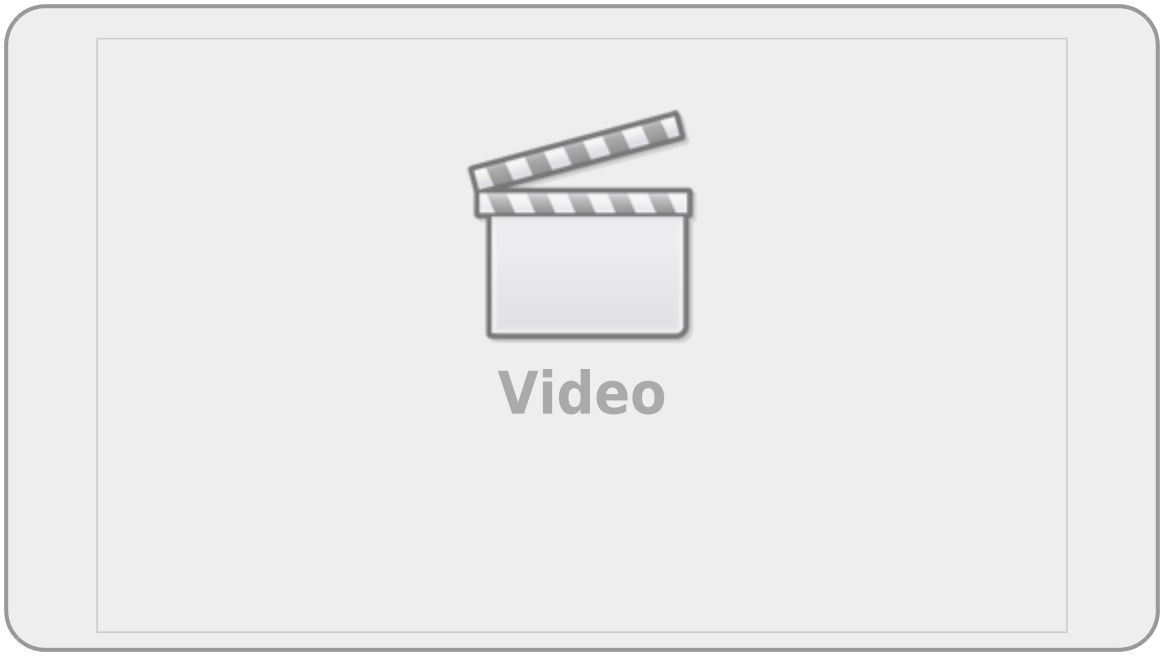
Modelos Generalizados: binomial

GLM: introdução

Essa introdução aos GLM é a mesma do tutorial Modelos Lineares Generalizados, caso já tenha feito, pode passar diretamente para o tópico GLM: binomial



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.

