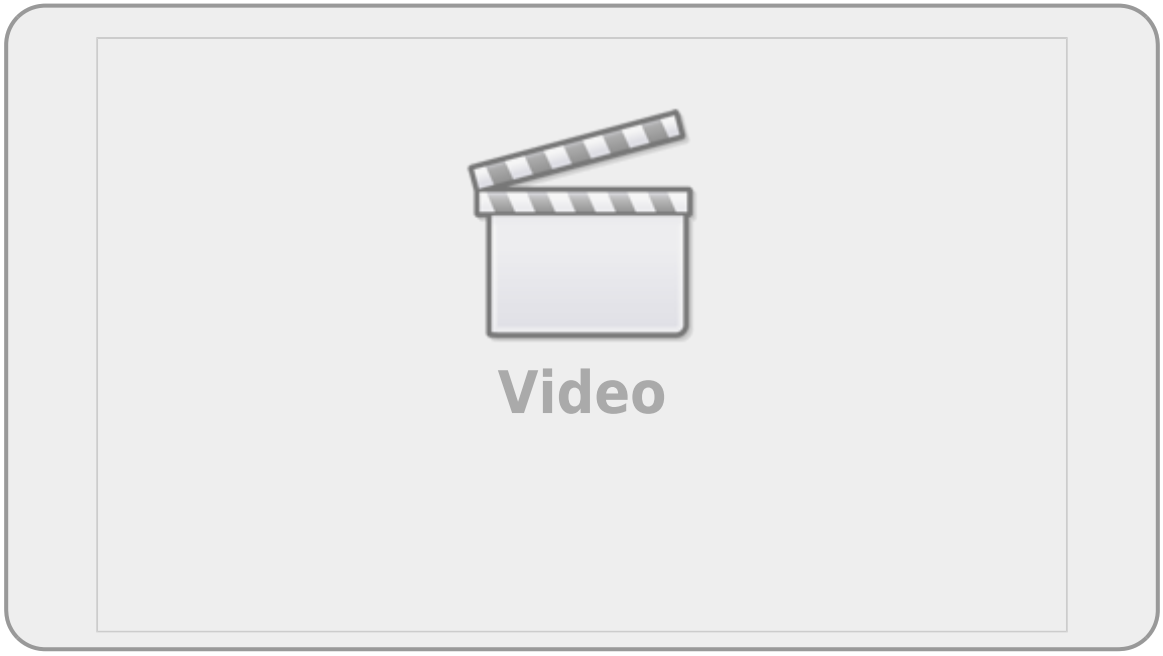
Modelos Lineares Generalizados

GLM: Introdução



Os modelos lineares generalizados (**GLMs**) são uma ampliação dos modelos lineares ordinários. Os **GLM's** são usados quando os resíduos (erro) do modelo apresentam distribuição diferente da normal (gaussiana). A natureza da variável resposta é uma boa indicação do tipo de distribuição de resíduos que iremos encontrar nos modelos. Por exemplos, variáveis de contagem são inteiras e apresentam os valores limitados no zero. Esse tipo de variável, em geral, tem uma distribuição de erros assimétrica para valores baixos e uma variância que aumenta com a média dos valores preditos, violando duas premissas dos modelos lineares. Os casos mais comuns de modelos generalizados são de variáveis resposta de contagem, proporção e binária, muito comum nos estudos de ecologia e evolução.

Devemos considerar os GLMs principalmente quando a variável resposta é expressa em:

- contagens simples
- contagem expressa em proporções
- número de sucesso e tentativa
- variáveis binárias (ex. morto x vivo)
- tempo para o evento ocorrer (modelos de sobrevivência)

GLM: componentes

Uma das formas de entendermos os modelos generalizados é separar o modelo em dois componentes: a relação determinística entre as variáveis (resposta e preditora) e o componente aleatório dos resíduos (distribuição dos erros). Em um modelo linear ordinário a relação entre as variáveis é uma proporção constante, o que define uma relação funcional de uma reta. Quando temos uma contagem, essa relação pode ter uma estrutura funcional de uma exponencial. Para esses casos, os modelos generalizados utilizam uma função de ligação `log` para linearizar a relação determinística entre as variáveis. Portanto, a estrutura determinística dos modelos **GLM's** é definida por um preditor linear, associada à função de ligação.

O componente aleatório dos resíduos, no caso de uma variável de contagem, segue, em geral, uma distribuição **poisson**. A distribuição **poisson** é uma variável aleatória definida por apenas um parâmetro (λ), equivalente à média, chamada de lambda. A distribuição **poisson** tem uma característica interessante, seu desvio padrão é igual à média. Portanto, se a média aumenta, o desvio acompanha esse aumento e a distribuição passa a ter um maior espalhamento.

Preditor linear e função de ligação

O preditor linear está associado à estrutura determinística do modelo e está relacionado à linearização da relação, aqui definido como η:

$$ \eta = \alpha + \beta x$$

A função de ligação é o que relaciona o preditor linear com a esperança do modelo:

$$ \eta = g^{-1}(E_{(y)}) $$

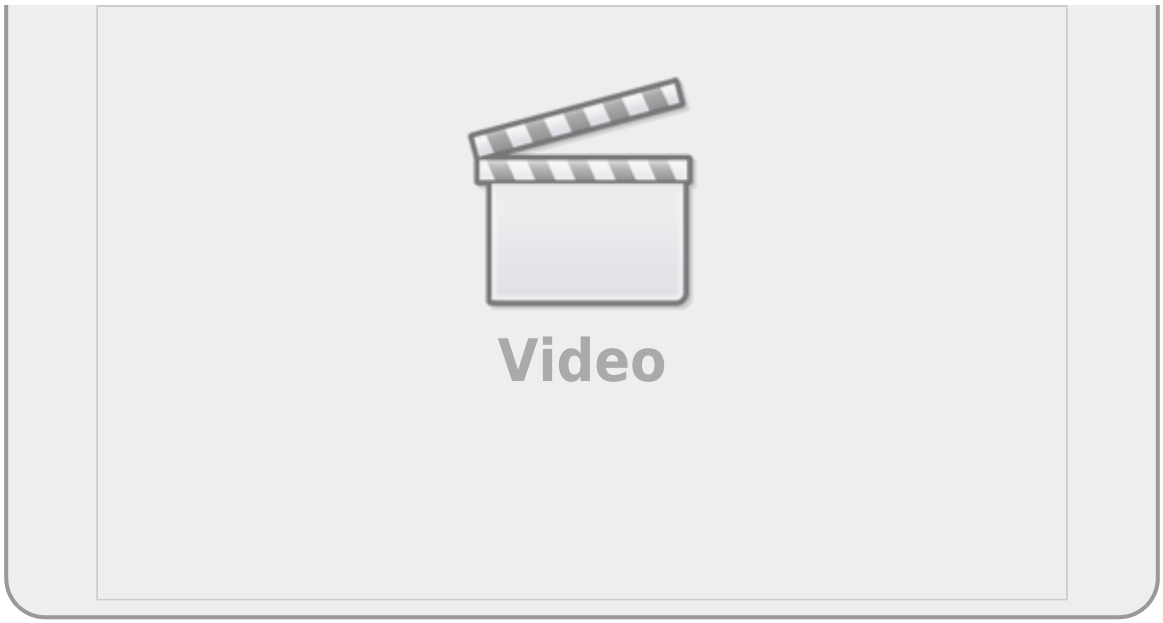
Ou seja, nos modelos generalizados não é a variável resposta que tem uma relação linear com a preditora, e sim o preditor linear que tem uma relação linear com as preditoras.

