

Modelos Lineares

Modelos Linearizados

Modelos não Lineares

Depto de Zoologia
09 de março de 2021

Modelos Lineares

- Variável dependente tem distribuição normal de resíduos e é função direta do modelo linear das variáveis independentes
- Variáveis independentes participam do modelo por meio de combinação linear dos parâmetros estimados
- As variáveis independentes podem ser contínuas ou categóricas (regressão, análise de variância, ou misto)
- O modelo é ajustado por meio do método de mínimos quadrados (podem existir outros). Os resíduos correspondem à variância não explicada pelo modelo
- Os parâmetros são testados com relação às hipóteses por meio de análise de variância

Exemplo de modelo linear

- Os três primeiros modelos são lineares
- Modelo linear significa *soma linear dos parâmetros*
- As *variáveis* podem ser multiplicadas, divididas, exponenciadas. Os parâmetros a b c etc não.
- Neste caso só o quarto modelo precisa ser linearizado

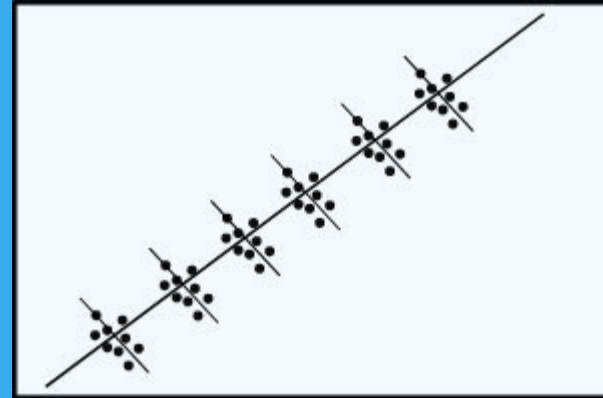
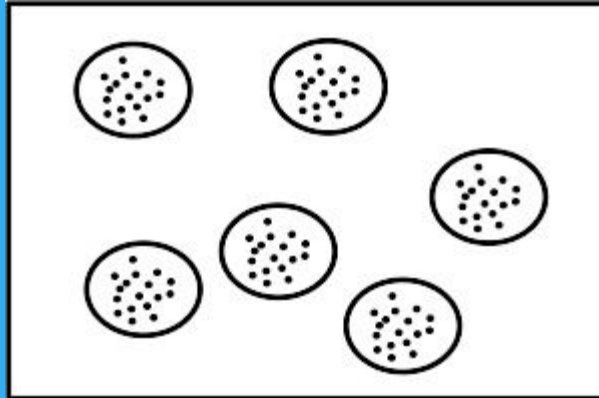
Regression Model	Equation
Simple linear	$Y = a + bX$
Quadratic	$Y = a + bX + bX^2$
Logarithmic	$Y = a + b \log X$
Exponential	$Y = ae^{bx}$ $e = 2.7183$

O que é Modelo Linearizado

- Ao fazer o ajuste observa-se que a variável dependente não se distribue de acordo com o modelo linear. Ou seja a distribuição dos resíduos da variável dependente no modelo não é normal (lembra do quadrante de Anscombe?)
- É aplicada uma função linearizante para reprojeter o modelo na escala da variável dependente, e normalizar os resíduos. A função linearizante é selecionada conforme a distribuição da variável dependente.
- Exemplos de funções linearizantes: poisson, binomial, gaussian(identidade), quasipoisson(log)
- Exemplos de variáveis dependentes que tem essas distribuições e usam este tipo de transformação: contagens de amostras (poisson), percentagens e sobrevivencia (binomial)