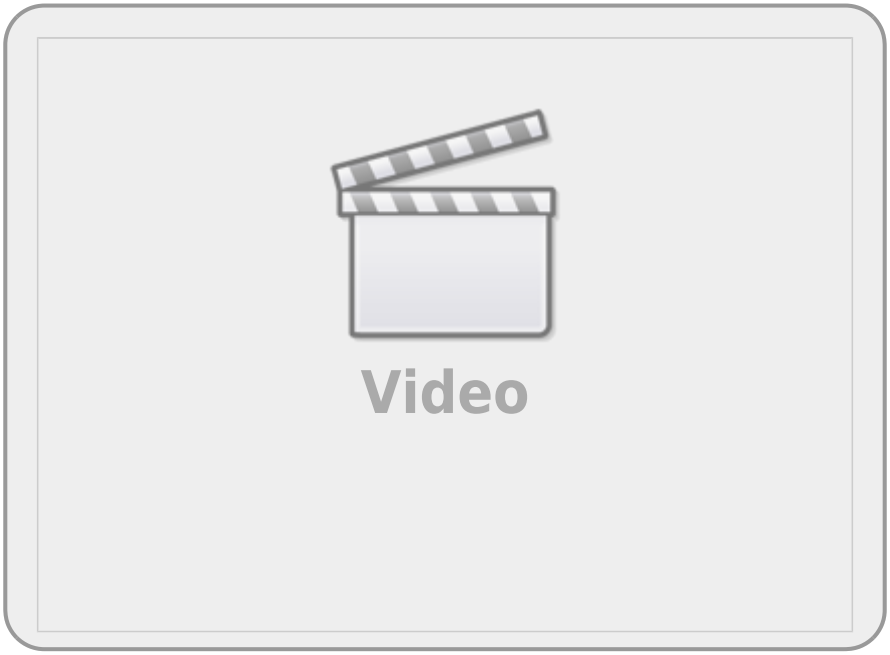
Funções de ligações canônicas

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: binomial

Em muitos fenomenos naturais a variável de interesse tem apenas dois estados possíveis. Essas variáveis de resposta binária (`presença` x `ausência`; `vivo` x `morto`, `germinou` x `não germinou`) tem uma natureza muito distinta das medidas contínuas ou as contagens que vimos anteriormente. Tradicionalmente essa natureza de variável binária foi definida como a probabilidade de sucesso. Neste tópico iremos tratar essas variáveis respostas em **Modelos Lineares Generalizados Binomiais**



Os modelos de proporção (ou probabilidade) de sucessos (`sucessos/tentativas`), proporção simple ou de resposta binária 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 conforme varia a probabilidade de sucesso. 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 apenas 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 [1)].

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 germinação (sucessos) de sementes em um experimento onde temos 20 sementes (número de tentativas). Neste experimento o que tentamos estimar é a probabilidade de germinação (`sucessos`). Elaborando um pouco mais esse experimento podemos incluir uma variável preditora contínua como a úmidade e/ou uma categórica como o tipo de solo, e perguntar como essas variáveis afetam a `probabilidade de sucesso`, no caso a germinação.

