

# Introdução aos modelos lineares generalizados

Prof. Caio Azevedo

# Motivação

- Toda metodologia de análise estatística de dados está baseada em suposições, sobre as quais construímos os resultados inferenciais.
- Se as suposições associadas à uma metodologia, que se pretende utilizar, não são satisfeitas, para o conjunto de dados que se quer analisar, podemos:
  - Verificar/pesquisar o quão robusto é tal metodologia. Se for satisfatoriamente robusta, podemos utilizá-la.
  - Caso não o seja, devemos utilizar uma alternativa, se existir.
  - Caso não exista, devemos desenvolver uma metodologia alternativa.

# Motivação

- Durante muitos anos os modelos de regressão normais lineares homocedásticos (MRNLH) foram utilizados na tentativa de descrever a maioria dos fenômenos aleatórios.
- Com efeito, a grande maioria das metodologias vistas na maioria das disciplinas regulares dos bacharelados em estatística, baseia-se na normalidade (e homocedasticidade) da variável resposta.

# Motivação

- Mesmo quando uma dessas duas suposições não se verificava (verifica), transformações na variável resposta eram (as vezes são) utilizadas a fim de que os dados transformados tivessem (tenham) comportamento aproximadamente normal/homocedástico, para então utilizar as metodologias que pressupõem tais características.
- As principais transformações utilizadas (ainda) são: logaritmica, raiz quadrada, arcoseno e Box-Cox (Paula, 2013).
- Exemplo de transformação. Se  $Y$  é a variável resposta então defini-se  $Y^* = \ln(Y)$  como a nova variável resposta.

# Motivação

- Problemas: perda de interpretabilidade dos parâmetros, aumento do viés e erros-padrão das estimativas, desrepeito da natureza dos dados, impossibilidade, em alguns casos, de se obter normalidade/homocedasticidade.
- Transformações também são utilizadas para se obter uma relação linear entre a resposta e as covariáveis.
- Deve-se incorporar o máximo possível de características do experimento (conjunto de dados).

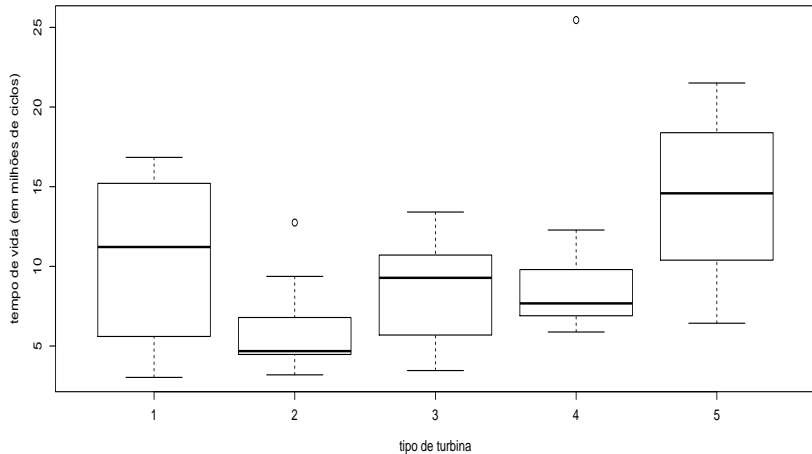
# Exemplo 1: dados da potência de turbina de aviões

- Conjuntos de dados relativos ao desempenho de 5 tipos de turbina de avião (ver Paula, 2013 e Lawless 1982, p. 201).
- Foram considerados dez motores de cada tipo nas análises e observado para cada um o tempo (em unidades de milhões de ciclos) até a perda da velocidade.
- Resposta: tempo até a perda de velocidade.
- Variável explicativa: tipos de turbina.
- Objetivo: comparar os tipos de turbinas.

# Banco de dados

U.E.	potência	tipo de turbina
1	3,03	1
⋮	⋮	⋮
11	3,19	2
⋮	⋮	⋮
21	3,46	3
⋮	⋮	⋮
40	25,46	4
⋮	⋮	⋮
50	21,51	5

# Análise descritiva





# Análise descritiva

Tipo de turbina	Média	DP	Var.	CV(%)	Mínimo	Máximo
1	10,69	4,82	23,23	45,07	3,03	16,84
2	6,05	2,92	8,50	48,18	3,19	12,75
3	8,64	3,29	10,83	38,10	3,46	13,41
4	9,80	5,81	33,71	59,26	5,88	25,46
5	14,71	4,86	23,65	33,07	6,43	21,51

# Observações

- A resposta é positiva, assimétrica e heterocedástica (em função dos tipos de turbina).
- Considerar um modelo de regressão normal linear homocedástico (MRNLH) e conduzir uma ANOVA pode ser problemático, uma vez que tais metodologias requerem normalidade e homocedasticidade da resposta.