Funções de ligações canônicas

Para alguns tipos de famílias de variáveis temos funções de ligações padrões. As mais usadas são:

Natureza da resposta	Estrutura dos resíduos (erro)	Função de ligação
contínua	normal	identidade
contagem	poisson	log
proporção	binomial	logit

GLM Contagem



Contagem: um exemplo simples

Um exemplo, apresentado no livro do Michael Crawley, The R Book, relata a contagem de espécies de árvores em unidades amostrais de florestas com diferentes biomassa e classificadas em três níveis de ph no solo: baixo, médio e alto. O objetivo desse experimento não manipulativo é verificar se há relação entre riqueza de árvores e as preditora biomassa da floresta e ph do solo.

ATIVIDADE

Modelo Linear Múltiplo (LM)

1. Importe o arquivo
   species.txt
   para o Rcmdr. Note que esse arquivo tem como separador de campo a tabulação e decimal como ponto.
2. Monte o modelo linear clássicos (`lm`) para esse dados, tendo como variável resposta a riqueza de espécies(`Species`) e como preditoras o pH e `Biomass` e as interações possíveis.
3. Reduza o modelo cheio ao modelo mínimo adequado utilizando como critério de comparação a tabela de anova.
4. Utilizando os coeficientes estimados do modelo, faça a predição do número de espécies para:
   - um nível alto de pH com `Biomass` de **3.2**
   - um nível médio de pH com `Biomass` de **15.5**
   - um nível baixo de pH com `Biomass` de **7.2**

Modelo Linear Generalizado (GLM)

1. Repita o procedimento de simplificação a partir do modelo cheio, agora com modelo linear generalizado (`glm`) e com `family = poisson`.
   - Caso o Rcmdr não retorne o `p-valor` na comparação de modelos

por anova, copie a linha de código que foi utilizada com `anova(…)` e cole novamente incluindo `anova(…, test="Chisq")`

2. Calcule as mesmas predições acima para o modelo, usando os coeficientes do preditor linear do `glm`.
3. Transforme os preditos pelo modelo de volta para a escala de observação [1)].
4. Faça os gráficos apresentados no tópico Gráfico no Rcmdr

- Para a predição no `glm` utilize os coeficientes estimados pelo modelo.
- Após estimar o predito na escala linear, transforme a predição para a escala de observação.
- Como usamos o `log` como função de ligação, para retornar a escala da observação devemos utilizar o antilog, no caso, a função exponencial.

Gráfico no Rcmdr

Gráfico dos dados

No menu **Graphs**, selecione **XY conditioningh Plot** e selecione as variáveis, definindo ***ph*** como variável de agrupamento, como no gráfico abaixo.

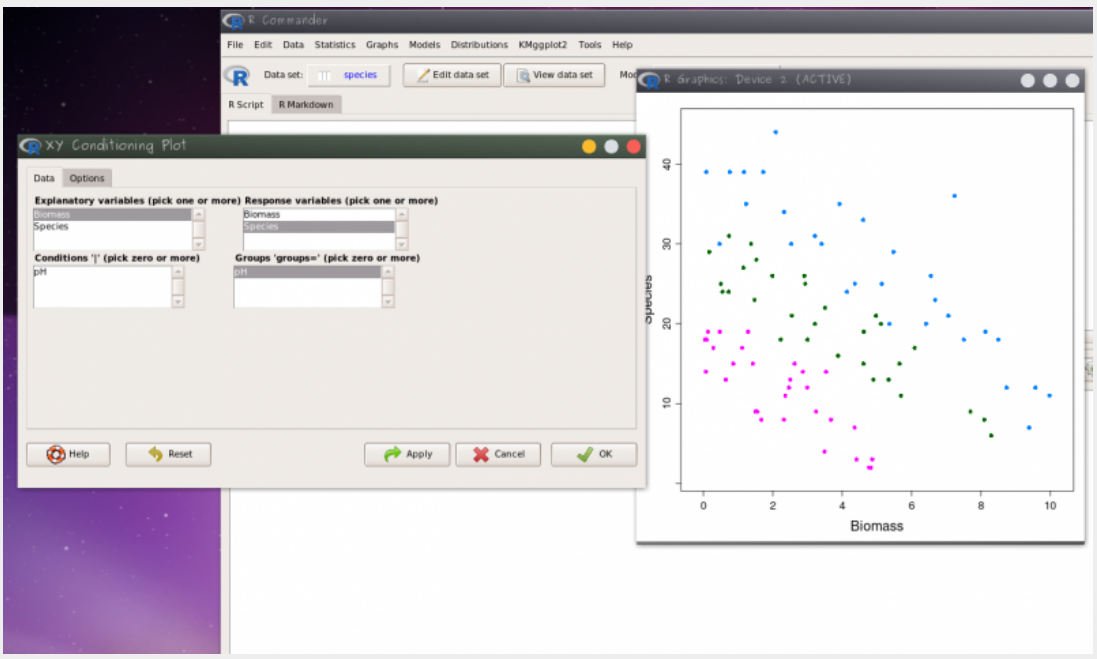
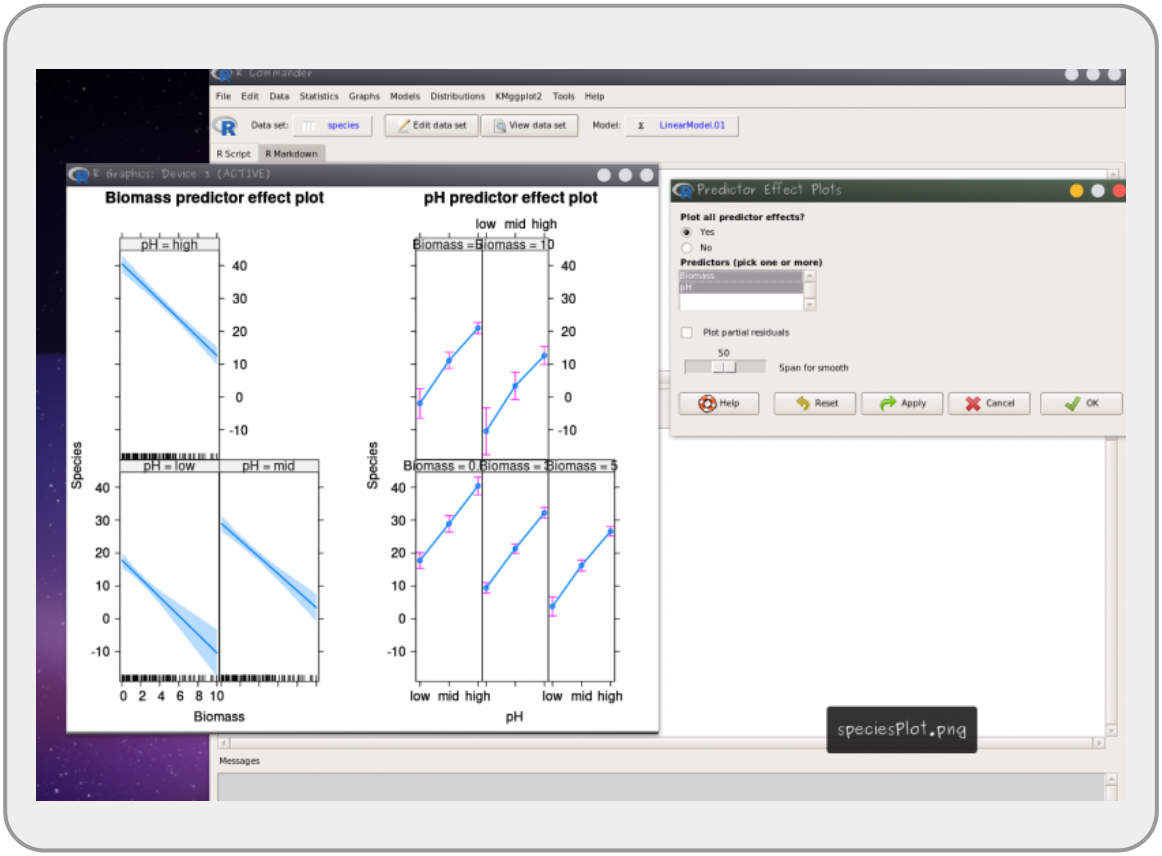


