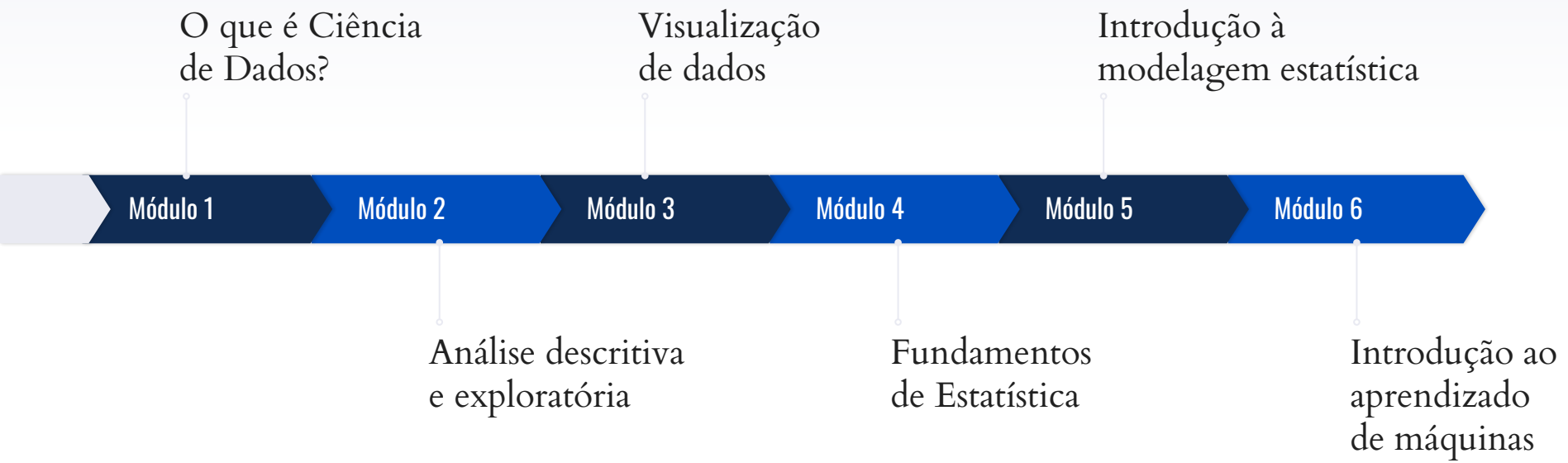
INTRODUÇÃO À CIÊNCIA DE DADOS

Sua jornada começa aqui

omegadatascience.com.br

Estrutura do curso



MÓDULO 5
INTRODUÇÃO À MODELAGEM ESTATÍSTICA

Objetivo

Motivar as aplicações dos modelos estatísticos, o que são e para que servem.

Componentes

O modelo estatístico e suas partes, como funcionam e como estão conectadas as classes de modelos.

Estimação

Como um modelo é ajustado aos dados e o que isso envolve.

Usando o modelo

De interpretações à predições, como usar modelos para tomar decisões e desenvolver produtos com dados.

MÓDULO 5

INTRODUÇÃO À MODELAGEM ESTATÍSTICA

1. Modelos estatísticos e seus componentes.
2. Objetivos do uso de um modelo.
3. Especificação e ajuste de modelos.
4. Interpretando os resultados e a tomada de decisões.
5. Diferentes tipos de modelos e como escolher o adequado.
6. Projeto V: Especificando e ajustando um modelo.

Vamos começar!

Telegram: t.me/omegadatascience

Instagram: @omegadatascience

Twitter: @omegadatascienc

YouTube: /OmegaDataScience

Ômega Data Science

PLATAFORMA DE CURSOS ONLINE

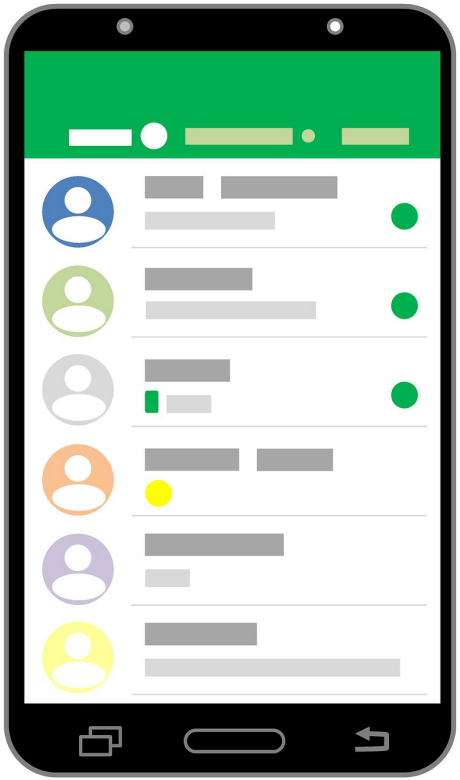
omegadatascience.com.br

Ômega Fly

- **Área de negócio:** Venda de passagens áreas para empresas.
- **Objetivos:**
  - Aumentar o portfólio de produtos.
  - Aumentar a satisfação e lealdade dos clientes.
- **Problema:** Voo atrasado.
- **Oportunidade:** Criar um produto que ajude os colaboradores a planejar melhor suas viagens.

Produto

- **Smart travel**
  - Cliente reporta cidade de origem e de destino e qual o horário gostaria de chegar.
  - Sistema busca qual é o voo mais indicado de modo antes do horário desejado.
- **Se o voo atrasar a Ômega Fly cobre:**
  - Custos adicionais de remarcação.
  - Acomodação + estadia no local de destino.



Resumo executivo

- **Suposições**
  - Venderemos 10000 tickets no ano.
  - Condições do passado serão mantidas aproximadamente constantes.
  - Atraso é independente entre os clientes.
  - Política trivial: 60 minutos antes para todos.

- **Resultados**
  - Qual o percentual de tickets teremos que ressarcir?
  - 6,94% (6,47% | 7,46%)
  - Qual o custo esperado total do produto?
  - 312.383,70 (291.150,00 | 335.700,00)
  - Quanto devemos acrescer em cada passagem para em média cobrir os custos do produto?
  - 31,23 (29,11 | 33,57).
  - Quanto devemos cobrar para não ter prejuízo em 95% das vezes?
  - R$ 33.21.

Por que um voo atrasa?

- Aeroporto de origem.
- Operadora do voo.
- Horário previsto para a saída.
- Dia da semana ou do mês.
- Mês do ano.