# Observações

- Ademais, assumir normalidade para a resposta levará à um modelo que atribui probabilidades positivas de ocorrência a valores negativos para resposta (os quais não podem ser observados).
- Além disso, a linearidade imposta entre a média e a variável explicativa (fator) pode levar à valores negativos para as médias preditas.

## Exemplo 2: Estudo sobre vasoconstrição

- Dados sobre um estudo de vasoconstrição (veja Paula, 2013, Finney, 1978 e Pregibon, 1981).).
- Nesse estudo, foram medidos de 3 pacientes o volume e a razão de ar inspirado, como também a ocorrência ou não de vasoconstrição (contração de vasos sanguíneos) na pele dos dedos da mão. O primeiro paciente contribuiu com 9 observações, o segundo com 8 e o terceiro com 22.
- Em princípio, não seria razoável assumir independência entre as observações. Contudo, por enquanto, assumiremos (metodologias mais apropriadas: modelos mistos, modelos hierárquicos).

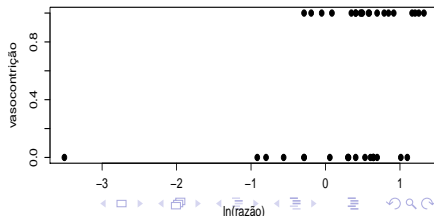
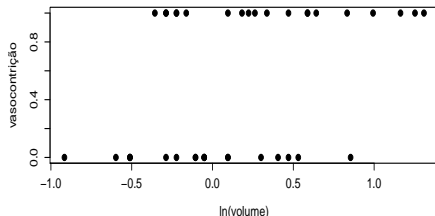
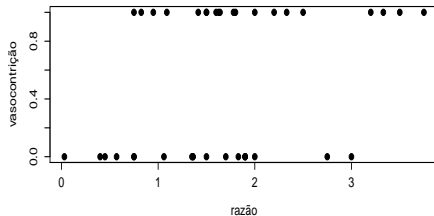
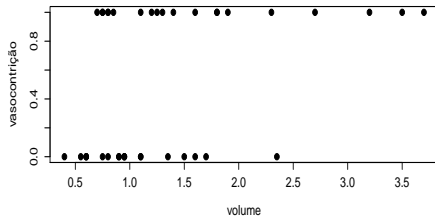
# Banco de dados

Medida	Ocorrência	Volume	Razão
1	1.00	3.70	0.82
2	1.00	3.50	1.09
⋮	⋮	⋮	⋮
7	0.00	0.60	0.75
8	0.00	1.10	1.70
⋮	⋮	⋮	⋮
39	1.00	1.30	1.62

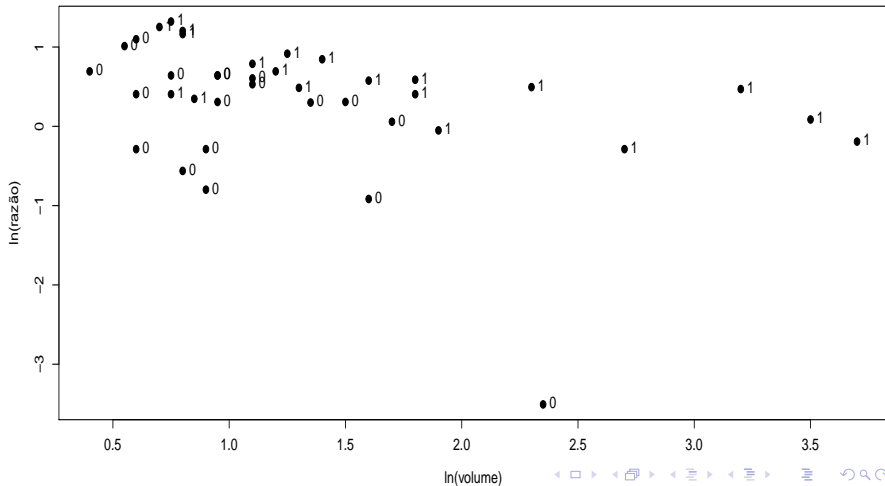
## Cont.

