Modelo Lineares Generalizados

Os modelos lineares generalizados (**GLMs**) são usado quando a variância não é constante ou o erro do modelo não tem uma distribuição gaussiana (normal). A natureza da nossa variável resposta indica os desvios que iremos encontrar em relação aos pressupostos dos modelos lineares (**regressões ordinárias**).

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

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

Preditor linear e função de ligação

O preditor linear está baseado na estrutura linear que temos visto nos modelos. Para uma variável preditora:

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

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

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

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

Exemplo de contagem

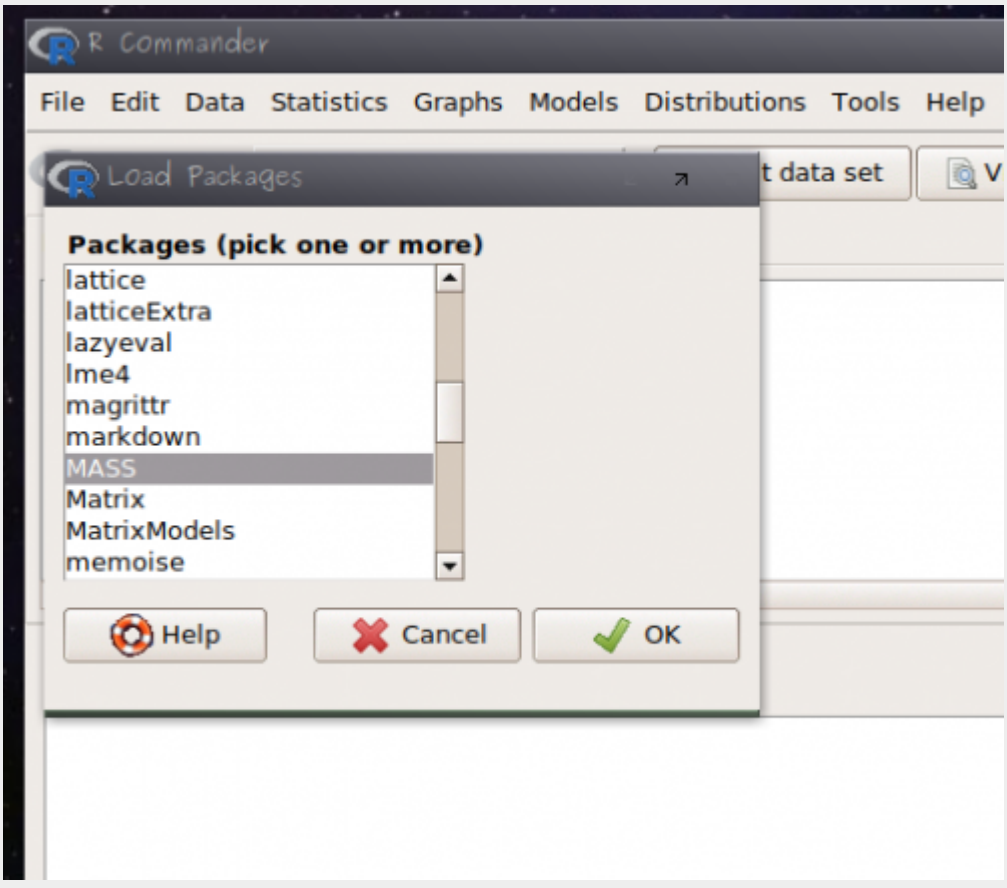
Sequência de ajuste de modelo de contagem

- 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.