Gráfico dos Modelos

No menu **Models>Graphs** selecione **Predict effect plots...** e selecione as variáveis.



<u>Ordenando uma categórica</u>

O padrão do R é ordenar as variáveis categóricas por ordem alfabética. No exemplo seria desejável reordenar a variável categórica ***ph*** em uma categórica ordenada **low>medium>high**. Para reordenar utilize o menu **Data>Manager variable in active data set> Reorder factor levels**. Caso não deseje sobrescrever a variável original, forneça um novo nome para a variável reordenada.

O que preciso entregar

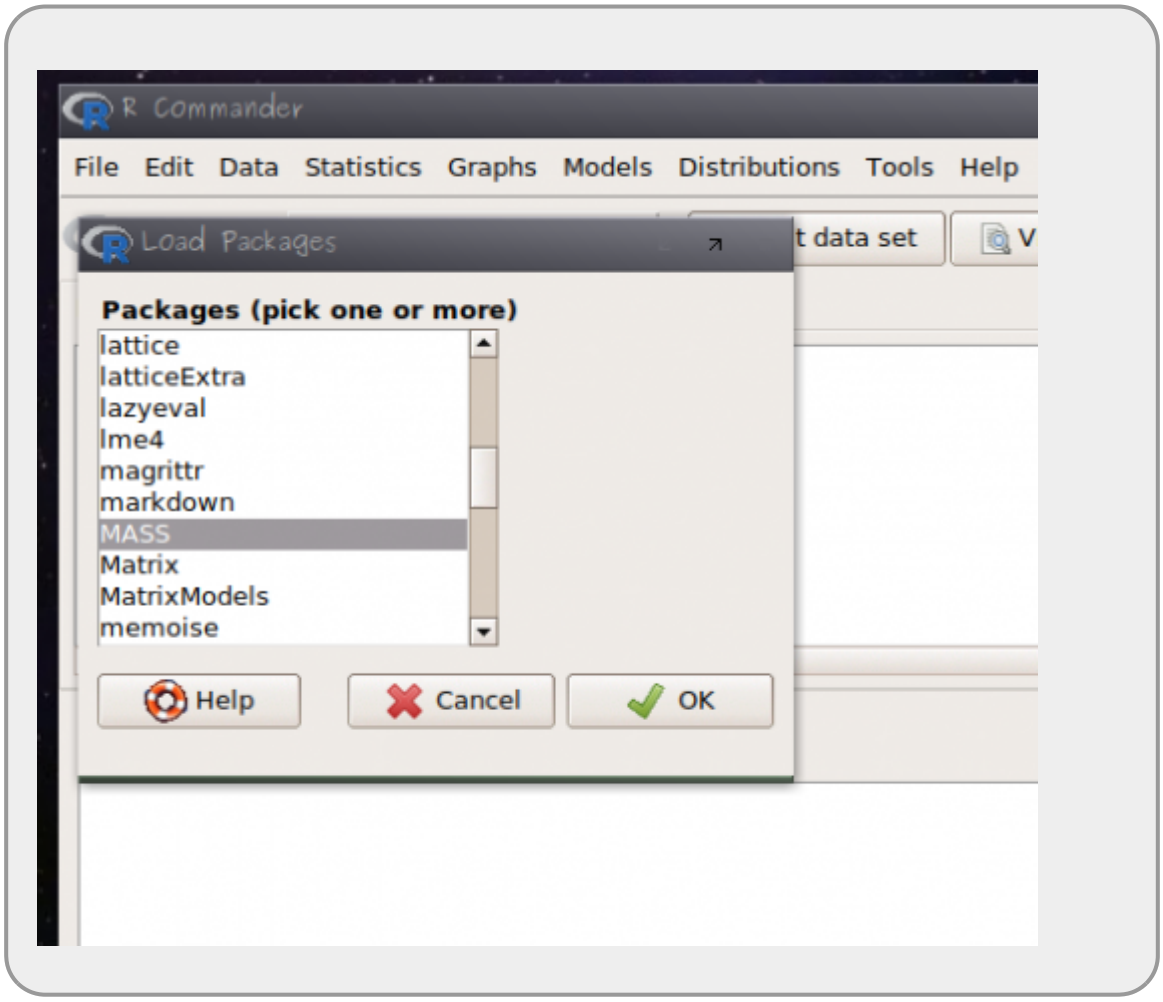
- Preencha as perguntas do quadro abaixo ou pelo link do formulário

Contagem: o que faz um aluno faltar às aulas

Vamos utilizar um exemplo que está presente no livro de W. Venables e B. Ripley, Modern Applied Statistics with S-PLUS[2], sobre o número de dias ausentes da escola de crianças na Austrália.

Carregando o pacote MASS

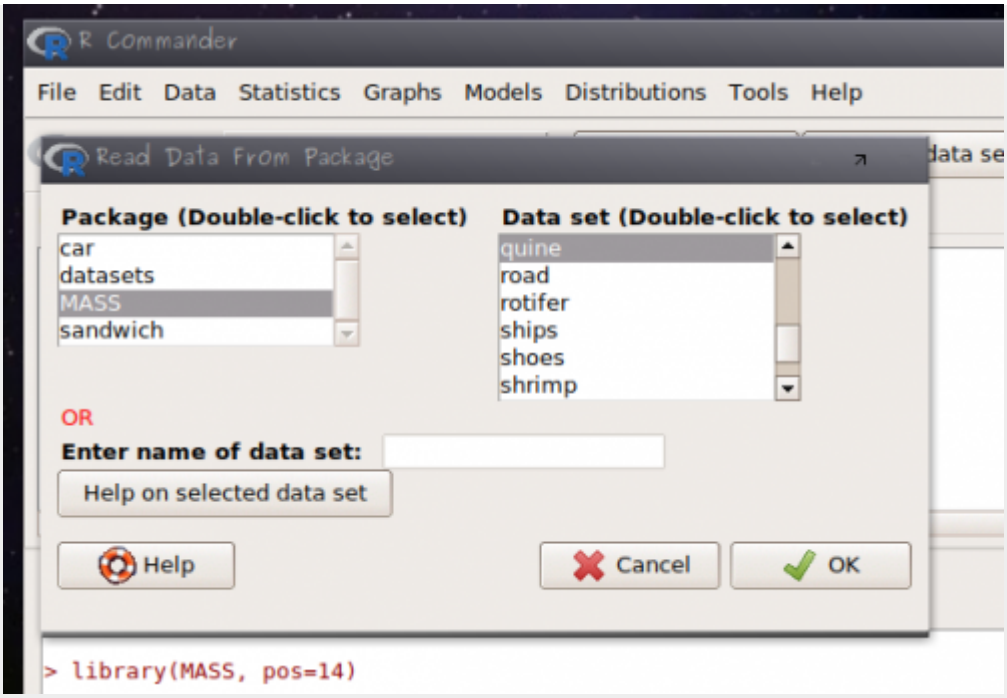
No Rcommader (Rcmdr) vá ao menu **Tools** > **Load package(s)** e selecione o pacote MASS.
Caso o pacote não apareça listado, significa que ele já está carregado, então pule esse passo.