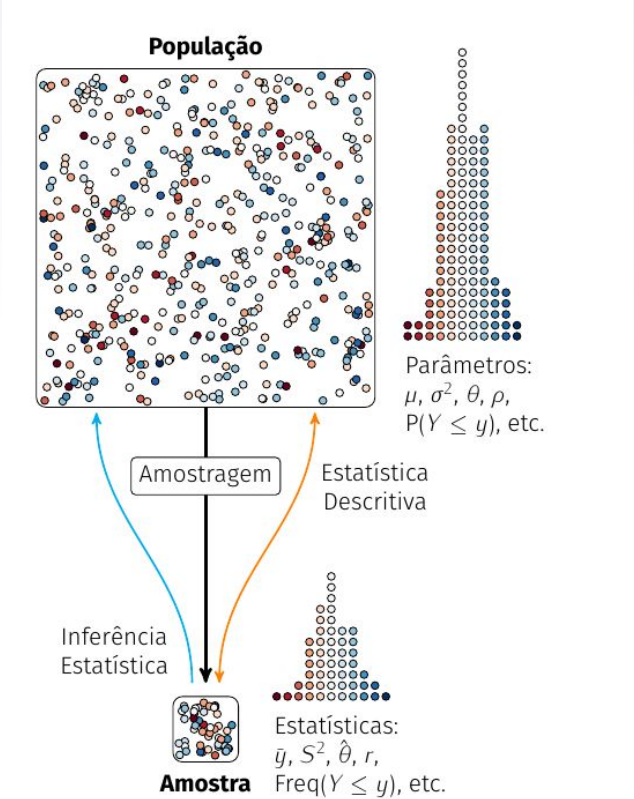
Hora do código!

Muitas possibilidades!

- A realidade é realmente complexa.
- Precisamos de uma forma de simplificar.
- Porém, mantendo as características importantes!
- Solução: Modelagem estatística!

Inferência estatística

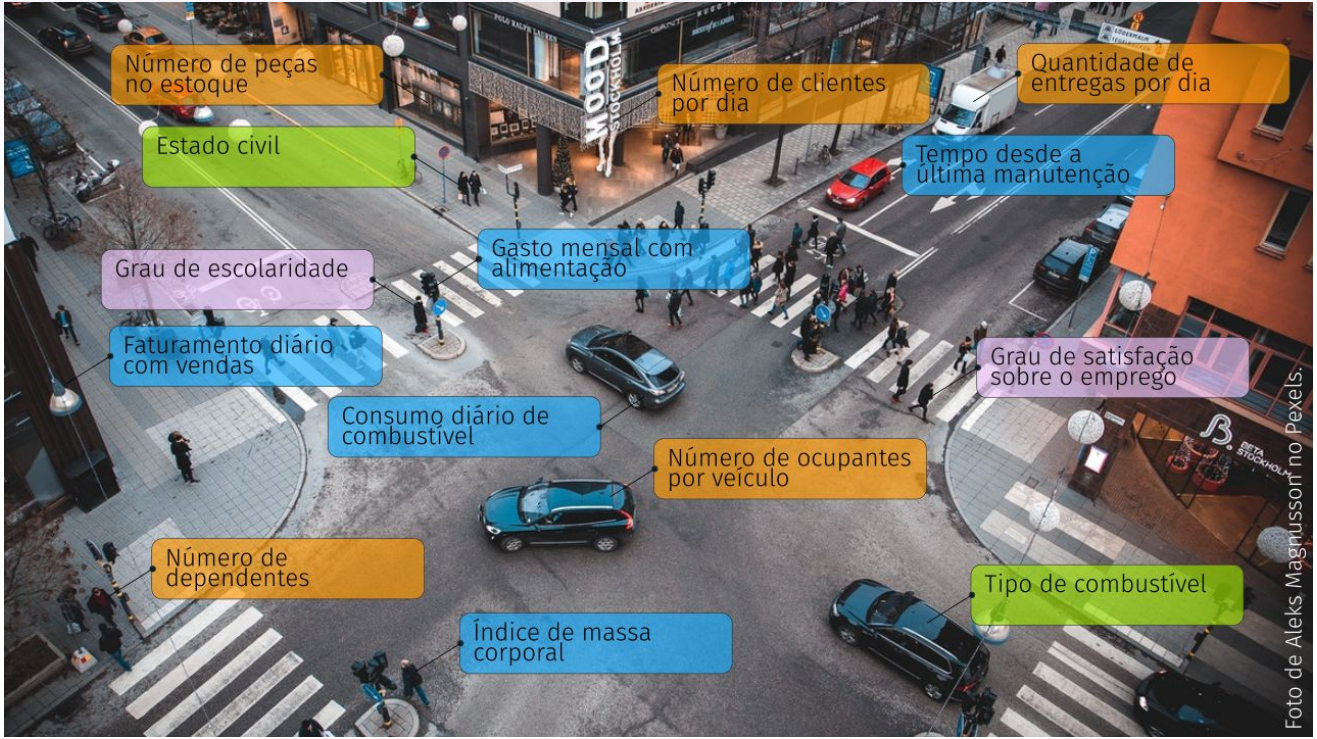
- Modelo -> comportamento da natureza.
- Parâmetros do modelo -> parâmetros populacionais de interesse.
- Qual modelo melhor descreve os dados?
- Assumimos um modelo -> parâmetros são desconhecidos.
- Baseado na amostra -> encontrar os parâmetros compatíveis com a amostra.
- Descrever a incerteza -> distribuição amostral.



O que é um modelo?

