Modelos Lineares Generalizados

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 desvios 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)

Modelo Generalizado: 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, o **glm** faz uso de uma função de ligação **log**, para linearizar a relação determinística entre as variáveis, temos portanto o estrutura deterministica do modelo definida por um preditor linear associado 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 da distribuição normal, 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:

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

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

- 1. Abra o arquivo
  species.txt
  no Rcmdr. Note que esse arquivo tem como separador de campo tabulação.
- 2. Monte os modelos lineares clássicos para esse dados, reduzindo ao modelo mínimo adequado a partir do cheio.
- 3. Faça o diagnóstico do resíduos do modelo.
- 4. Utilizando os coeficientes estimados do modelo, faça a predição do número de espécies para um nível baixo ph em uma floresta com biomassa de 7.
- 5. Repita o procedimento 2 a 4 agora com modelo generalizado (glm) e com **family = poisson**.
- 6. Calcule o predito para o modelo, usando os coeficientes do preditor linear do glm.
- 7. Transforme os preditos pelo modelo de volta para a escala de observação.
- 8. Compare os preditos com os observados.

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.

A partir dos gráficos e do modelo selecionado faça um relato (5 linhas) das interpretações biológicos. Esse relato, junto ao resultado e gráficos, deve ser enviado aos professores ao final da atividade.

Gráfico no Rcmdr

Gráfico dos dados

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

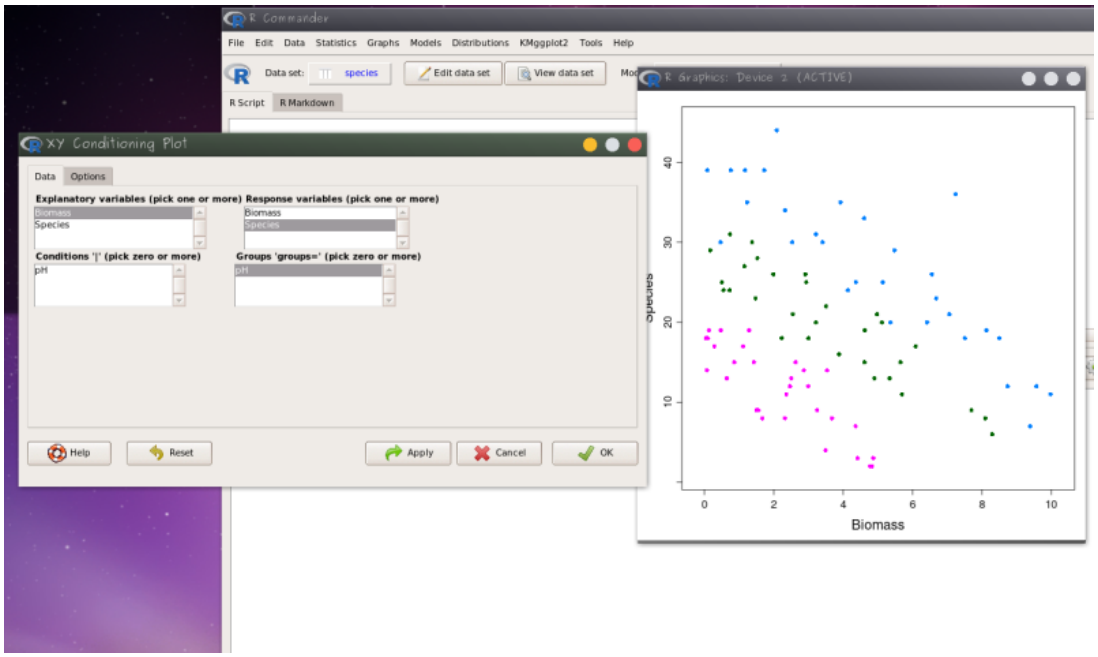
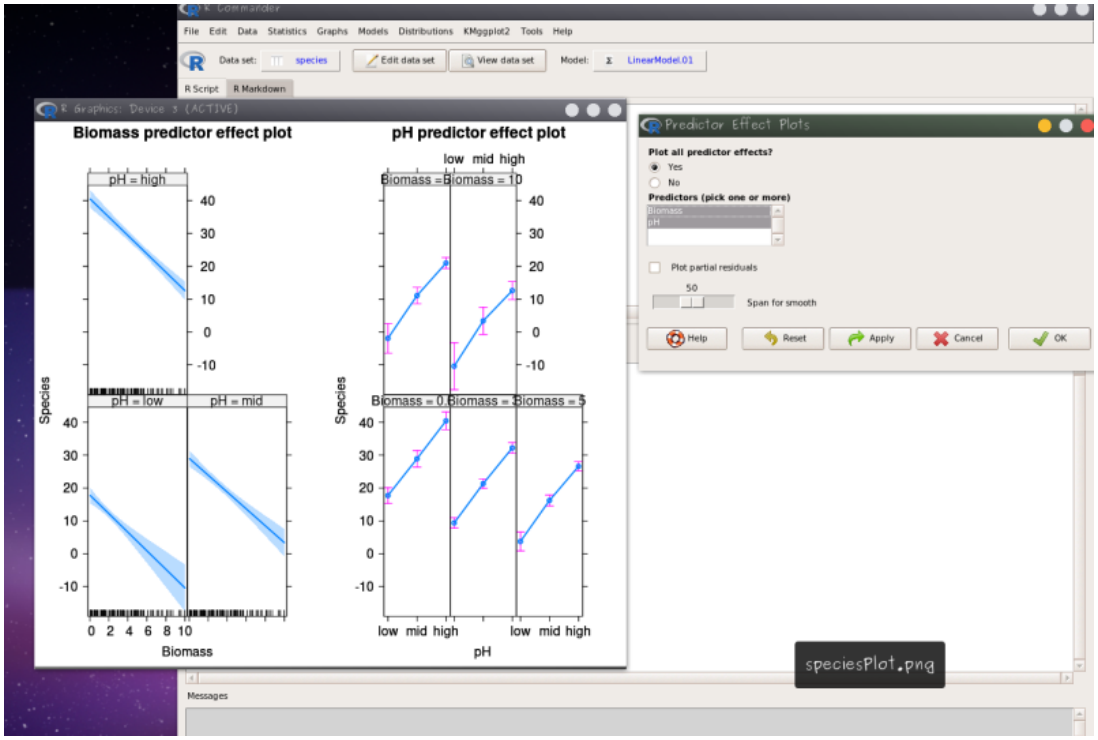