Lendo os dados: quine

Em sequida:

- abra o menu **Data** > **Data in packages** > **Read data from an attached package...**
- selecione o pacote **MASS** e os dados **quine** [3]



Entendendo os dados: quine

Os dados estão relacionados ao estudo para entender quais variáveis estão relacionados à ausência (falta) do aluno na escola. A observação está relacionada a alunos amostrados aleatoriamente de escolas na Austrália.

- **Days**: variável resposta, número de dias ausente da escola
- **Eth**: origem aborígene (A) ou não (N)
- **Sex**: homem (M) ou mulher (F)
- **Age**: estágio de educação F0(primário)... quatro níveis.
- **Lrn**: classificação de aprendizado do aluno médio (AL) e fraco (SL)[4)]

Gráfico dos dados

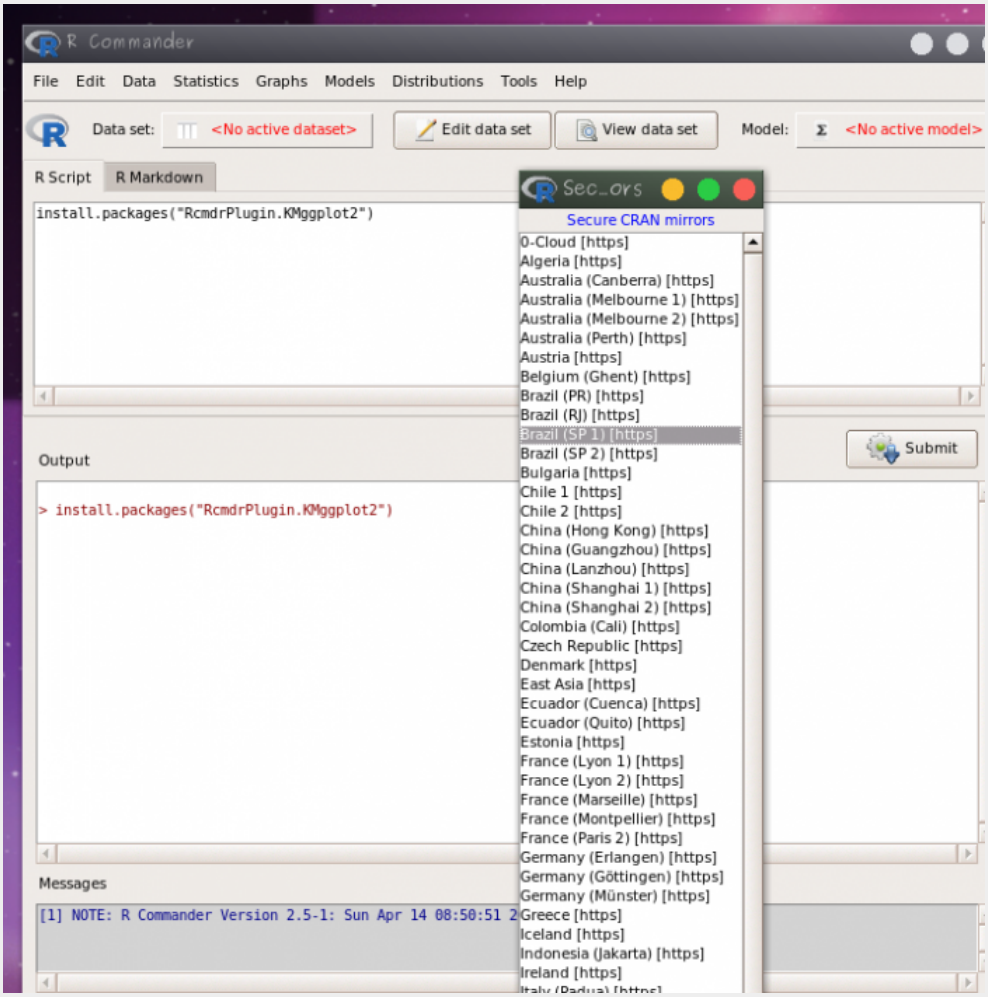
O pacote RcmdrPlugin.KMggplot2 é um plugin para Rcmdr que amplia as funções gráficas da interface. Instale o pacote copiando o comando abaixo no box superior do Rcmdr:

- **guarde os resultados dos modelos fora do Rcmdr pois a instalação e o carregamento do pacote solicita a reinicialização do Rcmdr**
- **após a instalação e carregamento do pacote, confira se os dados permanecem ativos, confira se precisará**

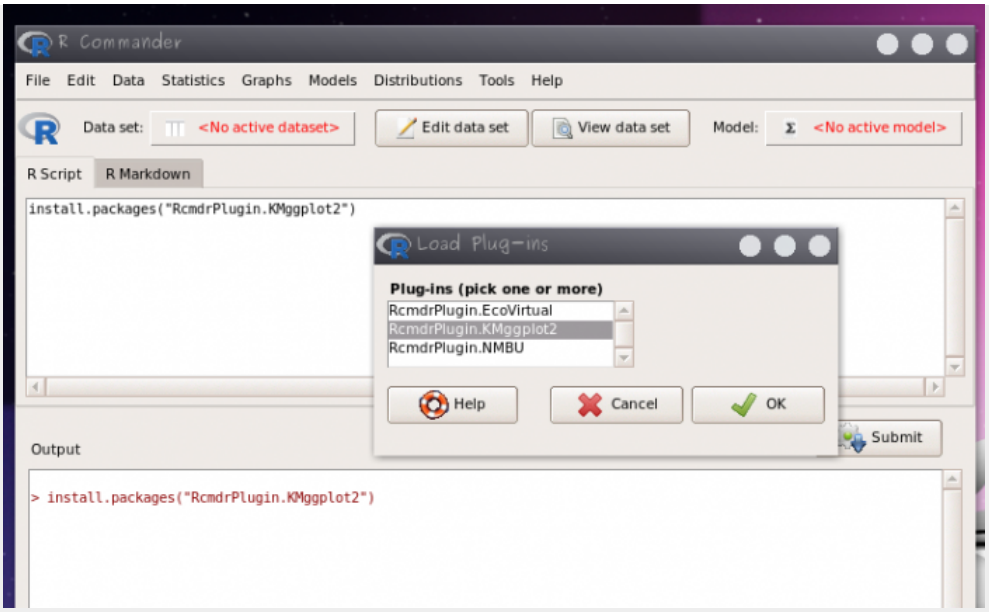
carregá-lo novamente

```
install.packages("RcmdrPlugin.KMggplot2")
```