Carregando o pacote MASS

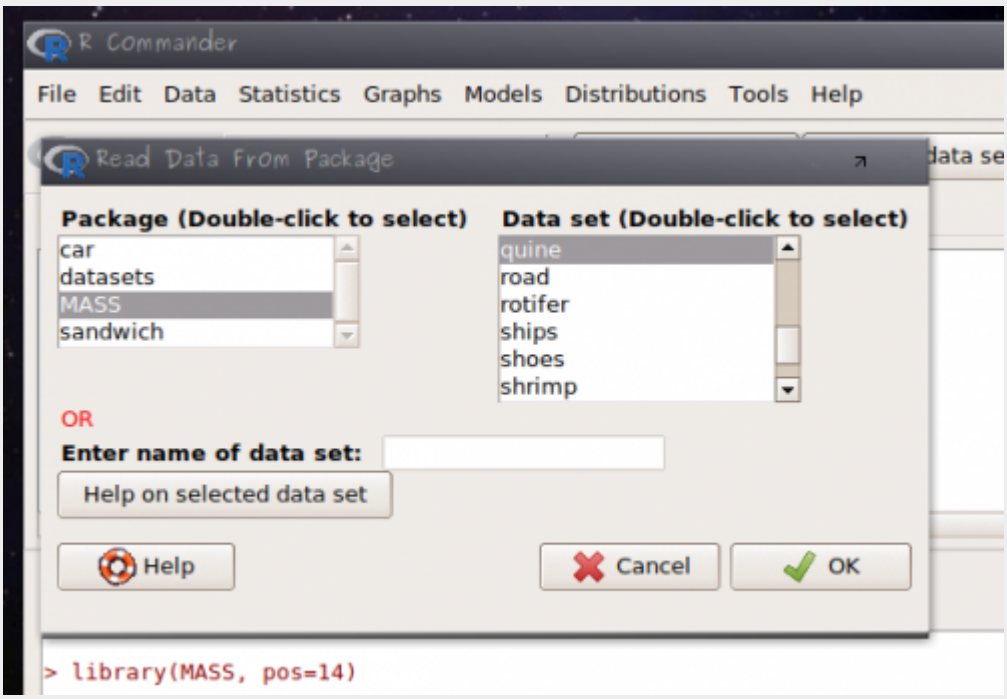
No Rcommader (Rcmdr) vá ao menu **Tools** > **Load package(s)** e selecione o pacote **MASS**



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** [1)]



Entendendo os dados: quine

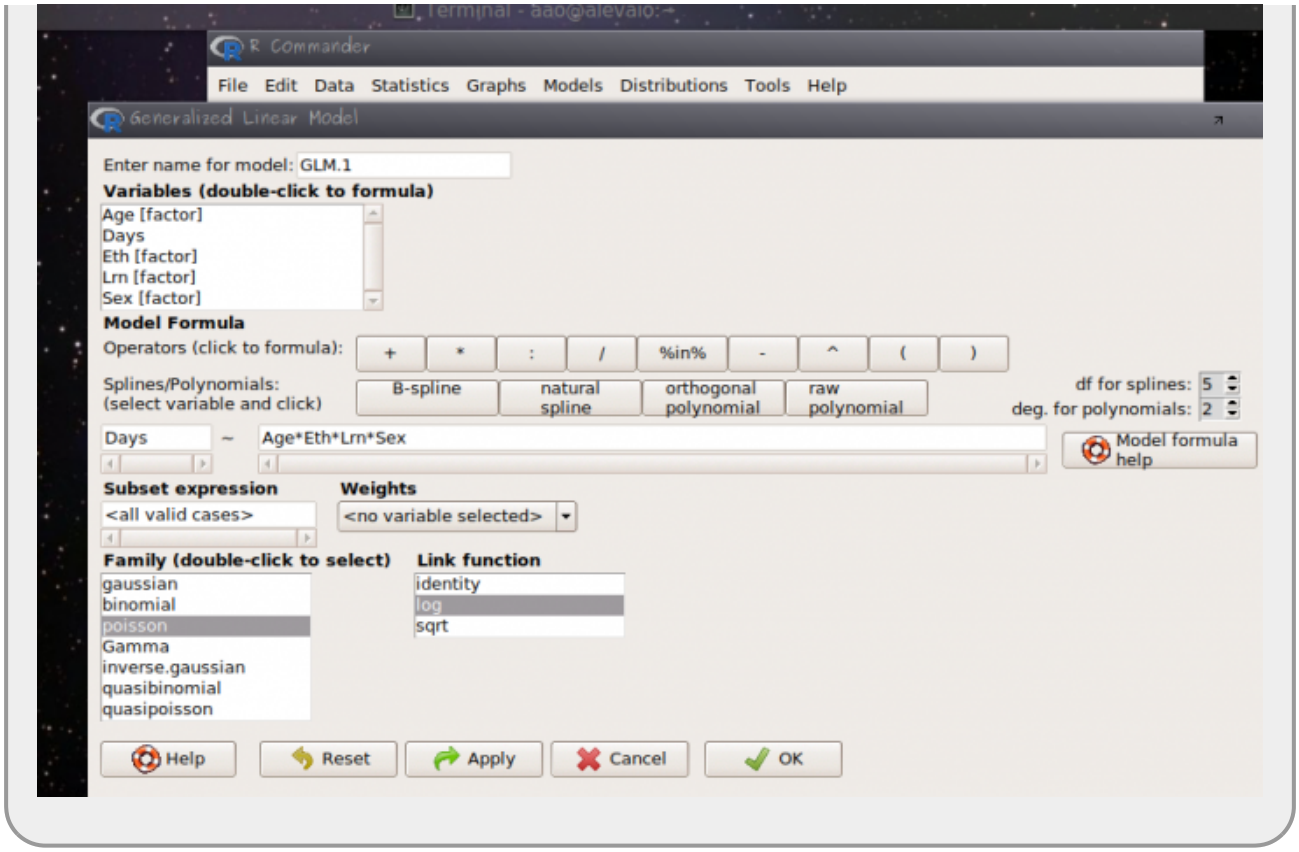
Os dados estão relacionados ao estudo para entender quais variáveis estão relacionados à ausência (falta) do aluno na escola. A observação está relacionada a alunos amostrados aleatoriamente de escolas na Austrália.