Gráfico dos Modelos

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



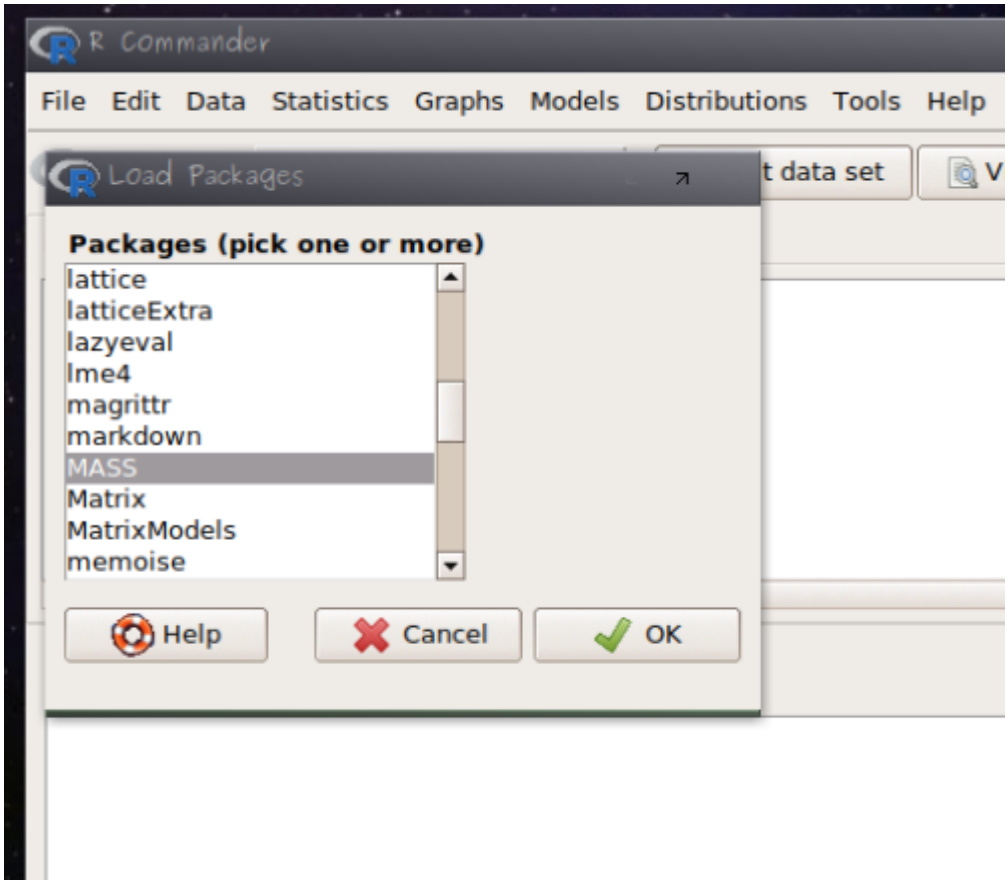
<u>Ordenando uma categórica</u> O padrão do R é ordenar as variáveis categóricas por ordem alfabética. No exemplo seria desejável reordenar a variável categórica ***ph*** em uma categórica ordenada **low>medium>high**. Para reordenar utilize o menu **Data>Manager variable in active data set> Reorder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