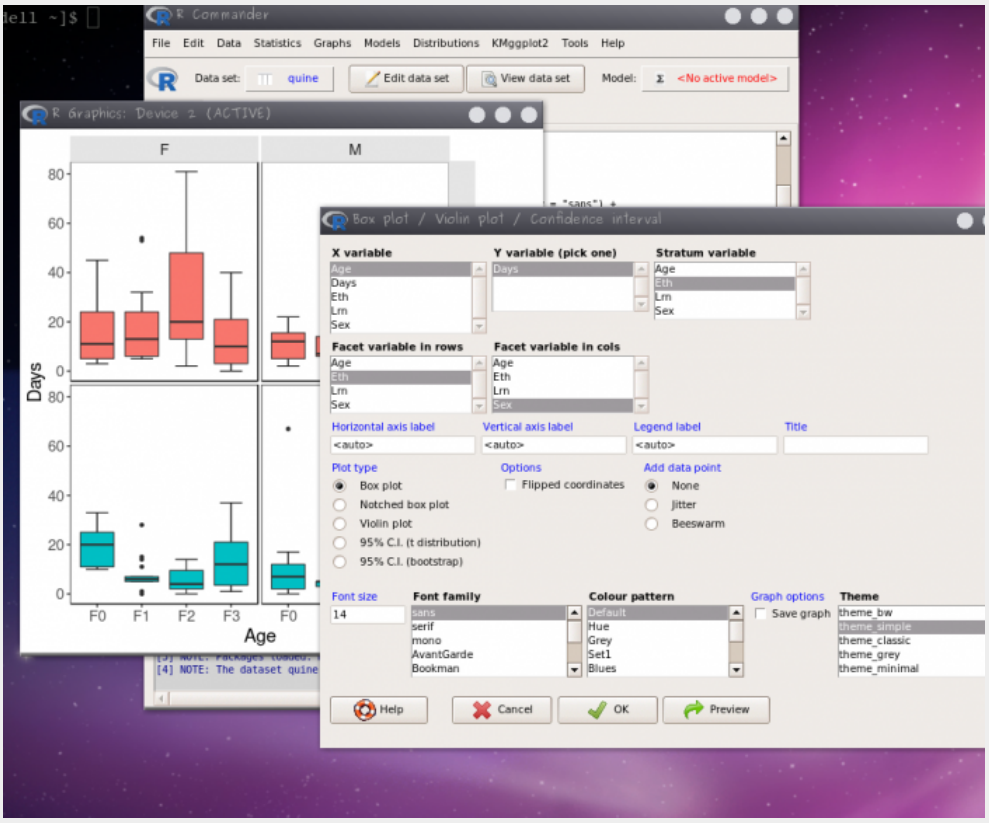
Em seguida, garanta que o cursor do mouse está na linha de comando e clique no botão **Submit**. Na janela que ira se abrir selecione o repositório **Brasil(SP1)**.



Para ativar o plugin selecione o menu **Tools> Load Rcmdr plug-in(s)...** e em seguida selecione o pacote `RcmdrPlugin.KMggplot2`.



- clique em sim na janela que solicita a reinicialização do Rcmdr
- clique na nova opção do menu **KMggplot2 > BoxPlot/...** e selecione as variáveis



Ajustando o GLM: dias fora da escola

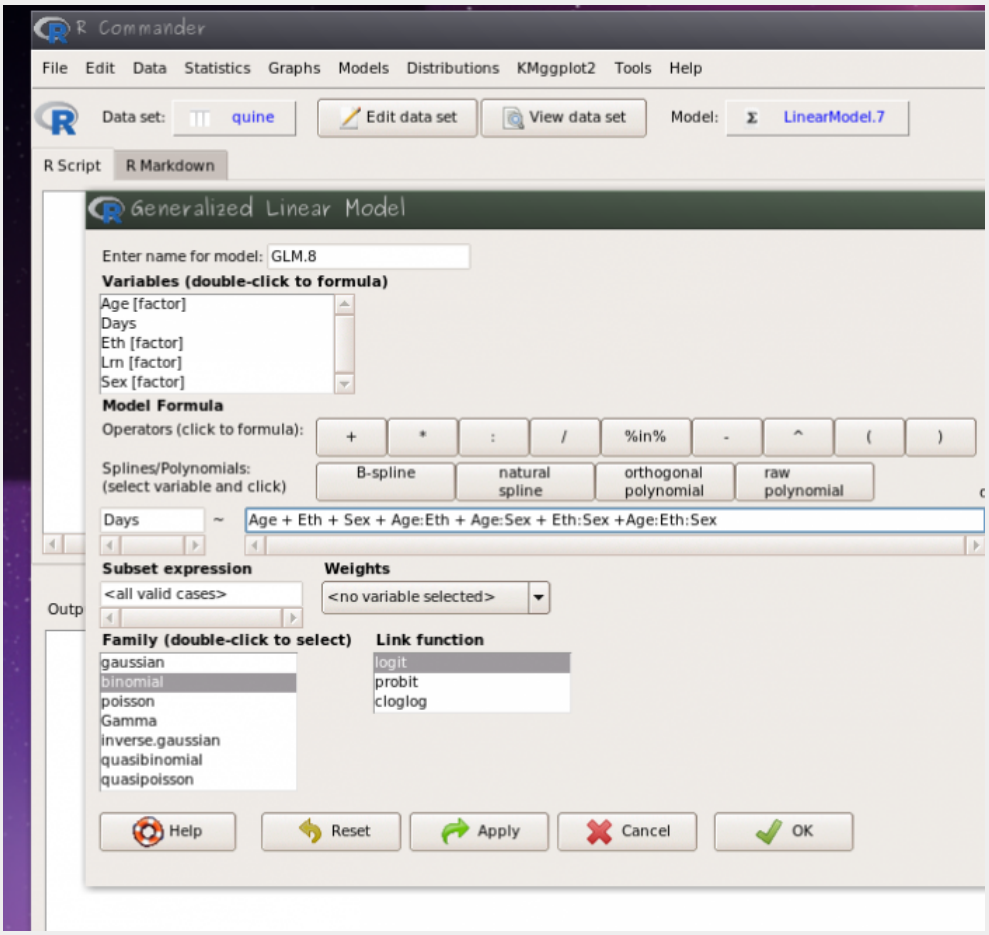
Atividade

Para nosso exercício vamos deixar de lado a variável `Lrn` por que há dados faltantes nela com relação a outras variáveis. Vamos construir o modelo cheio com a variável resposta `Days` e com as variáveis preditoras (`Eth`, Sex, Age) e todas as possibilidades de interações entre elas. Como estamos tratando de uma variável de contagem podemos partir direto para um modelo **GLM** indicando a família de distribuição de resíduos **POISSON** e a função de ligação ***log***.

