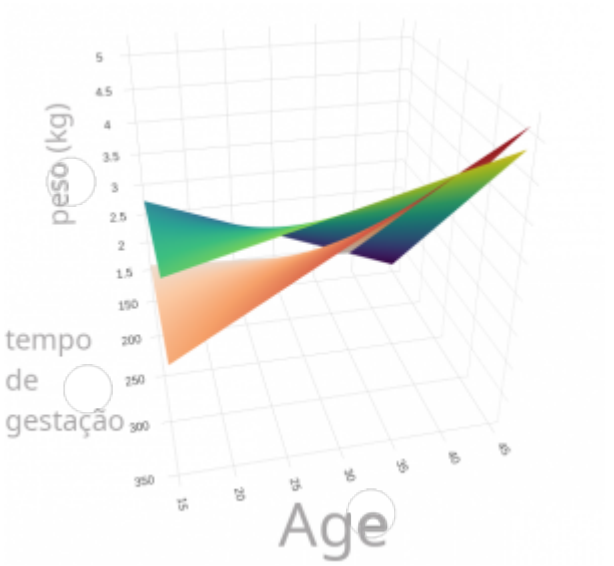
8. Repita os procedimentos anteriores até não haver nenhuma variável com VIF maior que 4.

9. E possível que algumas das variáveis remanescentes, mesmo que não sejam colineares entre si, não sejam relevantes para definir o estoque de carbono em manguezais. Então, para iniciar o procedimento de seleção de modelos, crie um modelo completo que inclua as variáveis remanescentes **e suas interações** (nomeie como “carbon_int”). Analise o *summary*do modelo.

10. Realize o procedimento de seleção do modelo mínimo plausível pelo método de simplificação para o mínimo adequado, conforme explicado no item Simplificando modelos do roteiro I de Modelos Lineares Múltiplos]]

11. Analise os resultados do modelo final.

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.

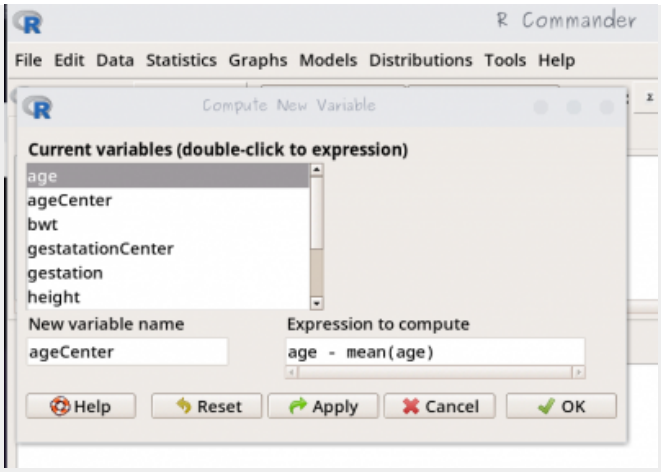
Peso de bebês ao nascer

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.

1)

dados fictícios, mas baseados em valores reais

2)

Os dados são predominantemente do hemisfério sul e por isso seriam esperados valores negativos de latitude, porém foram transformados em valores positivos para facilitar a interpretação. Alguns valores do hemisfério norte aparecem como negativos na planilha

3)

A função utilizada no *Rcmdr* para calcular o `VIF` utiliza uma variante chamada `GVIF`, uma generalização que pode ser aplicada também para variáveis categóricas com a mesma interpretação colocada acima o `GVIF` foi desenvolvida pelo John Fox, mesmo autor do Rcmdr. Veja o artigo no link https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475190