Contagem: o que faz um aluno faltar às aulas

Vamos utilizar um exemplo que está presente no livro de W. Venables e B. Ripley, <u>Modern Applied Statistics with S-PLUS</u>[1)], sobre o número de dias ausentes da escola de crianças na Austrália.

Carregando o pacote MASS

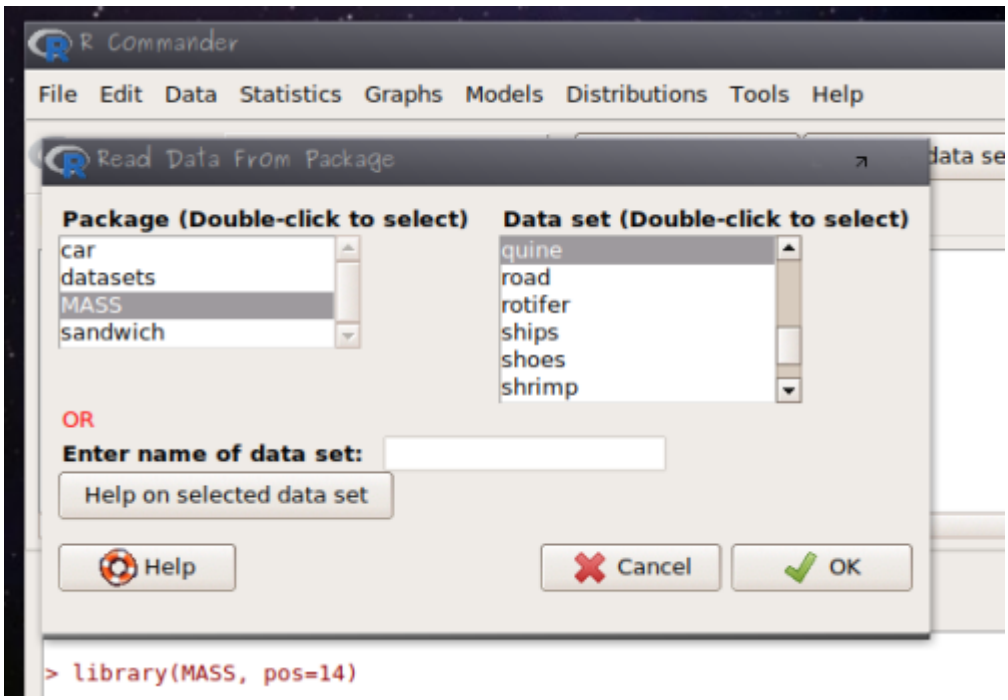
No Rcommader (Rcmdr) vá ao menu **Tools** > **Load package(s)** e selecione o pacote **MASS**



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** [2)]



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)[3)]