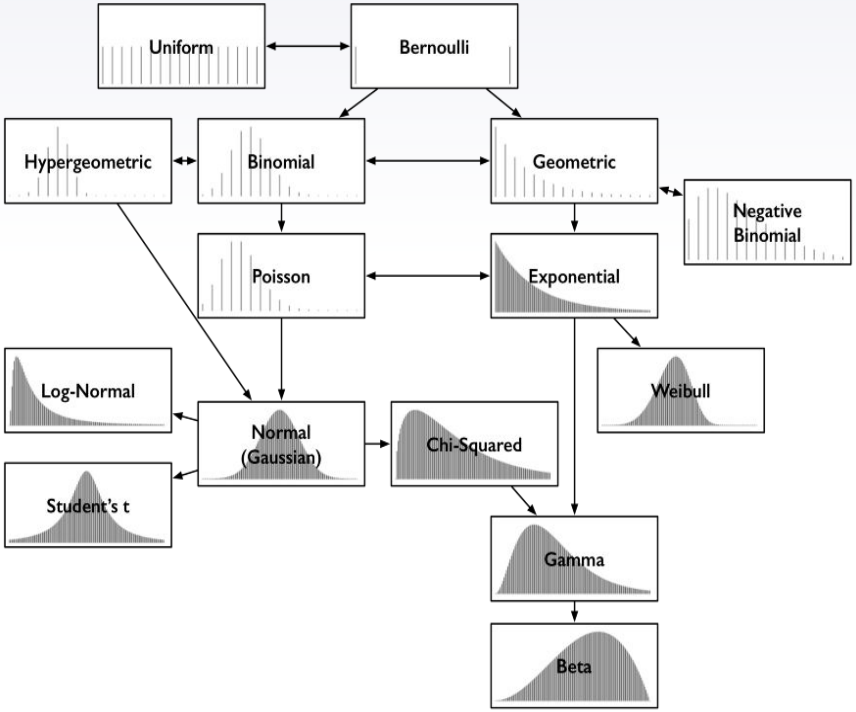
- É uma função parametrizada!!??
- Distribuição de probabilidade!!??
- Simplificação da realidade.
- Objetivo: Manter os aspectos relevantes da realidade.
- Simples de interpretar e generalizar.

Fenômenos aleatórios



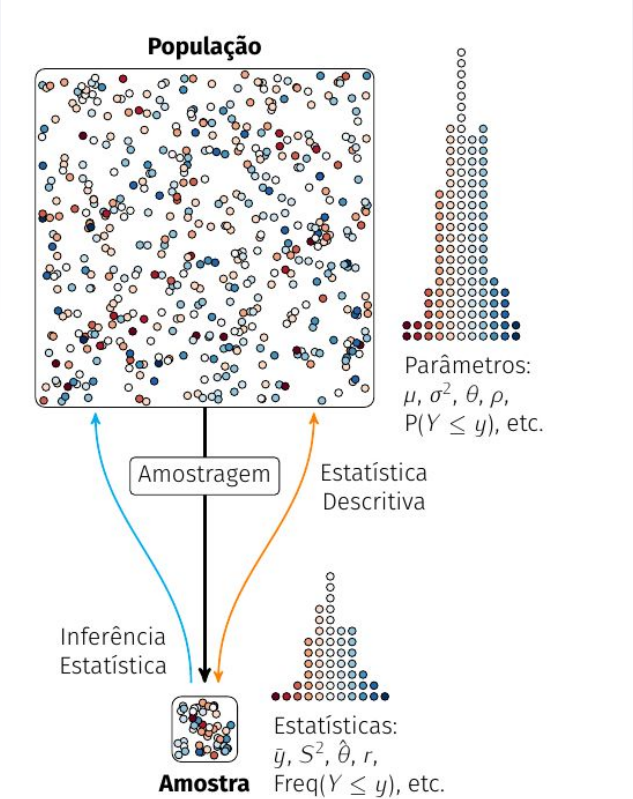
Modelos estatísticos

- Modelos fornecem **fórmulas gerais** para tratar situações similares.
- Permitem determinar **probabilidades** de eventos.
- Servem para **estimar** parâmetros.
- Permitem realizar **testes de hipóteses.**
- Permitem **compreender/acomodar** o efeito de covariáveis sobre o comportamento da variável de interesse.
- São empregados para fazer **previsão**.
- Fazem **suposições** que podem ou não ser razoáveis.
- Podem ser **flexibilizados** para acomodar diferentes situações.



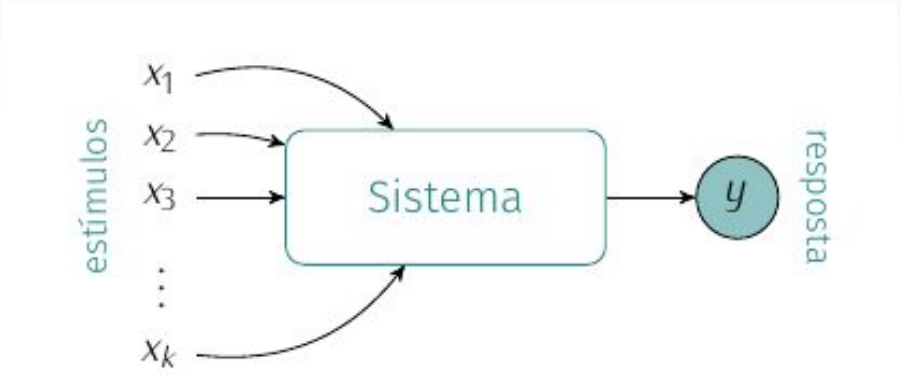
Inferência estatística

- População -> distribuição de probabilidade.
- Intuição -> Como a variável aleatória se comporta na população.
- Variável de interesse -> variável aleatória.
- Parâmetros da população -> parâmetros da distribuição de probabilidade.
- Como a partir da amostra estimar os parâmetros populacionais?
  - Método da máxima verossimilhança.
  - Método dos mínimos quadrados.
  - Métodos do momentos.
  - Funções de estimação.
  - E muitos outros!

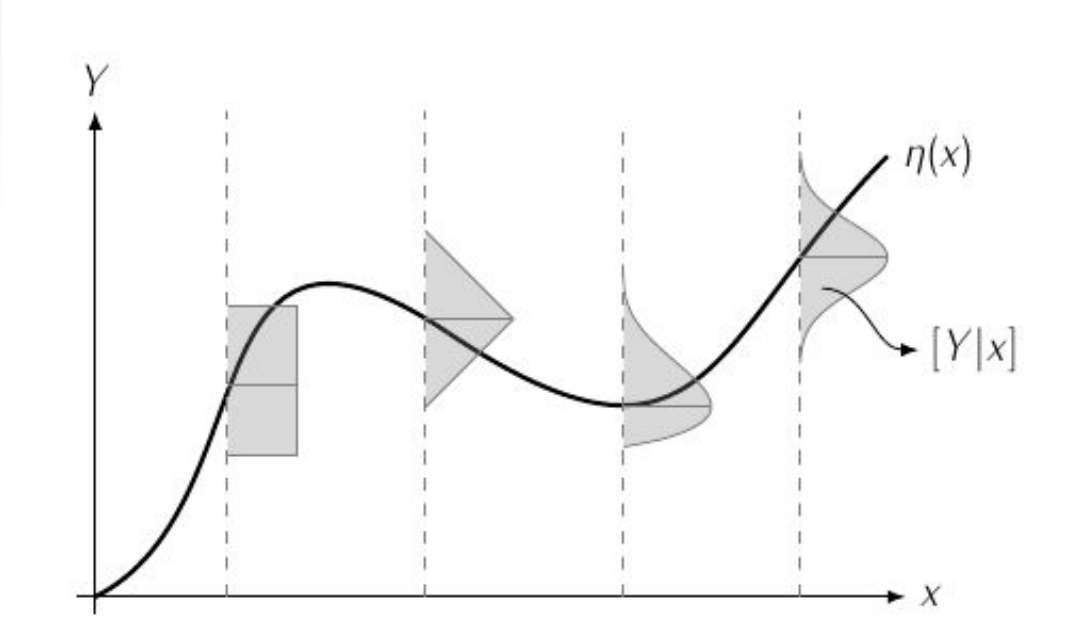


Modelos de regressão

- Conjunto de técnicas aplicadas na análise e modelagem da relação estatística entre variáveis. Generalização do que é chamado em aprendizado de máquina de aprendizagem supervisionada



