University of Zurich[UZH]

EBPI Epidemiology, Biostatistics and Prevention Institute

Generalised Regression

Torsten Hothorn & Luisa Barbanti, Lucas Kook

UZH Master Biostatistics, FS 2019

Introduction

The Central Dogma of Statistics

Everything is in the distribution:

$$Y \sim \mathbb{P}_Y$$

Y is called response (outcome, dependent, endogenous) variable (actually: “random” variable)

Regression Analysis

Everything is in the *conditional* distribution:

$$Y \mid \boldsymbol{X} = \boldsymbol{x} \sim \mathbb{P}_{Y \mid \boldsymbol{X} = \boldsymbol{x}}$$

$\boldsymbol{X}$ (typically multivariate) are called explanatory (design, independent, exogenous, predictor) variables or covariates

How do changes in $\boldsymbol{x}$ propagate to changes in $\mathbb{P}_{Y \mid \boldsymbol{X} = \boldsymbol{x}}$?

NB: Terminology

"Regression" classically means

$$Y \in \mathbb{R}$$

$$Y \mid \boldsymbol{X} = \boldsymbol{x} \sim \mathrm{N}(\mu(\boldsymbol{x}), \sigma^2) \Rightarrow \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \mu(\boldsymbol{x})$$

"Generalised Regression" means

$$Y \mid \boldsymbol{X} = \boldsymbol{x} \sim \mathbb{P}_{Y \mid \boldsymbol{X} = \boldsymbol{x}}$$

without these restrictions.

Univariate Distributions

Distribution Functions

Sample space: $Y \in \Xi$
σ-algebra: $\mathfrak{C}$ (in essense the set of "suitable" subsets of Ξ)
Probability measure: $\mathbb{P}_Y : \mathfrak{C} \to [0, 1]$
Distribution: $Y \sim \mathbb{P}_Y$
$A \in \mathfrak{C}$ is called *event* and $\mathbb{P}_Y(A)$ is a *probability*

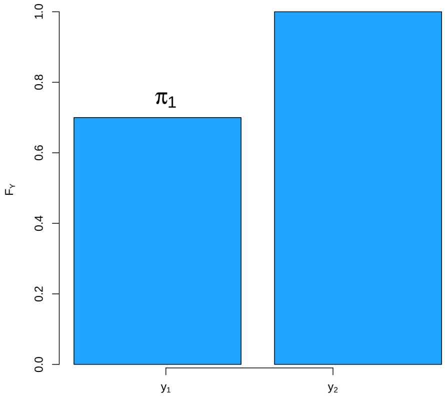
Cumulative distribution function: $F_Y : \Xi \to [0, 1]$ with

$$F_Y(y) = \mathbb{P}_Y(\{\nu \in \Xi \mid \nu \leq y\})$$

F_Y is monotone non-decreasing

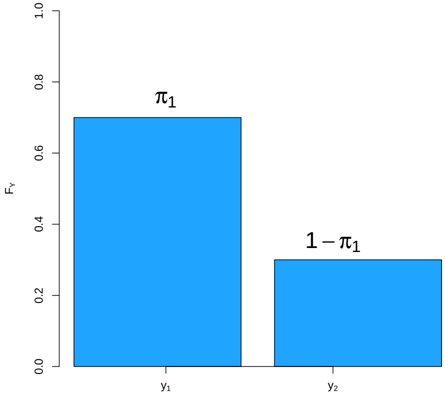
Dichotomous Variables

$$Y \in \{y_1, y_2\}, F_Y(y_1) = \pi_1, F_Y(y_2) = 1$$



Dichotomous Variables

Density function: $f_Y : \Xi \to \mathbb{R}^+$
$f_Y(y_1) = F_Y(y_1) = \pi_1$
$f_Y(y_2) = F_Y(y_2) - F_Y(y_1) = 1 - \pi_1$