Ajustando um Modelo Linear

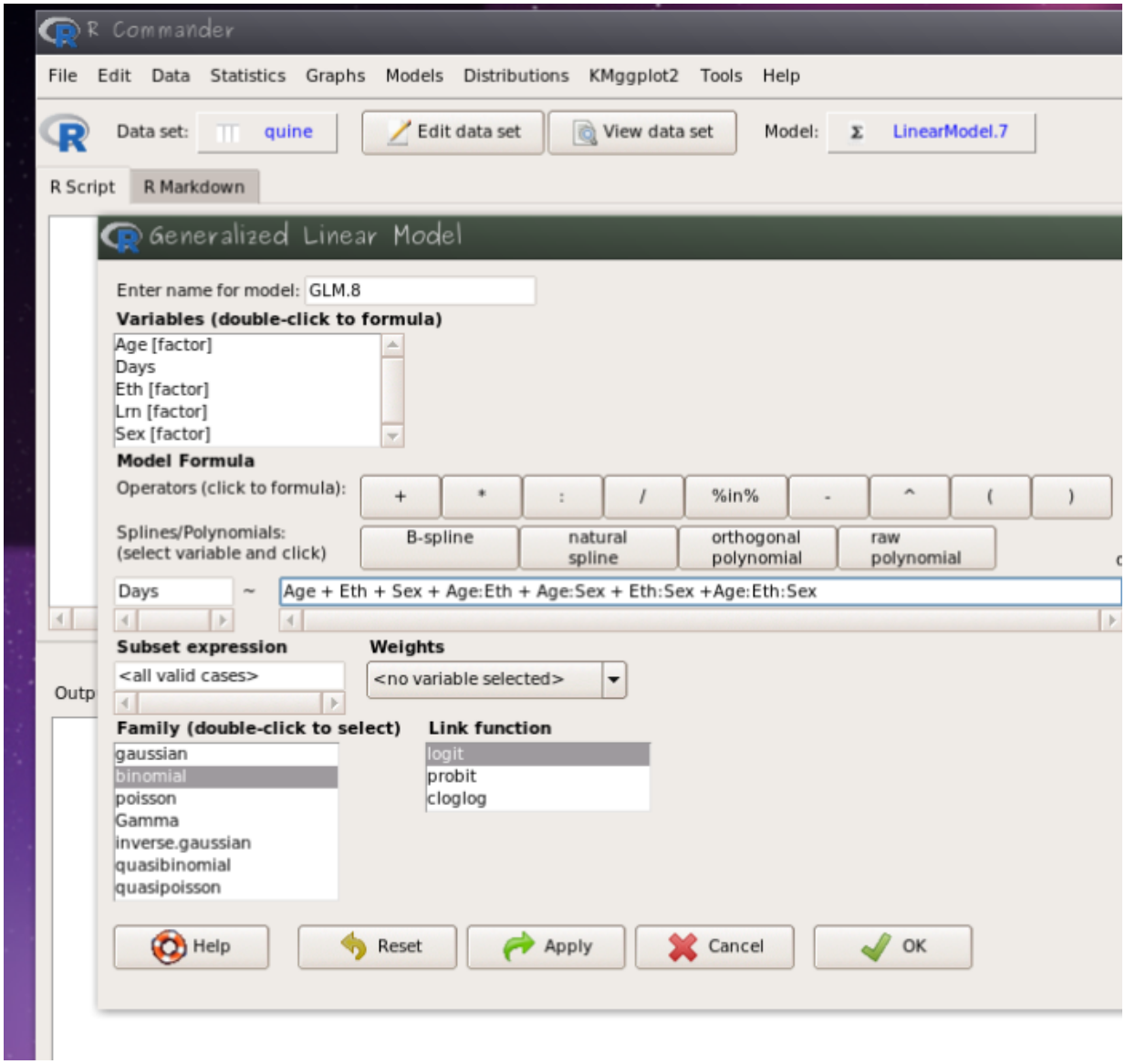
Para nosso exercício vamos deixar de lado a variável **Lrn** por que há dados faltantes nela com relação a outras variáveis. Vamos construir o modelo cheio com todas as outras variáveis (**Eth, Sex, Age**) e todas as possibilidades de interações entre elas. Começamos então com um modelo linear simples.

- abra o menu **Statistics** > **Fit model** > **Linear Model**
- construa um modelo cheio com (**Age, Eth e Sex**) e as suas interações possíveis
- faça a simplificação do modelo para obter o modelo mínimo adequado
- guarde o resultado do modelo selecionado para comparar com o GLM

Ajustando o GLM

Para isso vamos construir o modelo usando a família de erro **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
- faça a simplificação do modelo para obter o modelo 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`

- utilize a familia quasipoisson e
- siga em frente simplificando o modelo para o mínimo adequado
- interprete o modelo selecionado

A partir dos gráficos e do modelo selecionado faça um relato (5 linhas) das interpretações biológicos. Esse relato, junto ao resultado e gráficos, deve ser enviado aos professores ao final da atividade.