Modelos de regressão



O que pensam os criadores?

All models are wrong but some are useful.

- George Box

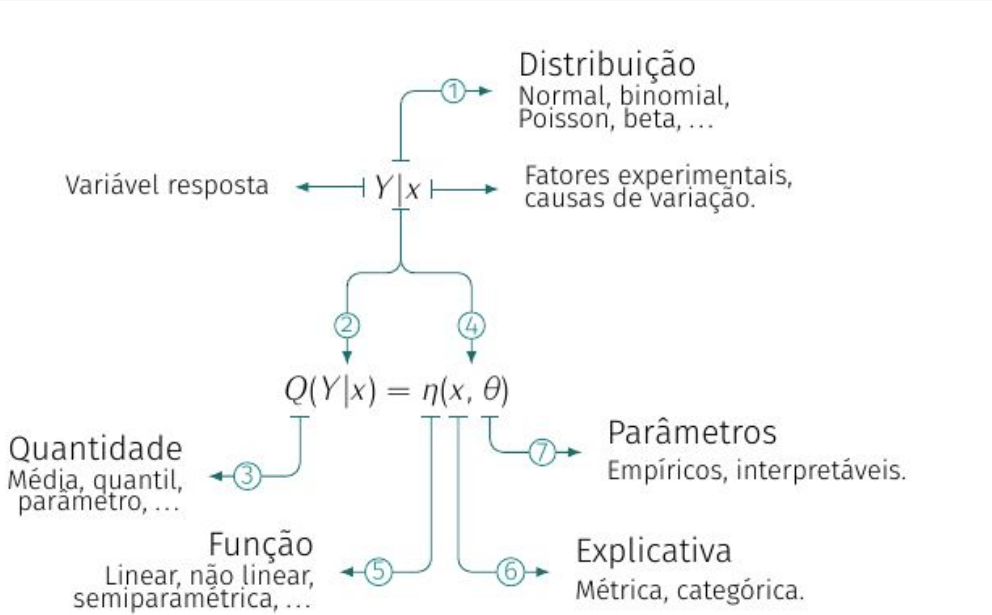
No matter how beautiful your theory, no matter how clever you are or what your name is, if it disagrees with experiment, it's wrong.

- Richard Feynman

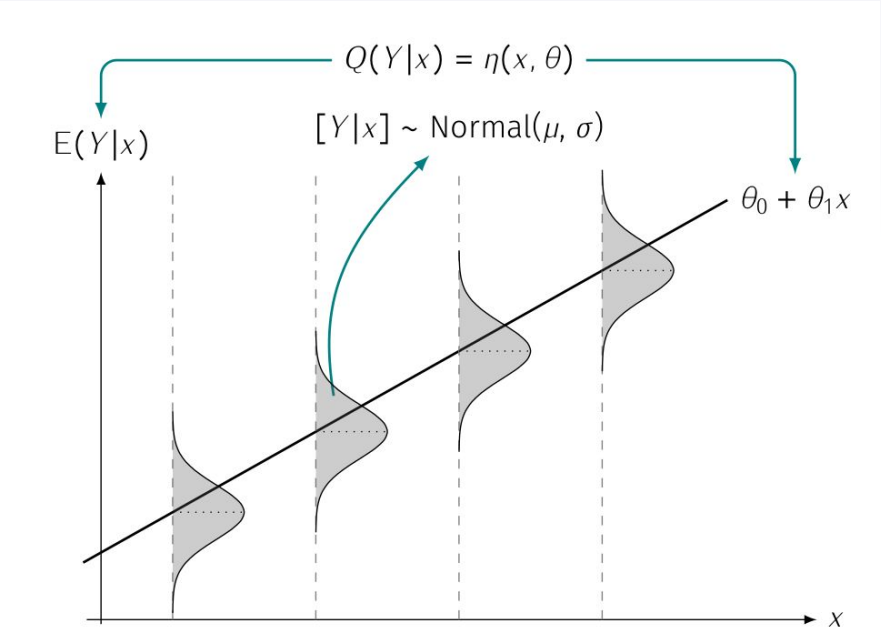
Far better an approximate answer to the right question, which is often vague,
than an exact answer to the wrong question, which can always be made precise.

