Generalized Linear Models

Davi Moura Alixandro Werneck

LAMFO
Laboratório de Aprendizado de Máquina em Finanças e Organizações

August 15, 2020

Sumário

Introdução

- Introduzido pelo artigo *Generalized Linear Models* de Nelder e Wedderburn (1972).
- Restrito à academia até os anos 90 (complexidade em trabalhar com os computadores e softwares da época).
- Função de unificar e gerar uma síntese de toda a modelação estatística até então desenvolvida.
- Utilizado quando a variável dependente segue uma distribuição da família exponencial.
- Funciona com modelos confirmatórios, dependência ou preditivos.

Características(I)

- Conceito: Generalização flexível de uma regressão linear ordinária que permite variáveis de resposta que têm modelos de distribuição de erro diferentes de uma distribuição normal.
- Máximo Verossimilhança e MQO.
- Variáveis de resposta Y podem ser contínua, discreta ou dicotômica.
- Covariáveis (determinadas ou estocásticas) podem ser de qualquer natureza.

Características (II)

- Utilização em diferentes distribuições para erros.
- Função de Ligação.
- 3 pontos precisa se definir quando se trata de problema de modelagem.
  - Comportamento (distribuição) da variável resposta.
  - Variáveis explicativas.
  - Função de ligação que irá ligar as variáveis explicativas à variável resposta.
- Limitação: Erros precisam ser independentes. Como consequência, não consegue modelar bancos de dados com estruturas longitudinais. Somente com GLM mistos ou Equações de Estimações Generalizadas para conseguir resolver esse problema.

Família exponencial

Uma distribuição é dita da família exponencial se sua densidade pode ser escrita na seguinte forma:

Estrutura

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

Onde:

- $a(\phi)$, $b(\theta)$ e $c(y,\phi)$ são funções específicas
- θ é o parâmetro natural (*função de ligação canônica*)
- ϕ é o parâmetro de dispersão

Propriedades:

- $\mu = E(Y) = b'(\theta)$
- $\sigma^2 = Var(Y) = b''(\theta)a(\phi)$

Componentes

- Todos os GLMs possuem três componentes
  - Componente aleatório - identifica a variável de resposta Y
  - Componente sistemático - Especifica as variáveis explicativas
  - Função de ligação - Especifica uma função do valor esperado de Y

Componente aleatório e sistemático

- Componente aleatório
  - Identificação da variável Y e seleção da distribuição de probabilidade.
  - GLMs padrões tratam $Y_1, ..., Y_n$ como independentes
  - Y pode tomar valores binários, uma contagem ou valores contínuos
- Componente sistemático
  - As variáveis explicativas entram linearmente na equação, sendo chamados de preditores lineares

Função de Ligação

- A função de ligação especifica uma função $g(.)$ tal que:

$$g(\mu) = \alpha + \beta_1 x_1 + ... + \beta_k x_k$$

  Dado que $E(Y) = \mu$
- a função de ligação mais simples é a identidade, tal que $g(\mu) = \mu$
- Outras funções de ligação permitem que μ seja relacionado não linearmente com as variáveis explicativas

Funções de ligação não lineares

- Tem-se a ligação log, que é utilizada quando os valores não podem ser negativos, log linearizando o modelo:

$$log(\mu) = \alpha + \beta_1 x_1 + ... + \beta_k x_k$$

- Para casos binários utiliza-se a ligação *logit*, onde:

$$g(\mu) = log \frac{\mu}{1-\mu}$$

Quadro com alguns modelos

Modelos de regressão	Características da variável dependente Y	Distribuição	Função de Ligação Canônica
Linear	Quantitativo	Normal	$\hat{Y}$
Logística Binária	Qualitativa com 2 categorias (Dummy)	Bernoulli	$\ln(\frac{p}{1-p})$
Logística Multinominal	Qualitativa com 3 ou mais categorias resposta	Binomial	$\ln(\frac{pm}{1-pm})$
Poisson	Quantitativo não inteiro	Poisson	$\ln(\lambda)$
Binomial Negativa	Quantitativa com dados de contagem	Poisson-Gama	$\ln(\lambda)$

Exemplo:Modelo de Poisson

- O modelo de Poisson é utilizado para dados estilo contagem, ou seja, o número de vezes que um evento ocorreu.
  - Número de pedidos de seguros
  - Quantidade de acidentes
  - Número de decaimentos de uma fonte radioativa

Exemplo:Modelo de Poisson

- Função de probabilidade:

$$f_Y(y|\lambda) = \frac{\lambda^y}{y!}e^{-\lambda}, y \in \mathbb{N}$$

- Forma da família exponencial

$$f_Y(y|\lambda) = \exp\left(ylog(\lambda) - \lambda - log(y!)\right)$$

Exemplo:Modelo de Poisson

- Teremos:

$$\theta = log(\lambda); b(\theta) = \lambda = \exp(\theta); a(\phi) = 1; c(y, \phi) = -log(y!)$$

- Média e Variância

$$\mu = b'(\theta) = \exp(\theta) = \lambda$$
$$\sigma^2 = b''(\theta)a(\phi) = \exp(\theta) = \lambda$$

Exemplo: Modelo de Poisson

Sob a ligação canônica temos que:

$$\theta = log(\lambda) = g(\mu) = \mathbf{X}'\beta$$

Então

$$g(E(Y)) = g(\lambda) = log(\lambda)$$

Logo, para o modelo de poisson a função de ligação é a log.

Exemplo:Modelo de Poisson

Possiveis problemas

- Como o modelo requer que a variância e a esperança da variável sejam iguais pode-se utilizar o modelo binomial negativo, que utiliza uma distribuição Poisson-Gamma
- O excesso de zeros, como em o número de cigarros fumados em um período com uma amostra que possui não fumantes

Estimação do modelo

A estimação do modelo linear generalizado é feito pela maximização do log verossimilhança:

$$L - lnl(y;\theta,\phi) = \sum lnf(y_i,\theta_i,\phi) = \sum \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi) \right\}$$

Para sua resolução é utilizado o método de Newton-Raphson ou Escore de Fisher