- Resposta: assume valor 1, se ocorreu vasoconstricção no  $i$ -ésimo paciente e 0, caso contrário.
- Variáveis explicativas: volume e razão de ar inspirado.
- Objetivo: verificar se a quantidade de ar (volume e razão) influenciam a ocorrência de vasoconstricção.
- As vezes é mais apropriado trabalhar como o  $\ln$  (logaritmo natural) das variáveis explicativas (para, por exemplo, medir melhor o impacto de cada uma na variável resposta, principalmente se esta não for contínua).

# Gráficos de dispersão individuais



# Gráficos de dispersão: $\ln(\text{razão}) \times \ln(\text{volume})$





## Medidas resumo $\ln(\text{razão})$ e $\ln(\text{volume})$

Medida resumo	$\ln(\text{volume})$		$\ln(\text{razão})$	
	Resposta			
	0	1	0	1
Média	-0,06	0,37	0,05	0,58
Mediana	-0,05	0,30	0,31	0,54
DP	0,45	0,54	1,03	0,46
Var.	0,20	0,29	1,07	0,22
$ \text{CV}(\%) $	723,00	147,00	2223,00	81,00
Min.	-0,92	-0,36	-3,51	-0,29
Max.	0,85	1,31	1,10	1,30

# Observações

- A resposta é binária, logo é inviável não assumir distribuição Bernoulli para a variável resposta.
- O ajuste de um MRNLH levará a inferências inapropriadas.
- Ademais a média da variável resposta jaz no intervalo  $(0,1)$ . Tal característica deve, necessariamente, ser considerada.

## Exemplo 3: tempo de sobrevivências de bactérias

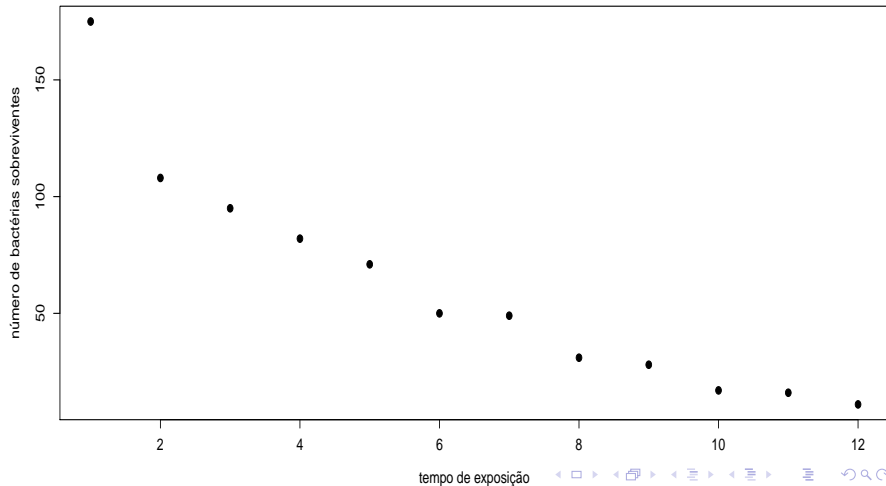
- Os dados correspondem ao número de bactérias sobreviventes em amostras de um produto alimentício segundo o tempo (em minutos) de exposição do produto à uma temperatura de  $300^{\circ}F$ .
- Resposta: número (contagem) de bactérias sobreviventes.
- Variável explicativa: tempo de exposição.
- Nessas amostras de alimentos foram feitas 12 medições, a cada minuto, contabilizando a quantidade de bactérias vivas (do total original) sobreviventes.
- Novamente temos uma situação de medidas repetidas e, assim, as observações podem ter algum tipo de dependência.

# Dados oriundos do experimento

número	175	108	95	82	71	50	49	31	28	17	16	11
tempo	1	2	3	4	5	6	7	8	9	10	11	12

número: número de bactérias sobreviventes; tempo: tempo decorrido em minutos.

# Gráfico de dispersão



# Observações

- A resposta é positiva e discreta. Assim, há possibilidade da resposta apresentar heterocedasticidade em função da variável explicativa.
- Considerar um MRNLH pode ser problemático uma vez que tal metodologia requer normalidade e homocedasticidade da resposta.