- John W. Tukey

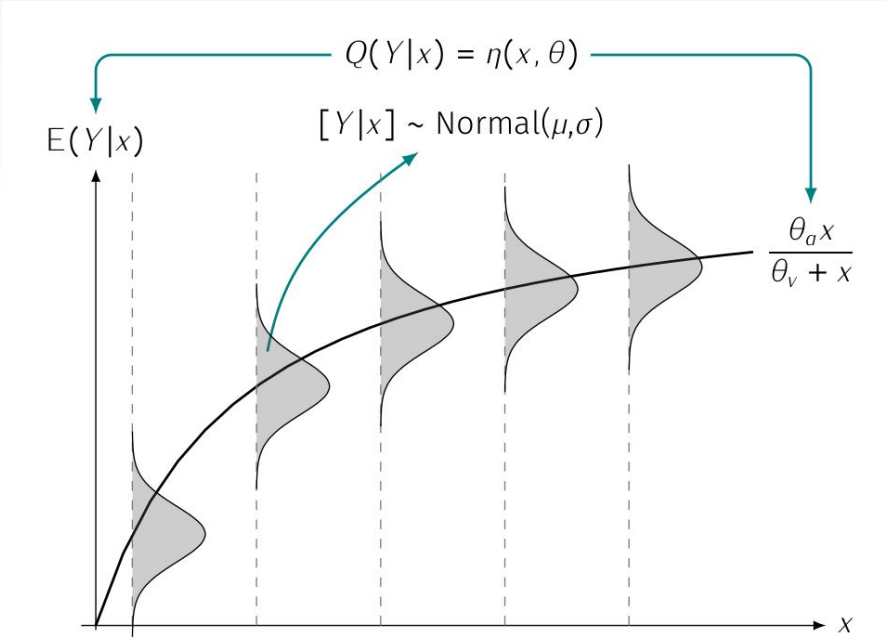
Componentes do modelo de regressão



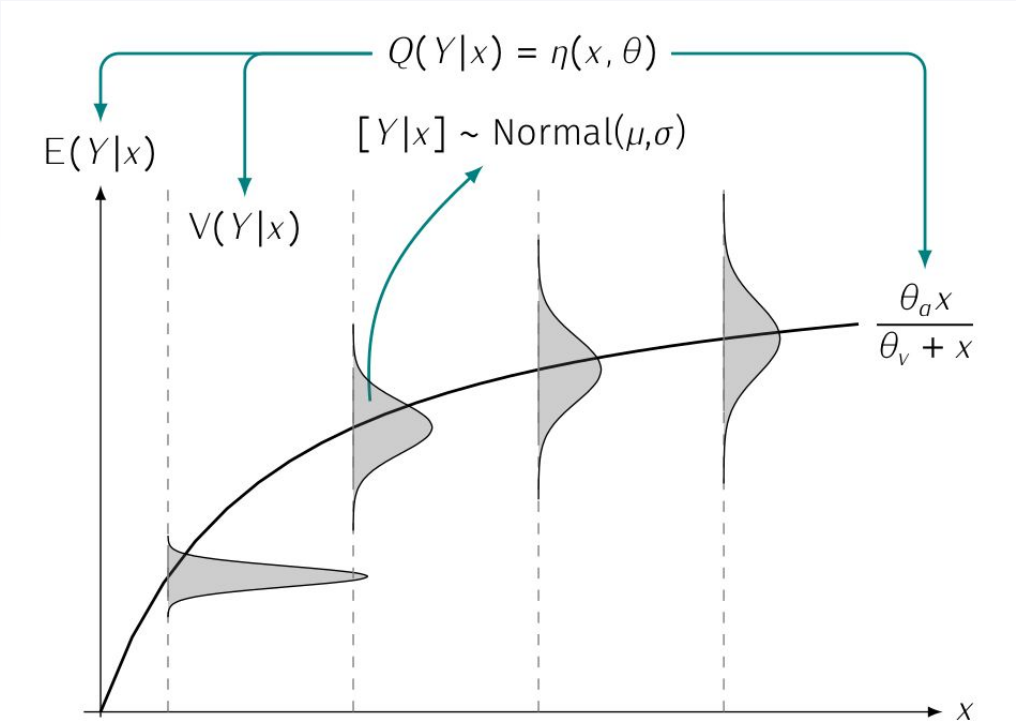
Modelos lineares



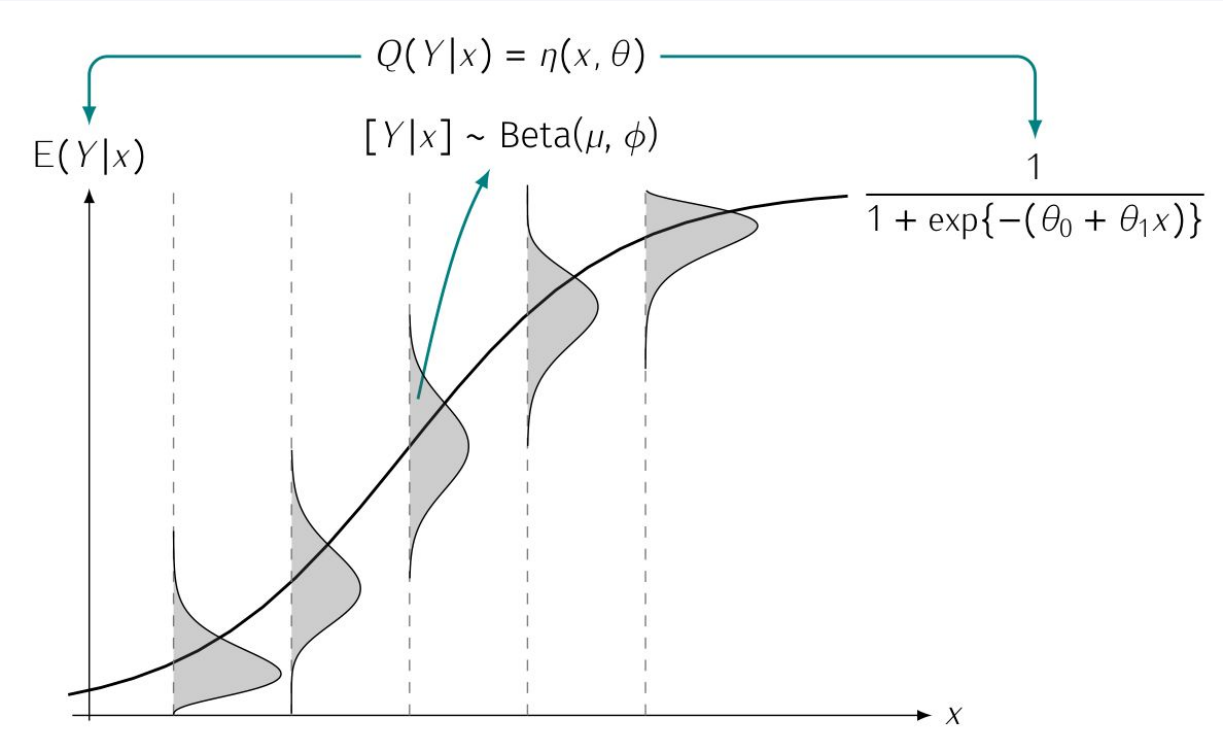
Modelos não lineares



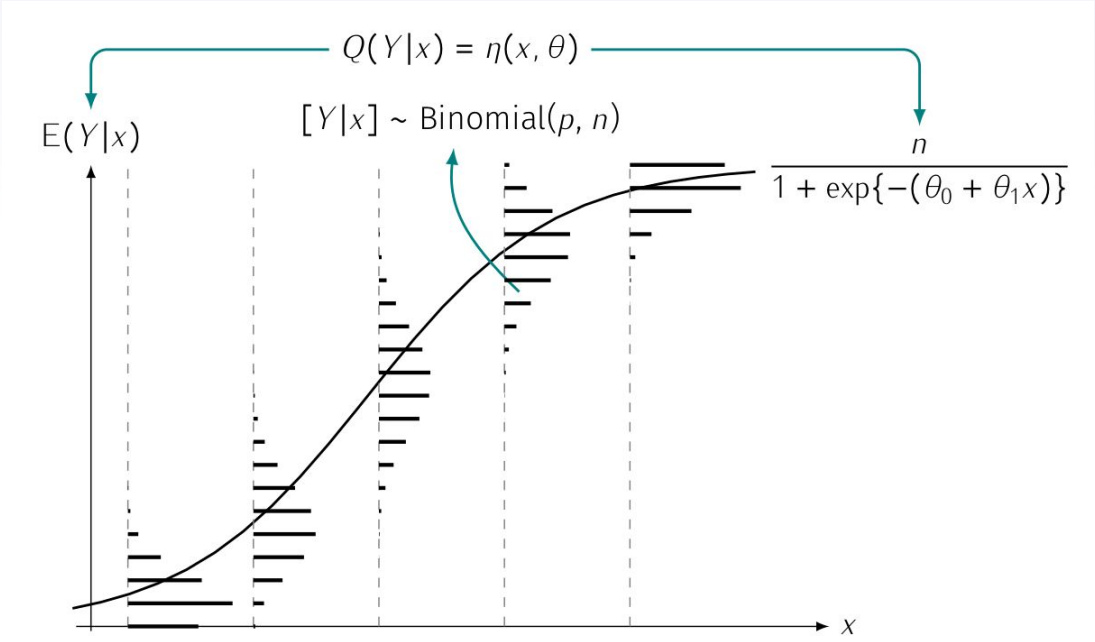
Modelos heterocedásticos



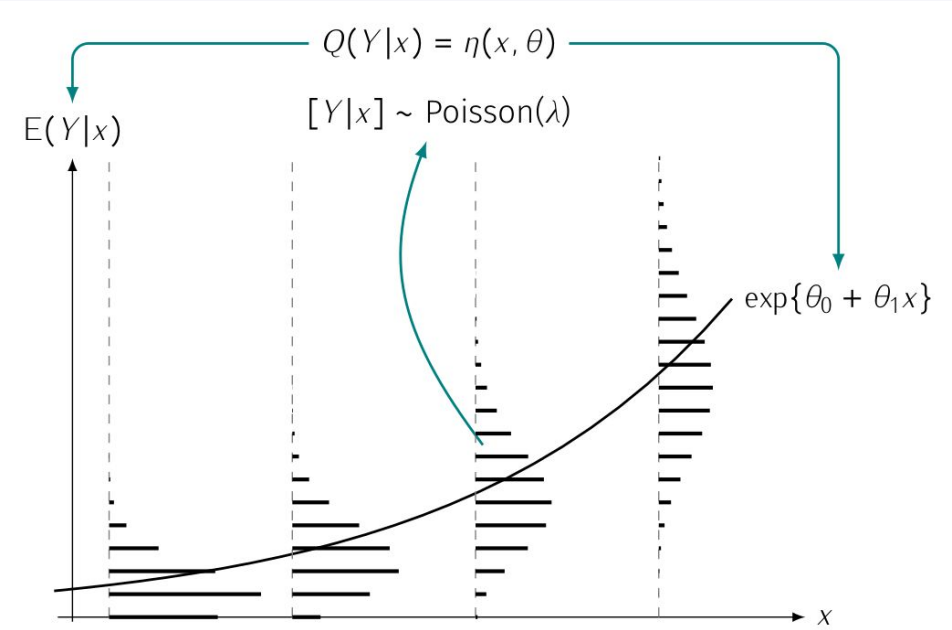
Assimétricos e/limitados



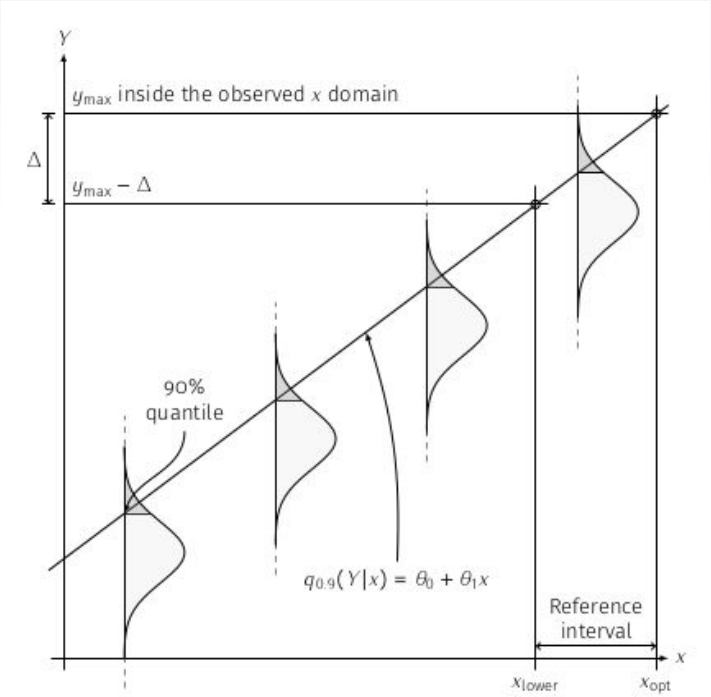
Dados binários



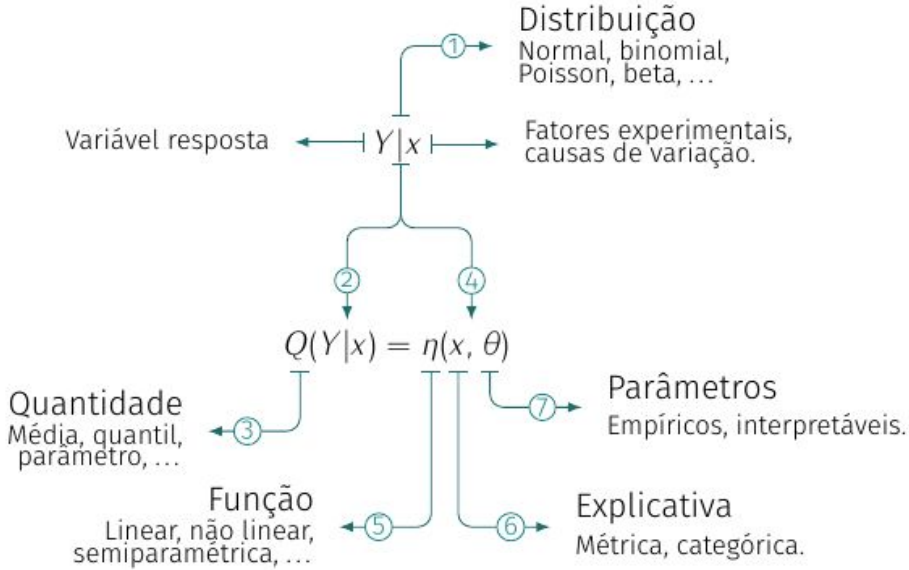
Dados de contagens



Regressão quantílica

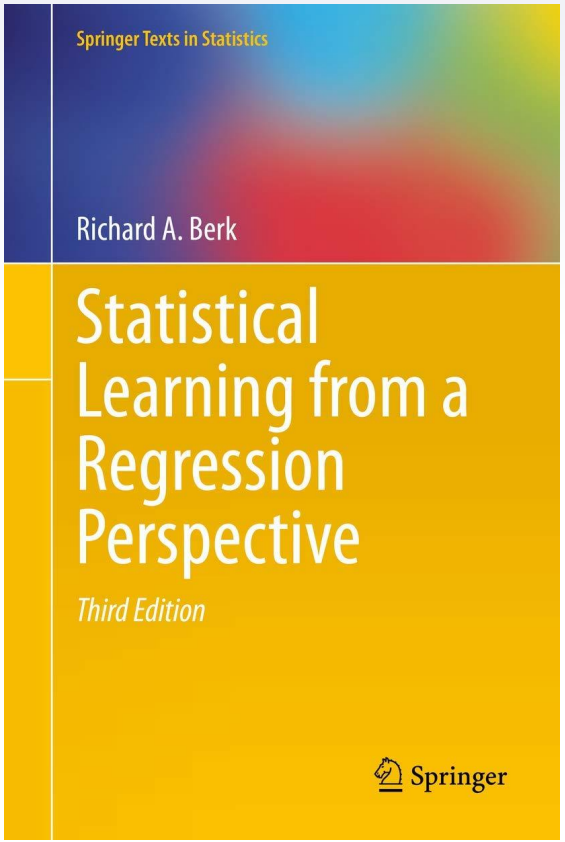


Componentes do modelo de regressão