Gráfico do Modelo

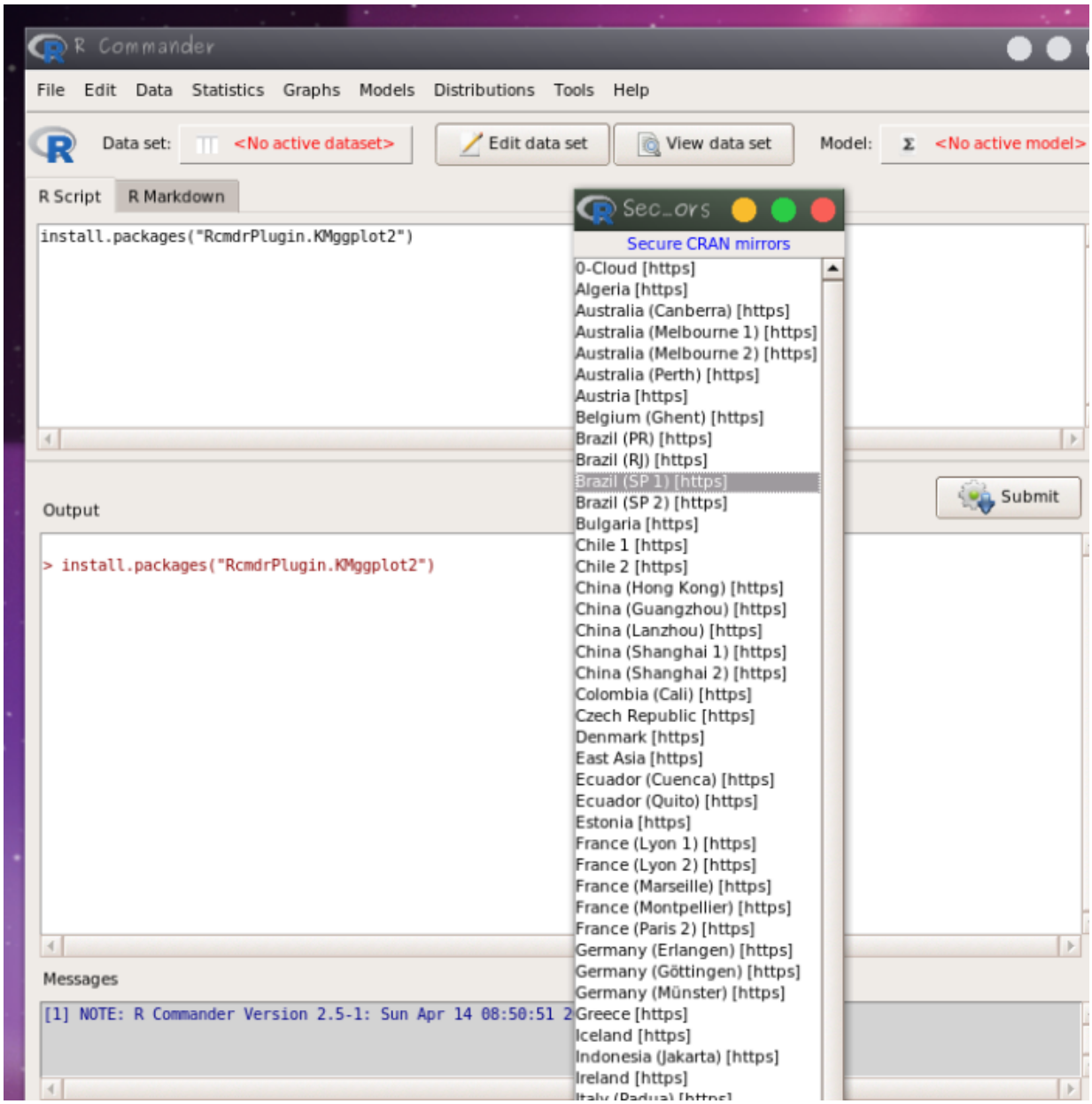
O gráfico do modelo pode ser obtido no Rcmdr da mesma forma indicada no modelo anterior, no menu: **Models>Graphs** selecione **Predict effect plots...** e selecione as variáveis.