# Observações

- Ademais, assumir normalidade para a resposta levará à um modelo que atribui probabilidades positivas de ocorrência a valores os quais não podem ser observados (negativos e não inteiros). Por outro lado, atribui probabilidade zero para eventos que são observados (valores pontuais).
- Além disso, a linearidade imposta entre a média e a variável explicativa (fator) pode levar à valores negativos para as médias preditas.
- Também, a relação entre a resposta e a variável explicativa parecer ser não linear.

## Exemplo 4: comparação do número de acidentes

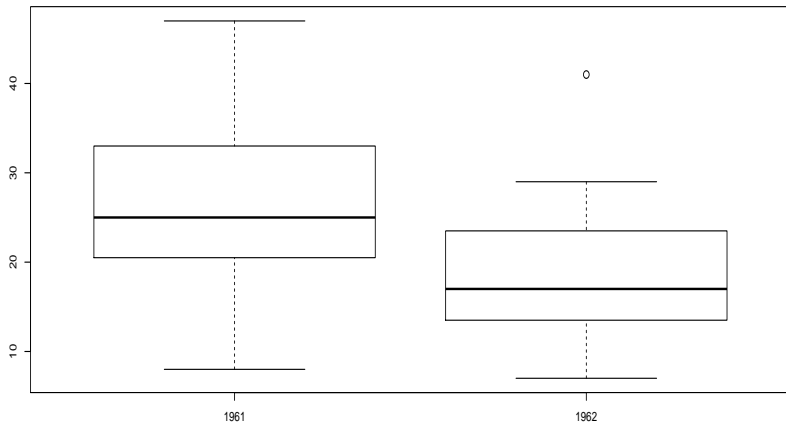
- Descrição: número de acidentes (com algum tipo de trauma para as pessoas envolvidas) em 92 dias (correspondentes) em dois anos distintos (1961 e 1962), medidos em algumas regiões da Suécia.
- Considerou-se apenas 43 dias, correspondendo a dias de 1961 em que não havia limite de velocidade e de 1962 em que havia limites de velocidade (90 ou 100 km/h).



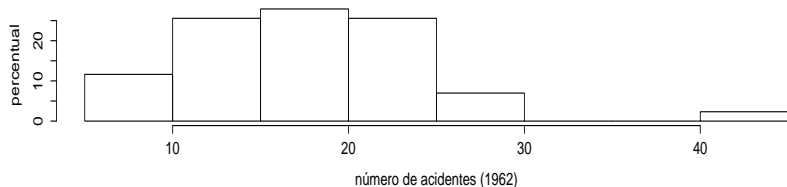
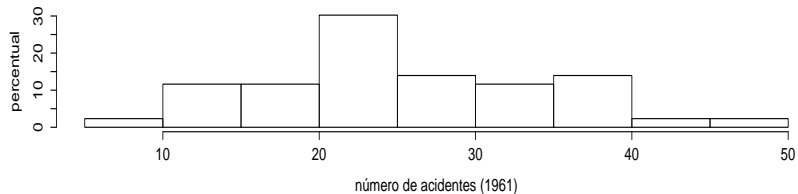
## Exemplo 4: comparação do número de acidentes

- Resposta: número de acidentes.
- Variável explicativa: ano (presença/ausência de limite de velocidade).
- Questão de interesse: a imposição dos limites de velocidade levou à redução do número de acidentes?

# Boxplots do número de acidentes por ano



# Histogramas do número de acidentes por ano



# Medidas Resumo

Ano	Média	Var.	DP	CV(%)	Mín.	Med.	Máx.
1961	26,05	82,66	9,09	34,91	8,00	25,00	47,00
1962	18,05	44,71	6,69	37,05	7,00	17,00	41,00

# Considerações

- Podemos concluir que faz-se necessário a utilização de metodologias (modelos de regressão) mais apropriados para se analisar os conjuntos de dados apresentados anteriormente.
- Nos concentraremos na classe de modelos de regressão lineares generalizados, ou simplesmente modelos lineares generalizados (MLG), proposta por Nelder e Wedderburn (1972).

# Considerações

- Cálculo diferencial integral
- Álgebra de matrizes.
- Métodos computacionais.
- Probabilidade (família exponencial).
- Inferência.