- abra o menu **Statistics** > **Fit model** > **Generalized Linear Model**
- construa um modelo cheio com (**Age, Eth e Sex**) e as suas interações possíveis:

```
 Days ~ Eth + Sex + Age + Eth:Sex + Eth:Age + Sex:Age +
Eth:Sex:Age
```

- faça a simplificação do modelo para reduzir o modelo ao mínimo adequado



Diagnóstico do modelo

Um dos pressupostos do modelo Poisson é que a variância aumenta linearmente com a esperança (média do modelo). Podemos avaliar isso dividindo a `Residual Deviance` pelo seu `degrees of freedom`. Essa razão deve ser próxima a 1. O que não é o caso do nosso modelo. Nesses casos uma das alternativas é:

- ajustar o modelo usando **Family**: `quasipoisson`

Ajustando o GLM com sobredispersão

- monte o modelo cheio utilizando a família `quasipoisson` e
- siga em frente simplificando o modelo para o mínimo adequado
- interprete o modelo selecionado

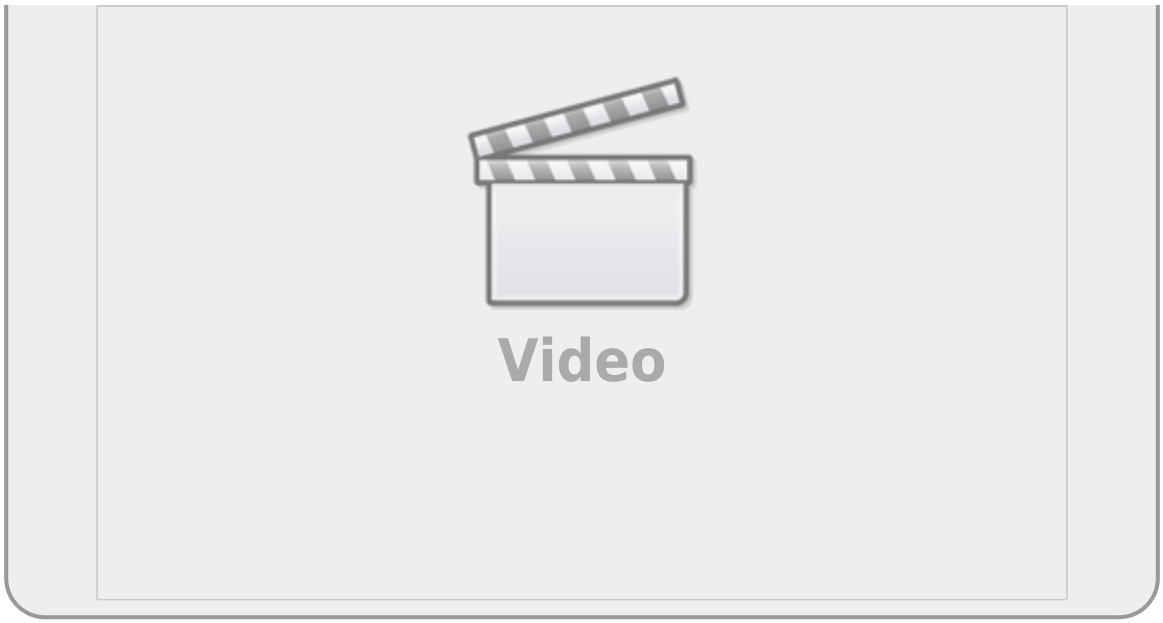
Gráfico do Modelo

O gráfico do modelo pode ser obtido no Rcmdr da mesma forma indicada no modelo anterior, no menu: **Models>Graphs** selecione **Predict effect plots...** e selecione as variáveis.

O que preciso entregar

- Preencha as perguntas do quadro abaixo ou pelo link do formulário

GLM Binário ou proporção



Os modelos de proporção (ou probabilidade) de sucessos (sucessos/tentativas), proporção simple (%) ou de resposta binária (presença/ausência, vivo/morto) são modelados, normalmente, com estrutura do erro binomial. Nesses casos, os limites dos valores da variável resposta é bem definido: entre 0 e 1. Além disso, a variância não é constante e seu valor difere com a probabilidade de sucessos. Estas características fazem com que os resíduos apresentem uma estrutura que aumenta e depois diminuí com o aumento da probabilidade de sucessos e o máximo de variância é encontrado nos valores intermediários (probabilidade de sucesso = 0.5).

A função Bernoulli, que é a base para a binomial, é definida pelo parâmetro de probabilidade de sucesso em um evento com duas possibilidades de resultado (binário). O parâmetro da função Bernoulli é a probabilidade de sucesso. No caso de uma moeda justa seria a probabilidade de 0,50 de sair coroa [5)].

A binomial é uma generalização da Bernoulli, definida pelo número de sucessos em certo número (n) de tentativas (número de eventos Bernoulli). Um exemplo de um experimento binomial seria a estimativa da probabilidade de germinação (sucessos) de um experimento onde temos 20 sementes (número de tentativas).

Conceitos Importantes

- n = número de tentativas
- s = número de sucessos
- f = número de falhas

```
 Probabilidade de sucesso
```

$$p = \frac{s}{n} $$

```
Probabilidade de falha
```

$$q = \frac{f}{n} $$

$$q = 1 - p$$

`Chance de sucesso`

$$\text{odds} = \frac{s}{f} $$

$$\text{odds} = \frac{p}{1-p} $$

Função de ligação

A estrutura da função de ligação é a mesma para qualquer modelo generalizado, o que muda é o tipo de função:

O preditor linear η está associado à estrutura determinística do modelo e relacionado à linearização da relação.

$$ \eta = \alpha + \sum\beta_{x_i}$$

A função de ligação é o que relaciona o preditor linear com a esperança do modelo:

$$ \eta = g(E_{(y)}) $$

A função de ligação $g()$ canônica ou padrão para modelos com resposta binária ou proporção é chamada de `logit` ou `logaritmo da chance`[6], definida como:

$$ \eta = \log(\frac{p}{1-p})$$

$$ \log(\frac{p}{1-p}) = \alpha + \sum\beta_{x_i} $$

Sendo $\frac{p}{1-p}$ a **chance** ou **odds** em inglês.

Para reverter o preditor linear da função logit para a escala de observação usa-se a função inversa:

$$ g^{-1} = \text{logit}^{-1} = \frac{e^{\eta}}{(1+ e^{\eta})} $$

Chance e Razão de Chance

O predito pelo modelo na escala do preditor linear do modelo binário com função de ligação ***logit*** está na escala de logaritmo da chance ($\log(\frac{p}{1-p})$). A ***razão de chance*** é uma medida muito popular em outras áreas da ciência, como medicina e mede o quanto uma chance é proporcionalmente diferente de outra, geralmente comparando com um nível controle. Ou seja, qual a proporção de mudança na chance do tratamento em relação a chance do controle. Dado que, em variáveis categóricas os coeficientes do modelo são relacionados às diferenças entre o nível do tratamento e o controle:

$$\text{exp}(\log(\text{odds}_{\text{trat}}) - \log(\text{odds}_{\text{control}})) = \frac{\text{odds}_{\text{trat}}}{\text{odds}_{\text{control}}} $$