Gráfico dos dados

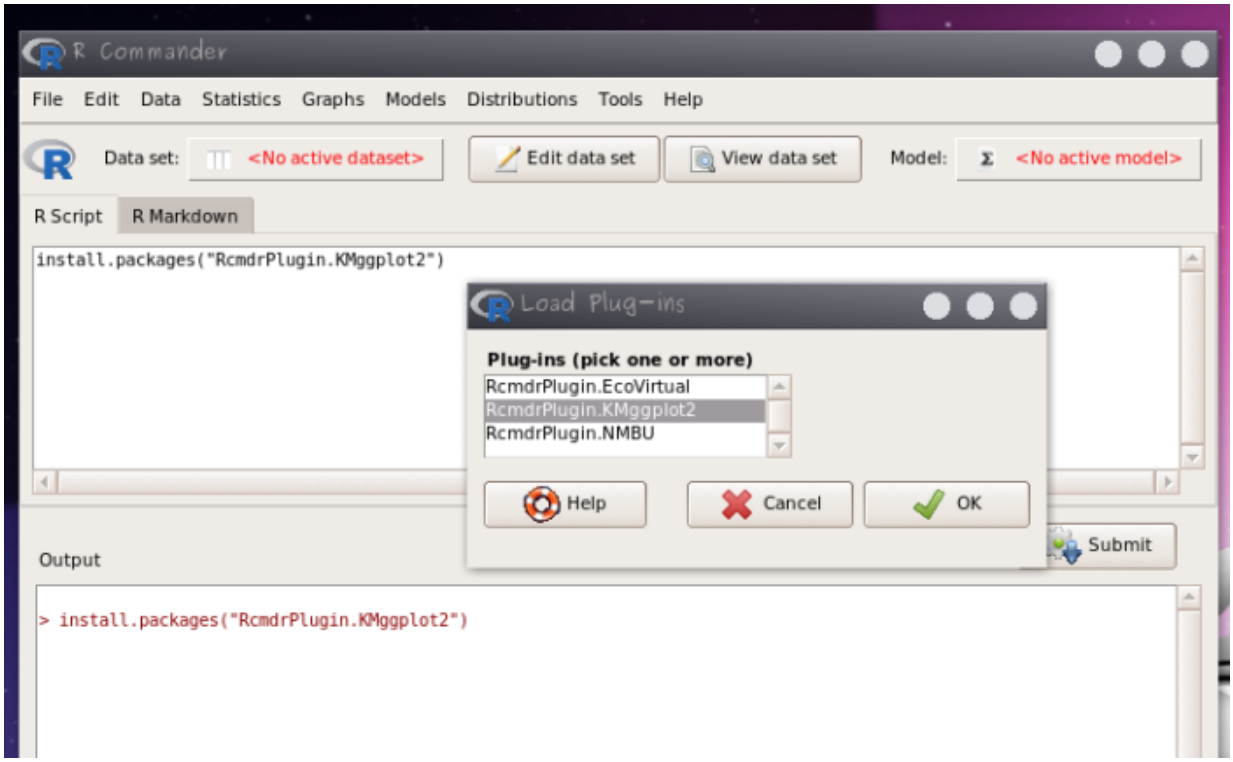
O pacote RcmdrPlugin.KMggplot2 é um plugin para Rcmdr que amplia as funções gráficas da interface. Instale o pacote copiando o comando abaixo no box superior do Rcmdr:

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

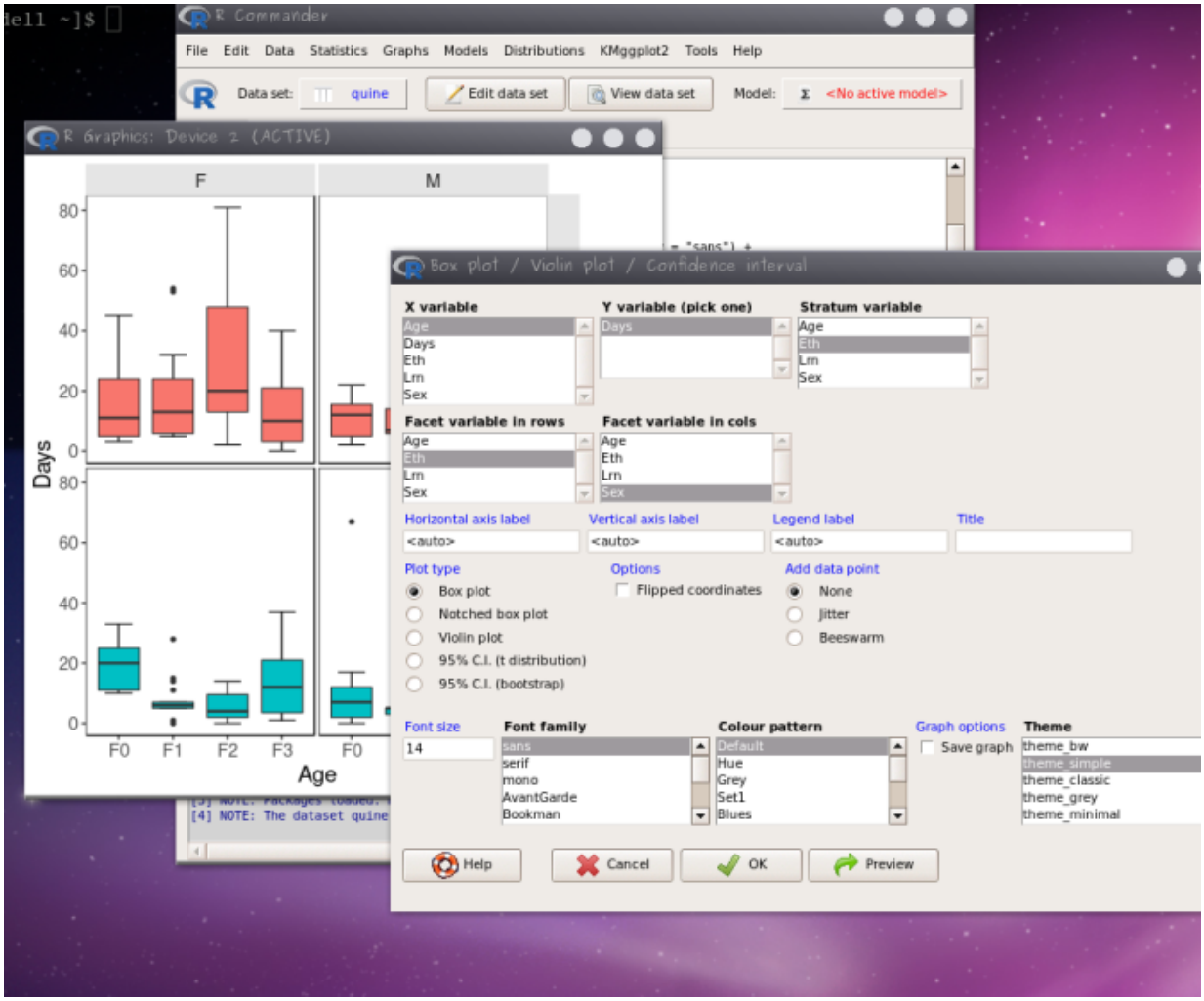
Em seguida, garanta que o cursor do mouse está na linha de comando e clique no botão **Submit**. Na janela que ira se abrir selecione o repositório **Brasil(SP1)**.



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



Ajuste do Modelo Poisson

Sequência de ajuste de modelo de contagem

- faça o modelo cheio usando a familia de ligação **poisson(log)**
- avalie o sobre-dispersão do erro pela razão `Residual deviance` e `degrees of freedom`
- se o valor da razão for muito maior que 1, ajuste o modelo cheio novamente com a família `quasipoisson`
- compare os modelos simplificados com o mais complexo usando `anova`
    - com `poisson` use o argumento `test = "Chisq"`
    - com `quasipoisson` use o argumento `test = "F"`
- retenha o modelo mínimo adequado

GLM Binomial

Os modelos de proporção 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 varia conforme a média. Essas características fazem com que os

resíduos apresentem uma estrutura que aumenta e depois diminuí, e normalmente o máximo de desvios é encontrado nos valores intermediários.

Função de ligação

A estrutura da função de ligação é a mesma para qualquer modelo:

O preditor linear está associado à estrutura determinística do modelo e 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:

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

A função de ligação $g()$ para modelos com resposta binária ou proporção é chamada de `logit` ou `log odds-ratio`, definida como:

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

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

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

Resposta: proporções

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

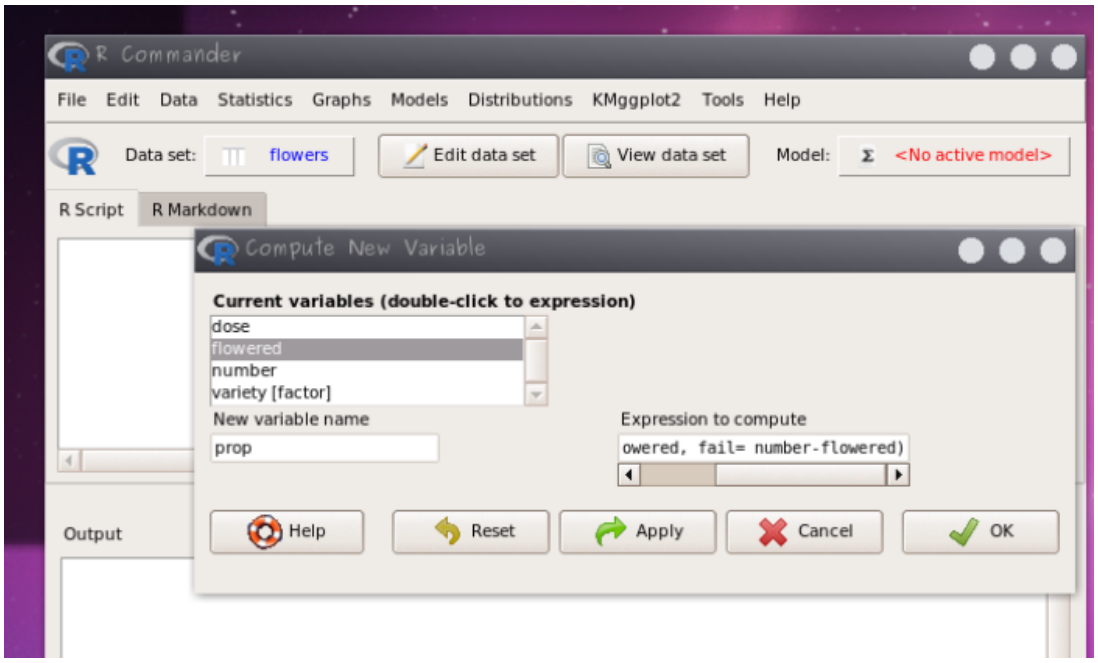
O objetivo do estudo que gerou esses dados é saber se o evento de floração é influenciado pelo dose de hormônio e a variedade da planta.

- baixe o arquivo

flowering.txt

- abra os dados no Rcmdr (a separação de campo é espaço) 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 anterior

Interpretação do resultado

Para interpretar tanto os coeficientes quanto 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

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
```

Gráfico e interpretação dos resultados

Para um gráfico dos resultados use o menu:

Models > Graphs > Predict effect plots...

A partir dos gráficos e do modelo selecionado faça um relato (5 linhas) das interpretações biológicos. Esse relato, junto ao resultado e gráficos, deve ser enviado aos professores ao final da atividade.

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 (reprodução) está relacionada com o isolamento e tamanho da ilha.

- abra os dados `isolation.txt` no Rcmdr (a separação de campo é espaço)
- 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 anterior

Interpretação do resultado

O modelo prevê a ocorrência da ave na escala de logaritmo da chance (log odds-ratio). Para interpretar tanto os coeficientes quanto os valores previsto é necessário aplicar a função inversa do `logit`, como no exercício anterior:

- calcule o predito pelo modelo e os coeficientes na escala original
- interprete o efeito do tamanho e distância na ocorrência da espécie

O que deve entregar?

Para cada exercício feito, deve ser entregue, em um único arquivo:

- o resultado do modelo mínimo adequado
- os coeficientes estimados, na escala de observação
- gráficos que apresentem os resultados principais
- um relato de no máximo 5 linhas, ou em tópicos, da interpretação biológica dos resultados

Sobredispersão e acumulo de zeros

Os modelo GLM poisson e binomial preveem o aumento da variância acompanhando a média dos valores. Caso haja uma variação maior nos dados, o modelo não consegue dar conta, da maneira como os modelos gaussianos normais, que tem um parâmetro específico para a variância ou desvio

padrão. Essa sobre-dispersão dos dados indica que temos mais 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
- há muitos zeros em decorrência de processos diferentes que geram a ausência ou a variação no processo

Soluções para a sobre-dispersão e acumulo de zeros

A solução mais simples para lidar com sobre-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 sobre-dispersões moderadas e não indicam qual a fonte dela. Há algumas alternativas ao modelo `quasi` para a sobre-dispersão dos dados, alguns deles estão listados abaixo:

- modelo binomial negativo
- modelo de mistura
- modelos mistos
- 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.

1)
já não tão moderno assim, já que foi publicado pela primeira vez em 1999
2)
deixe o nome do dado como quine
3)
essa variável tem algumas complicações adicionais e por isso vamos deixá-la de lado