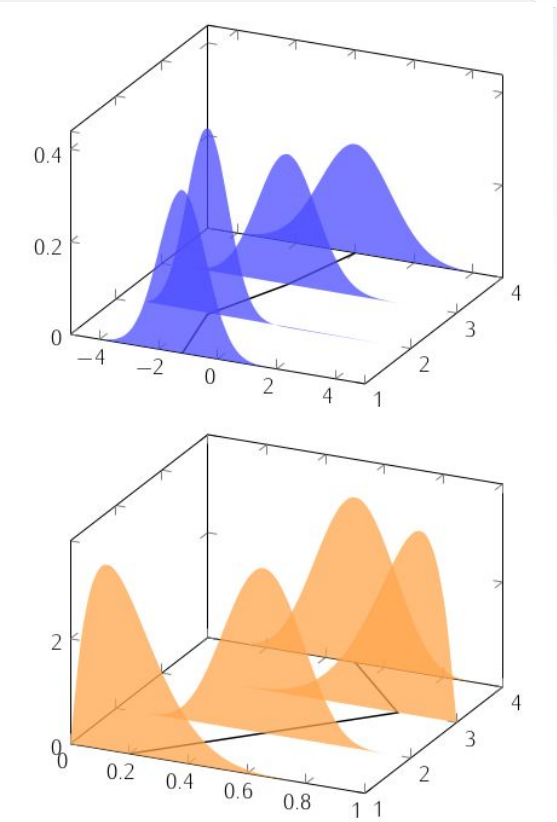
Objetivos

- Nível 1: O objetivo é meramente descritivo.
  - Livre de suposições.
  - Não envolve inferência estatística.
- Nível 2: Inferência estatística é o principal objetivo.
  - Intervalos de confiança e testes de hipóteses.
  - Núcleo da estatística convencional.
  - Uso correto com dados observacionais é um grande desafio.
- Nível 3: Inferência causal.
  - Demanda uma especificação conceitual do modelo.
  - Baseado em conhecimento específico do fenômeno.
  - Dados experimentais.
  - Aspectos confundidores devem ser cuidadosamente controlados.