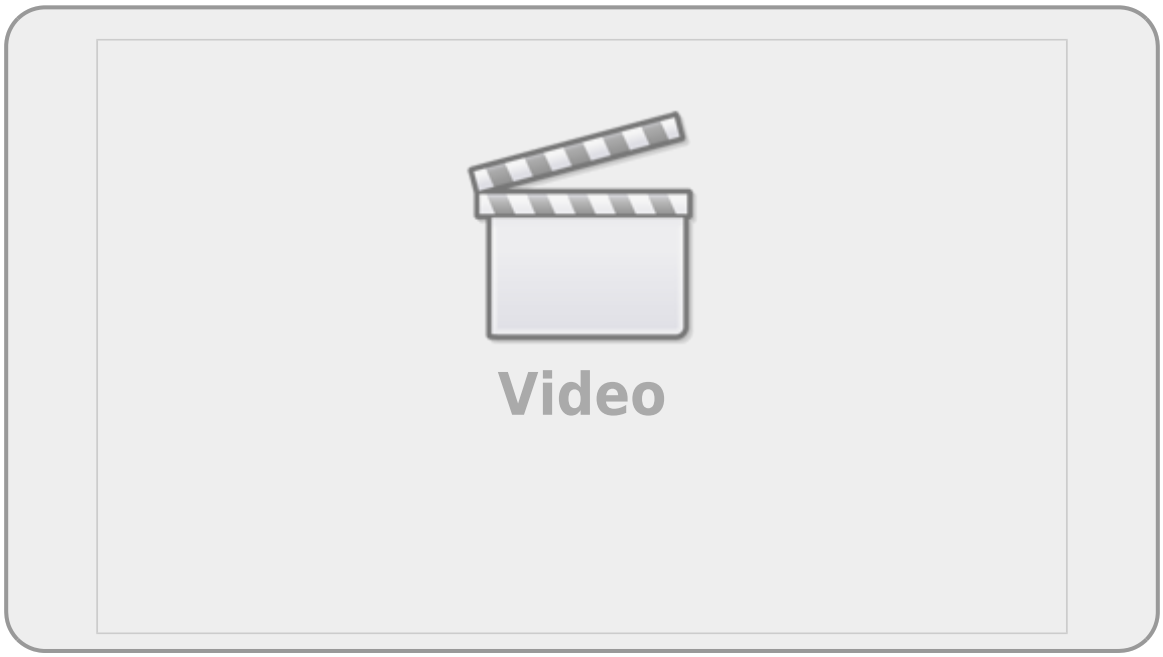
então, exponenciar os coeficientes do modelo binomial com preditora categórica transforma os coeficientes em razão de chance comparado com o nível basal [7].

No caso de variáveis contínuas a **razão de chance** é relacionada à chance de x+1 comparada com x, ou seja, qual a proporção de mudança na chance com o aumento de uma unidade da variável contínua preditora.

Portanto, uma forma de interpretar os coeficientes do modelo binomial é exponenciá-los e interpretá-los como razão de chance, sendo o intercepto a chance do nível basal da variável categórica ou a chance quando a variável contínua é zero.

GLM: resposta binária

Exemplo: pássaro na ilha



O conjunto de dados que vamos usar,

isolation.txt

tem como variável:

Conjunto de dados: `isolation.txt`

- ***incidence***: presença/ausência da espécie de ave (reprodução)
- ***area***: área total da ilha (km^2)
- ***isolation***: distância do continente (km)

Hipótese

O objetivo do estudo que gerou esses dados é saber se a ocorrência da ave está relacionada com o isolamento e tamanho da ilha.

ATIVIDADE

- abra os dados `isolation.txt` no Rcmdr (a separação de campo é tabulação)
- monte o modelo cheio com todas a variáveis preditoras e interações
- simplifique o modelo para o mínimo adequado

Importante:

- lembre-se que a `family` nesse caso é `binomial`
- os modelos com variáveis resposta binárias não tem problema com sobre-dispersão!!!

Interpretação do resultado

O modelo prevê a ocorrência da ave na escala de logaritmo da chance (log odds-ratio). Para os coeficientes estimados pelo modelo o melhor é aplicar a função `exp` e interpretá-los como `razão de chance` entre categorias ou entre x+1 e x. Para interpretar os valores previsto é necessário aplicar a função inversa do `logit`, ou seja, nosso modelo faz previsões na escala de `log(odds-ratio)`, nosso preditor linear η, e precisamos retornar para a escala de observação que é a probabilidade de ocorrência ($\hat{y}$):

$$\hat{y} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}} $$

ATIVIDADE

- calcule o predito pelo modelo na escala de probabilidade de ocorrência para uma ilha de 5.6 Km² e distante 7.2 Km da costa.
- quanto varia a chance de ocorrência se aumentar 1 Km² no tamanho da ilha?
- e se aumentar 1 Km no isolamento?
- faça uma interpretação biológica do modelo selecionado baseado nos seus coeficientes.

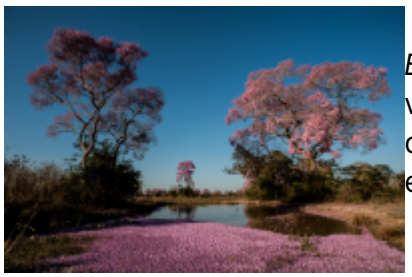
O que preciso entregar

- Preencha as perguntas do quadro abaixo ou pelo link do formulário

GLM binomial: resposta em proporções

Video

Exemplo: floração



Mais um exemplo apresentado no livro do Michael Crawley, *The R Book*. Neste experimento o objetivo foi avaliar a floração de 5 variedades de plantas tratadas com hormônios de crescimento (6 concentrações). Depois de seis semanas as plantas foram classificadas em floridas ou vegetativas.

Conjunto de Dados: `flowering.txt`

- ***flowered***: número de plantas que floresceram
- ***number***: número de plantas acompanhadas
- ***dose***: concentração da dose de hormônio
- ***variety***: variedade da planta (categórica 5 níveis)

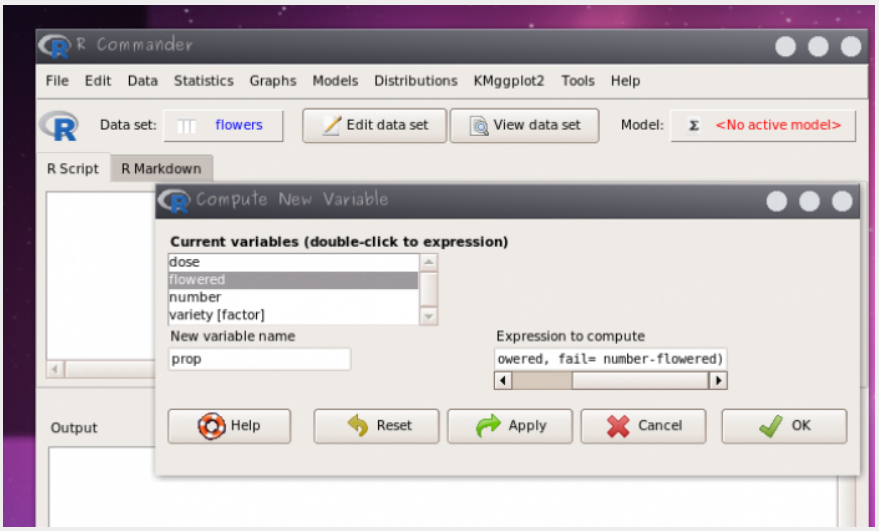
Hipótese

O objetivo do estudo que gerou esses dados é saber se o evento de floração é influenciado pelo dose de hormônio e a variedade da planta.