Modelos lineares mistos

- A variável independente foi amostrada com repetições não totalmente independentes (várias réplicas são relacionadas, por exemplo amostra de frutos de uma espécie inclui repetições da mesma árvore)
- Também chamado de modelo hierárquico



$$y = X\beta + Zu + \epsilon$$

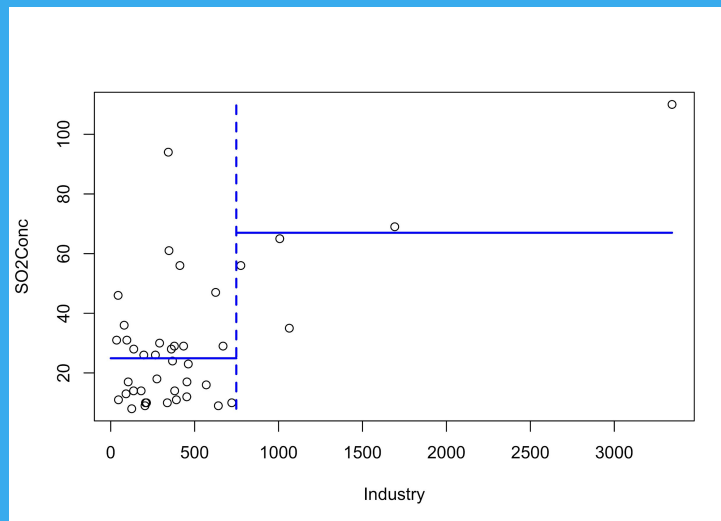
Modelos não lineares: GAM

- GAM - modelos aditivos generalizados. Não há premissa do modelo ser uma combinação linear ou linearizada dos parâmetros
- É estimado o ajuste usando métodos não paramétricos
- O modelo tem a fórmula geral. A variável Y tem distribuição da família exponencial e é transformada pela função g (pode ser log ou outra). As funções f podem ser bastante diversas e estimadas não parametricamente

$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m).$$

Modelos não lineares: CART

- CART - Classification and Regression Trees
- Ao invés de ajustar o modelo contínuo ou discreto, é feita uma partição dos dados e representada na forma de uma árvore. Os pontos de separação na árvore representam os valores da variável dependente que separam melhor a amostra total em subgrupos.



Funções no R para estimar modelos

- modelo linear: `lm`
- Modelo linear generalizado: `glm`
- Anova: `aov`
- Modelo não linear: `gam`
- Modelos de regressão ou classificação em árvore: `tree`
- **IMPORTANTE:** todas estas funções produzem objetos do tipo `list` em R contendo numerosos outros objetos como vetores que especificam os parâmetros estimados, os resíduos, os valores previstos, entre outros. Estes objetos podem ser inspecionados usando funções como `summary`, `anova`, `plot`, etc

Exemplo: alguns objetos gerados pela função `lm`

- `coefficients`: vetor com os parâmetros estimados
- `residuals`: vetor com o valor dos resíduos Y para cada X
- `fitted.values`: vetor com os valores estimados de Y para X
- `df.residual`: graus de liberdade dos resíduos
- `terms, model`: dados sobre o modelo usado
- ESTES objetos podem ser visualizados usando funções como `summary`, `View`, `plot`, `anova`, ou então usados como dados de entrada para análises posteriores

Como Selecionar Modelos

- Podemos testar várias combinações de variáveis independentes para ajustar o modelo.
- Como selecionar o melhor modelo?
- O método mais usado hoje é o AIC (Akaike Information Criteria)
- Cada modelo tem seu índice de Akaike. Selecionamos o modelo com o menor índice ou o melhor ajuste

Como interpretar os resultados dos modelos

- A qualidade do ajuste pode ser medida pelos resíduos (mínimos quadrados)
- A intensidade da associação entre variáveis independentes e dependentes pode ser avaliada pelo valor de cada parâmetro estimado e sua variável correspondente
- A significância do modelo (se é diferente de zero) pode ser avaliada pelo teste de significância de cada parâmetro (hipótese nula é de que o parâmetro é igual a zero, ou seja não há efeito).

Dados e exemplos usados na aula de hoje

- Artigo Abreu et al 2020 (J Appl Ecol) sobre biodiversidade da Serra do Facão com os dados originais e script em R depositado no Dryad. Modelo bayesiano hierárquico.
- Artigo Hoover et al 2020 (Condor) sobre modelos mistos para estimar determinantes de parametros reprodutivos de aves.
- Tutorial sobre GLM GAM E CART do Prof Pat Bartlein da Univ of Oregon:
<https://pjbartlein.github.io/GeogDataAnalysis/lec15.html>