Frameworks populares

- Modelos lineares(`lm()` pacote car).
- Modelos não lineares (`nls()`).
- Modelos lineares generalizados (`glm()`).
- Modelos aditivos generalizados (`gam()`).
- Modelos aditivos generalizados para locação, escala e corpo (`gamlss()`).
- Aspectos em comum:
  - Ajuste é baseado no método da Máxima Verossimilhança ou alguma aproximação deste.
- Principais limitações:
  - Conjunto limitado de distribuições.
  - Assume que as observações são independentes.
  - Apenas uma variável resposta.

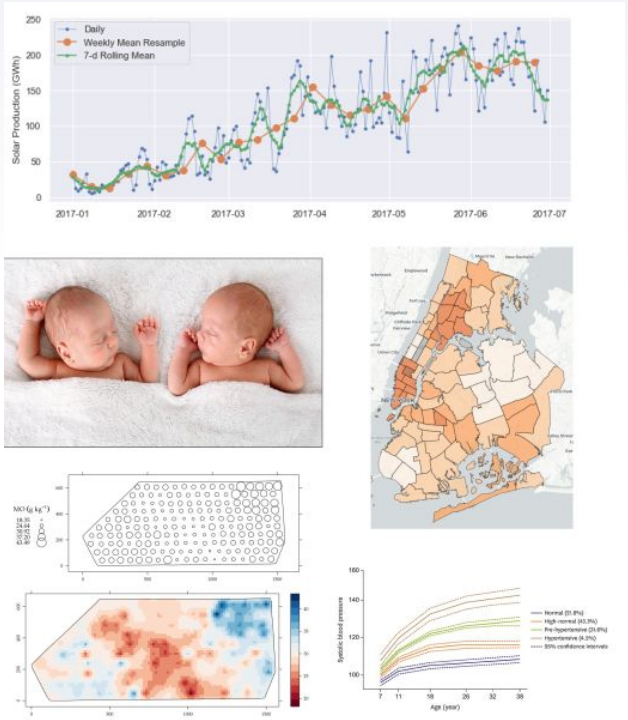


Hora do código!