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 (Odds)
```

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

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

Note como a **chance** de ocorrência de um evento é a probabilidade de ocorrência deste evento dividida pela probabilidade da não ocorrência do mesmo evento.

A **chance** é muito usada em apostas, quando, por exemplo, dizemos que a chance de um time vencer é de `4:1` [2], ou seja, a probabilidade de vencer é 4x maior do que a de perder. O conceito de chance é muito importante nos modelos binomiais e devemos evitar confundi-lo com probabilidade. Chance e probabilidade são escalas distintas para medir a ocorrência de sucessos.

Função de ligação

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

O preditor linear η está associado à estrutura determinística e relacionado à sua estrutura linear.

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

Note que:

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

É a estrutura determinística do modelo modelo linear, agora não mais relacionado diretamente à escala da variável resposta y e sim a um preditor linear η.

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`[3], 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 ($\log(\frac{p}{1-p})$). Dado que, para 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 [4].

A ***razão de chance*** 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. Pensando em nosso experimento de germinação tendo o `solo arenoso` como nível de referência, a razão de chance do `solo argiloso` seria o quanto a chance de germinar no `argiloso` é proporcionalmente maior/menor que a chance de germinar no solo `arenoso`.

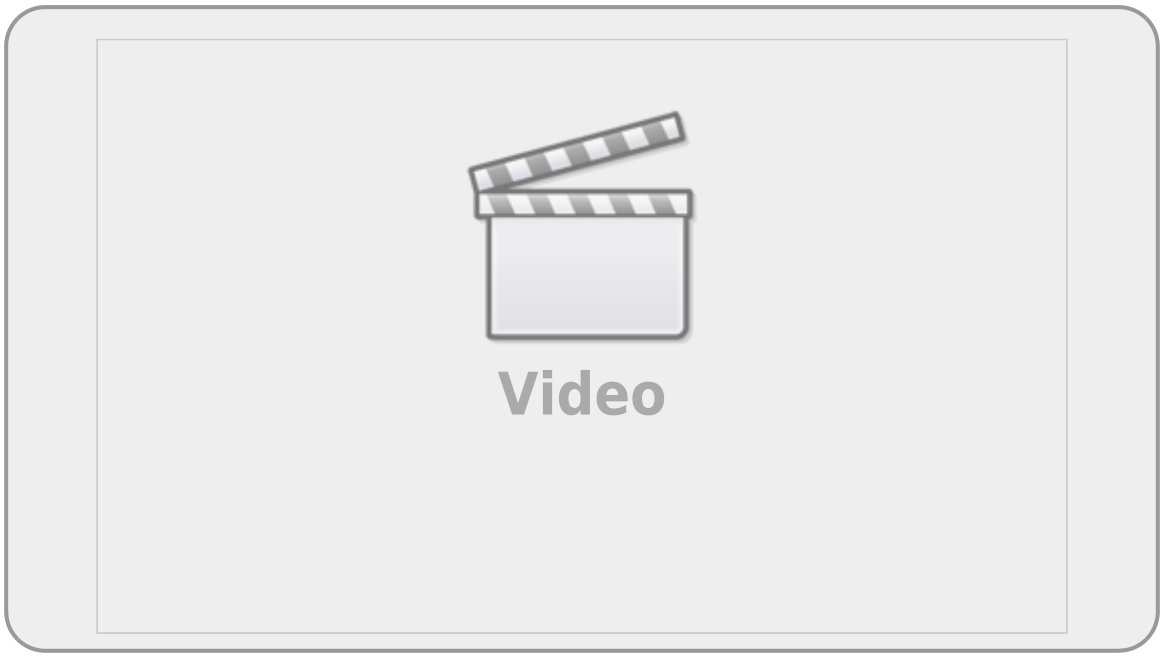
Parece complicado, mas é apenas por falta de intimidade com essas escalas, a razão de chance é uma medida muito popular em outras áreas da ciência, como medicina. Importante lembrar que a `razão de chance` mede o efeito proporcional em relação ao nível de referência.

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 exponenciar e interpretar 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 coletou esses dados foi 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 bernoli (apenas uma tentativa) 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 o 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)

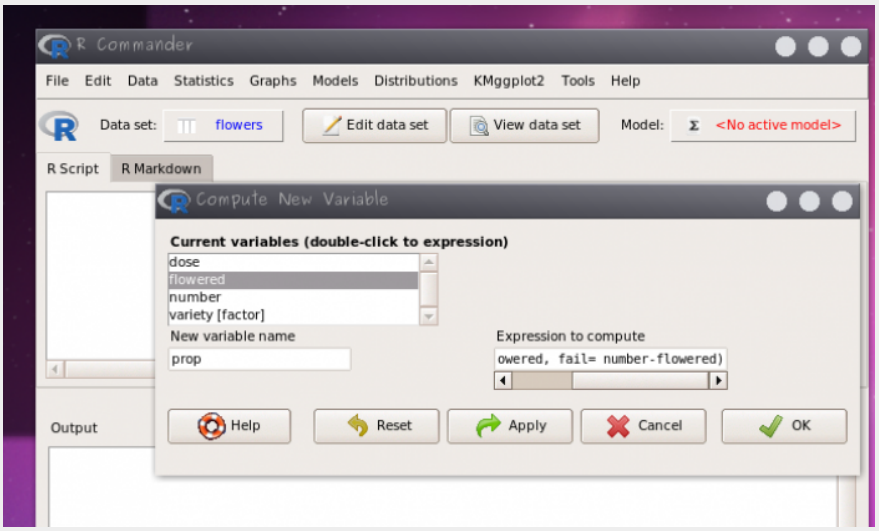
Hipótese

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[5].

Para interpretar os valores previsto[6] pelo modelo é necessário aplicar a função inversa do `logit`. O modelo faz previsões na escala de log(odds-ratio), o preditor linear η, para interpretar é necessário retornar os valores para a escala de observação: **probabilidade de florescer** ($\hat{y}$):

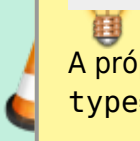
$$\hat{y} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}} $$

- calcule o predito pelo modelo na escala de probabilidade de floração para os valores das variáveis preditoras `dose` e `variety` dos dados originais ;

Transformar os coeficientes e valores preditos pelo GLM:

Para transformar o valor predito pelo modelo (`log(odds-ratio)`) na escala de medida (proporção ou probabilidade) é preciso transformar os preditos pelo modelo. Para gerar as predições do modelo usamos a função `predict`, como no código abaixo. O predito pelo modelo está na escala do preditor linear, portanto é necessário 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
```

O **Rcmdr** não poderia ficar sem essa funcionalidade para interpretar os valores do predito pelo modelo na escala de observação: 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.

- calcule o predito pelo modelo para todas as variedades com doses de hormônio de: 5.5, 12, 25;
- interprete o efeito da concentração na floração das variedades a partir dos coeficientes do modelo selecionado

Gráfico para interpretação dos resultados

Para um gráfico dos resultados use o menu:

Models > Graphs > Predict effect plots...

O que preciso entregar

- Responda as perguntas 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 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, 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.

1)
ou coroa, dependendo do que chamamos de sucesso
2)
ou que está pagando 4 a cada 1 apostado
3)
log odds ou log chance
4)
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
5)
O Rcmdr apresenta os valores dos coeficientes exponenciados após o resumo do modelo na sua construção
6)

esperança