- baixe o arquivo flowering.txt

- abra os dados no Rcmdr (a separação de campo é tabulação) com o nome `flower`
- crie a variável `prop` pelo menu **Data> Manage variables in active data set> Compute new variable...**, colocando no campo **Expression to compute**:

```
cbind(sucess = flowered, fail = number - flowered)
```



Esse comando acima cria uma nova variável nos dados **flower** chamada **prop**. Essa nova variável tem duas colunas (**sucess e fail**) contendo o número de plantas floridas e o número de plantas que não floresceram, respectivamente.

- use a variável `prop` como resposta (sucessos, falhas)
- monte o modelo cheio com todas a variáveis preditoras e interações
- simplifique o modelo para o mínimo adequado

Use os mesmos passos do modelo anterior no Rcmdr

- lembre-se que a `family` nesse caso é `binomial`
- o procedimento para a sobre-dispersão é o mesmo que no exemplo de contagem, com a diferença que a família aqui é o `quasibinomial`

Interpretação do resultado

Para interpretar os coeficientes use o mesmo procedimento do exercício anterior, que é aplicar a função exponencial (exp) nos coeficientes previstos e interpretar como chance e razão de chance[8].

Para interpretar os valores previsto é necessário aplicar a função inversa do `logit`, ou seja, nosso modelo faz previsões na escala de log(odds-ratio), nosso preditor linear η, e precisamos retornar para a escala de observação que é a probabilidade de florescer ($\hat{y}$):

$$\hat{y} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}} $$

- calcule o predito pelo modelo e os coeficientes na escala original
- interprete o efeito da concentração na floração das variedades a partir dos coeficientes do modelo selecionado

Transformar os coeficientes e valores preditos pelo GLM:

Para transformar o valor predito pelo modelo (log(odds-ratio)) na escala de medida (proporção) é preciso transformar os preditos pelo modelo. Para predizer na escala de medida usamos a função `predict`, como no código abaixo. O predito pelo modelo, está na escala do preditor linear, portanto devemos transformar essa medida com a função inversa da logit, como no código abaixo. Lembre-se de mudar, no código, o "nomedomodelo" pelo nome que usou quando construiu o glm.

```
(preditoLinear <- predict("nomedomodelo"))
(preditoProp <- exp(preditoLinear)/(1+ exp(preditoLinear)))
```

A própria função `predict`, também faz o serviço completo se colocarmos o argumento `type="response"`, como abaixo:

```
predito <- predict("nomedomodelo", type = "response")
predito
```

Mas o **Rcmdr** não poderia ficar sem essa funcionalidade para interpretar os valores do predito pelo modelo na escala de observação, para isso utilize o menu **Models> add observation statistic to data...>** e selecione apenas o **Fitted values**. O Rcmdr adiciona uma coluna nos dados chamada `fitted."nome_do_modelo"`, com os previstos na escala de observação, nesse caso probabilidade.

Gráfico para interpretação dos resultados

Para um gráfico dos resultados use o menu:

Models > Graphs > Predict effect plots...

O que preciso entregar

- Preencha as perguntas do quadro abaixo ou pelo link do formulário

Dispersão e acumulo de zeros

Os modelo GLM poisson e binomial apresentam a variância acoplada à média dos valores, diferentemente dos modelos com distribuição normal onde a média e a variância são independentes. Caso haja uma variação maior ou menor nos dados do que o previsto por essas distribuições, o modelo não consegue dar conta. Essa sobre-dispersão ou sub-dispersão dos dados indica que temos mais ou menos variação do que é predito pelos modelos. Isso pode ser decorrência de vários fontes de erro na definição do modelo, alguns exemplos são:

- o resíduo dos dados pode não ter sido gerado por um processo aleatório poisson ou binomial
- há mais variação do que predito pela ausência de preditoras importantes
- muitos zeros, além do predito pelas distribuições, em decorrência de diferentes processos: um que gera a ausência e outro que gera a variação nas ocorrências de sucesso

Soluções para a sobre-dispersão e acumulo de zeros

A solução mais simples para lidar com a dispersão são os modelo quasipoisson e quasibinomial, que estimam um parâmetro a mais, relacionando a média à variância, o parâmetro de dispersão. Entretanto, os modelos `quasi` dão conta apenas de dispersões moderadas e não indicam qual a fonte dela. Há algumas alternativas ao modelo `quasi` para a dispersão dos dados, alguns deles estão listados abaixo:

- modelo binomial negativo
- modelo de mistura, considerando dois processos distintos
- modelos mistos, considerando a ausência de independência das observações
- modelos com acúmulos de zeros (Zero Inflated Models).

Não é objetivo deste curso mostrar todas essas alternativas, mas caso se deparem com esse problema, muito frequente na área da biológica, saibam que existem alternativas robustas para solucioná-lo.

Variável resposta binária é um caso especial da binomial com apenas uma tentativa, chamado de distribuição de Bernoulli, e não tem problema com sobre-dispersão

Os modelo GLM poisson e binomial apresentam a variância acoplada à média dos valores, diferentemente dos modelos com distribuição normal onde a média e a variância são independentes. Caso haja uma variação maior ou menor nos dados do que o previsto por essas distribuições, o modelo não consegue dar conta. Essa sobre-dispersão ou sub-dispersão dos dados indica que temos mais ou menos variação do que é predito pelos modelos. Isso pode ser decorrência de vários fontes de erro na definição do modelo, alguns exemplos são:

- o resíduo dos dados pode não ter sido gerado por um processo aleatório poisson ou binomial
- há mais variação do que predito pela ausência de preditoras importantes
- muitos zeros, além do predito pelas distribuições, em decorrência de diferentes processos: um que gera a ausência e outro que gera a variação nas ocorrências de sucesso

Soluções para a sobre-dispersão e acumulo de zeros

A solução mais simples para lidar com a dispersão são os modelo quasipoisson e quasibinomial, que estimam um parâmetro a mais, relacionando a média à variância, o parâmetro de dispersão. Entretanto, os modelos `quasi` dão conta apenas de dispersões moderadas e não indicam qual a fonte dela. Há algumas alternativas ao modelo `quasi` para a dispersão dos dados, alguns deles estão listados abaixo:

- modelo binomial negativo
- modelo de mistura, considerando dois processos distintos
- modelos mistos, considerando a ausência de independência das observações
- modelos com acúmulos de zeros (Zero Inflated Models).

Não é objetivo deste curso mostrar todas essas alternativas, mas caso se deparem com esse problema, muito frequente na área da biológica, saibam que existem alternativas robustas para solucioná-lo.

Variável resposta binária é um caso especial da binomial com apenas uma tentativa, chamado de distribuição de Bernoulli, e não tem problema com sobre-dispersão

link para páginas GLMs

- Modelos Lineares Generalizados: binomial
- Modelos Lineares Generalizados: contagem

1)

note que é preciso primeiro calcular o predito na escala do preditor linear e depois transformar, o que não é a mesma coisa que transformar os coeficientes e depois calcular o predito

2)

já não tão moderno assim, já que foi publicado pela primeira vez em 1999

3)

deixe o nome do dado como quine
4)

essa variável tem algumas complicações adicionais e por isso vamos deixá-la de lado
5)

ou coroa, dependendo do que chamamos de sucesso
6)

log odds ou log chance
7)

lembre-se que as categóricas são transformadas em variáveis indicadoras ou dummy e um dos níveis é transportado para o intercepto do modelo, sendo esse o nível basal ou controle
8)

O Rcmdr apresenta os valores dos coeficientes exponenciados após o resumo do modelo na sua construção