Tópicos avançados

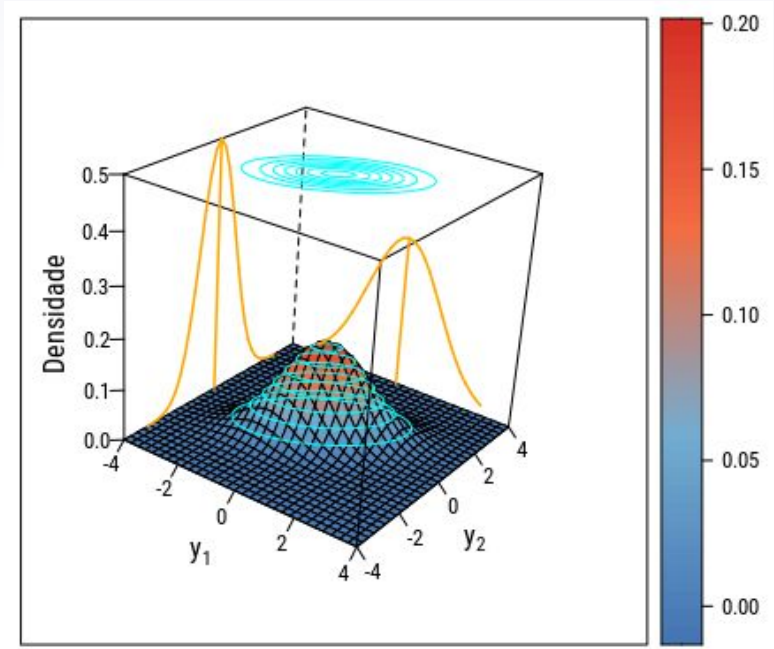
O que são observações não independentes?

- Observações com alguma referência espacial.
  - Dados de área.
  - Geostatística.
  - Processos pontuais.
- Séries temporais.
- Dados longitudinais.
- Dados de família (gêmeos).
- ...



Múltiplas variáveis resposta

- Descrição do fenômeno aleatório é feita por meio de várias variáveis aleatórias.
  - Experiência do usuário em um app: Tempo de permanência, número de cliques, número de acessos.
  - Fertilidade: Número de ovos, número de embriões, nível de estradiol e gonatropim.
  - Eficácia de fisioterapia respiratória: Taxa de respiração, batimento cardiáco, saturação de oxigênio.



Modelos multivariados de covariância linear generalizada

- McGLMs é um framework genérico para especificação de modelos estatísticos para dados não independentes e múltiplas respostas.
- Proposto em 2016 por Bonat e Jorgensen.
- Generaliza e unifica uma série de *statistical modelling frameworks*:
  - Modelos lineares generalizados (GLMs).
  - Modelos mistos (LMM).
  - Double GLMs.
  - Modelos para dados longitudinais e medidas repetidas.
  - Séries temporais.
  - Dados espaciais e espaço-temporais.
  - Dados genéticos, família e gêmeos.
  - Etc ...

Modelos multivariados de covariância linear generalizado

Appl. Statist. (2016)

Multivariate covariance generalized linear models

Wagner Hugo Bonat

Paraná Federal University, Curitiba, Brazil, and University of Southern Denmark, Odense, Denmark

and Bent Jørgensen†

University of Southern Denmark, Odense, Denmark

[Received April 2015. Revised December 2015]

Summary. We propose a general framework for non-normal multivariate data analysis called multivariate covariance generalized linear models, designed to handle multivariate response variables, along with a wide range of temporal and spatial correlation structures defined in terms of a covariance link function combined with a matrix linear predictor involving known matrices. The method is motivated by three data examples that are not easily handled by existing methods. The first example concerns multivariate count data, the second involves response variables of mixed types, combined with repeated measures and longitudinal structures, and the third involves a spatiotemporal analysis of rainfall data. The models take non-normality into account in the conventional way by means of a variance function, and the mean structure is modelled by means of a link function and a linear predictor. The models are fitted by using an efficient Newton scoring algorithm based on quasi-likelihood and Pearson estimating functions, using only second-moment assumptions. This provides a unified approach to a wide variety of types of response variables and covariance structures, including multivariate extensions of repeated measures, time series, longitudinal, spatial and spatiotemporal structures.

Keywords: Generalized Kronecker product; Linear covariance model; Matrix linear predictor; Non-normal data; Pearson estimating function; Quasi-likelihood; Spatiotemporal data

Journal of Statistical Software
April 2018, Volume 84, Issue 4. doi: 10.18637/jss.v084.i04

Multiple Response Variables Regression Models in R: The mcglm Package

Wagner Hugo Bonat
Paraná Federal University
University of Southern Denmark

Abstract

This article describes the R package **mcglm** implemented for fitting multivariate covariance generalized linear models (McGLMs). McGLMs provide a general statistical modeling framework for normal and non-normal multivariate data analysis, designed to handle multivariate response variables, along with a wide range of temporal and spatial correlation structures defined in terms of a covariance link function and a matrix linear predictor involving known symmetric matrices. The models take non-normality into account in the conventional way by means of a variance function, and the mean structure is modeled by means of a link function and a linear predictor. The models are fitted using an estimating function approach based on second-moment assumptions. This provides a unified approach to a wide variety of different types of response variables and covariance structures, including multivariate extensions of repeated measures, time series, longitudinal, genetic, spatial and spatio-temporal structures. The **mcglm** package allows a flexible specification of the mean and covariance structures, and explicitly deals with multivariate response variables, through a user friendly formula interface similar to the ordinary `glm` function. Illustrations in this article cover a wide range of applications from the traditional one response variable Gaussian mixed models to multivariate spatial models for areal data using the multivariate Tweedie distribution. Additional features, such as robust and bias-corrected standard errors for regression parameters, residual analysis, measures of goodness-of-fit and model selection using the score information criterion are discussed through six worked examples. The **mcglm** package is a full R implementation based on the **Matrix** package which provides efficient access to **BLAS** (basic linear algebra subroutines), **Lapack** (dense matrix), **TAUCS** (sparse matrix) and **UMFPACK** (sparse matrix) routines for efficient linear algebra in R.

Keywords: multivariate regression models, estimating functions, mixed outcomes, semi-parametric models, Tweedie distribution, Poisson-Tweedie distribution, R.

Projeto V: Ajustando um modelo

- Reconhecimento de gênero baseado em característica da voz.
- Aplicação da regressão logística.
- Conjunto de dados: voice.csv
- Kaggle: https://www.kaggle.com/primaryobjects/voicegender
- Disponível em leg.ufpr.br/~wagner/data