- **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)

Ajustando o modelo cheio

Como entendemos que todas as variáveis e interações são possíveis e interpretáveis para a tomada de decisão sobre o permanência do aluno na escola, vamos construir o modelo cheio com todas as possibilidades de interações. 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**
- complete os campos como na figura abaixo



O nosso modelo cheio não conseguiu estimar alguns dos parâmetros. Isso se deveu ao fato de algumas combinações de níveis de fatores não foram encontradas na amostra. Por exemplo não há nenhum:

- aluno de etnia Aborígene do sexo feminino no nível máximo de escolaridade com desempenho fraco!
- aqui o primeiro passo é jogar fora a interação de quarto nível e prosseguir

Avaliando o modelo cheio

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

GLM Binomial

Os modelos de proporção ou de resposta binária (presença/ausência, vivo/morto, sucessos/falhas) são modelados, normalmente, com estrutura do erro binomial. Nesses casos os limite dos valores da variável resposta é bem definido: entre 0 e 1. Essa característica faz com que o erro apresente uma estrutura que aumenta e depois diminuí, e normalmente o máximo de desvios é encontrado nos valores intermediários.

Função de ligação

A função de ligação para modelos com resposta binária ou proporção é chamada de `logit` ou `log odds-ratio`. Pode ser definida como:

$$ p = \log{(\frac{a+bx}{1-(a+bx)})} $$

Exemplo: passaro 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)

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

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 model cheio com todas a variáveis preditoras e interações
- simplifique o modelo para o mínimo adequado

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`:

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

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

Modelo Linear Misto

Para construir modelos onde as observações têm dependência espacial ou temporal, é preciso contemplar a variável com dependência como variável aleatória.

Pacotes para modelos mistos

Os pacotes para trabalhar os modelos mistos no R não são instalados junto com os pacotes básicos como os que contem as funções `lm` e `glm`. Os dois principais pacotes para realizar modelos misto são: `lme4` e `nlme`. Ambos funcionam muito bem para a grande parte do modelo que usamos, mas diferem um pouco quanto à sintaxe. Nesse roteiro iremos usar o `lme4`. Antes de iniciar o roteiro, instale e carregue o pacote com os seguintes comandos:

```
install.packages("lme4")
library(lme4)
```

Para um modelo onde a relação entre a preditora e a resposta não mudam, mas há um efeito relacionado aleatório relacionado à localidade ou o objeto da medida, construímos usamos o LMM da seguinte forma.

- baixe o arquivo
  rikz.txt
- ajuste um modelo com a variável aleatória Beach afeta apenas o intercepto
- compare o modelo com e sem a preditora NAP para tomar a decisão de qual modelo reter
- interprete o resultado do modelo mínimo adequado

```
lmm01r <- lmer(Richness ~ NAP + (1|Beach), data=praia, REML =  FALSE)
lmm00r <- lmer(Richness ~ 1 + (1|Beach), data=praia, REML = FALSE)
anova(lmm00r, lmm01r)
lmm01 <- lmer(Richness ~ NAP + (1|Beach), data=praia, REML =  TRUE)
summary(lm001)
```

Para incluir a variável aleatória Beach também na inclinação é preciso mudar o termo de interação $$(1|Beach) $$ para:

$$(1 + NAP|Beach) $$

- construa o modelo com a variável aleatória NAP afetando o intercepto e a inclinação
- interprete o resultado

1)
deixe o nome do dado como quine