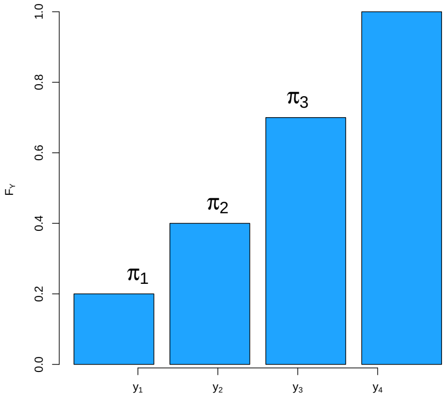
Dichotomous Variables

Odds function: $O_Y : \Xi \to \mathbb{R}^+$
$O_Y(y) = \frac{F_Y(y)}{1-F_Y(y)}$
$O_Y(y_1) = \frac{\pi_1}{1-\pi_1}$

NB: This is equivalent to $Y \sim \mathrm{B}(1, \pi_1)$

Polytomous Variables

$$Y \in \{y_1, y_2, \dots, y_K\}, F_Y(y_k) = \pi_k, F_Y(y_K) = 1$$



Polytomous Variables

Density function: $f_Y(y_1) = \pi_1$,
$f_Y(y_k) = F_Y(y_k) - F_Y(y_{k-1}) = \pi_k - \pi_{k-1}$

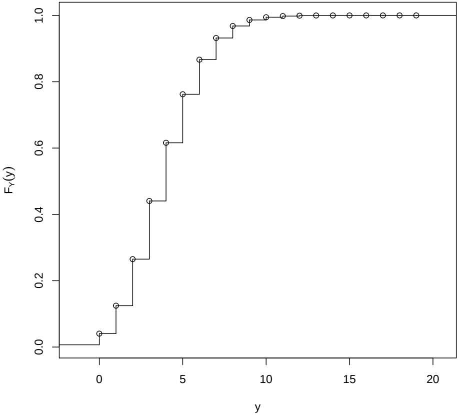
NB: $F_Y(y_k) = \sum_{i=1}^{k} f_Y(y_i)$

Odds function: $O_Y(y_k) = \frac{F_Y(y_k)}{1-F_Y(y_k)} = \frac{\pi_k}{1-\pi_k}$

NB: ordered polytomous means $y_1 < y_2 < \cdots < y_K$
NB: This is equivalent to a multinomial distribution

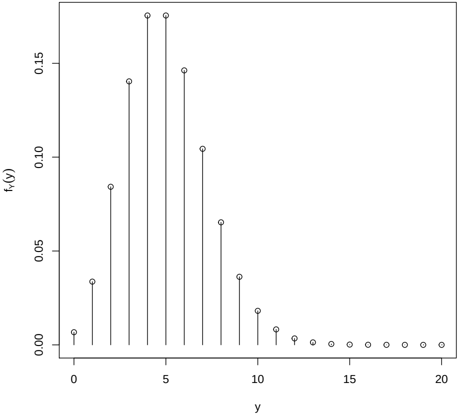
Count Variables

$\Xi = \mathbb{N}$, $F_Y(i) = \pi_i$, $F_Y(\infty) = 1$



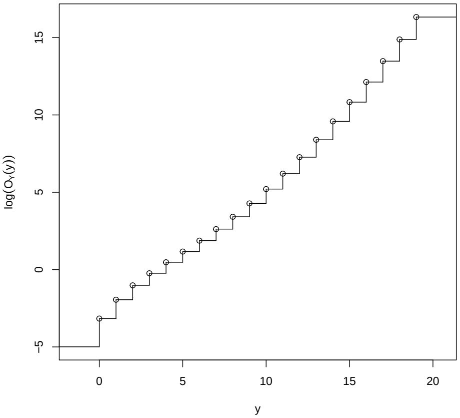
Count Variables

Density function: $f_Y(0) = \pi_0, f_Y(i) = \pi_i - \pi_{i-1}$



Count Variables

Odds function: $O_Y(i) = \frac{\pi_i}{1-\pi_i}$



Continuous Variables

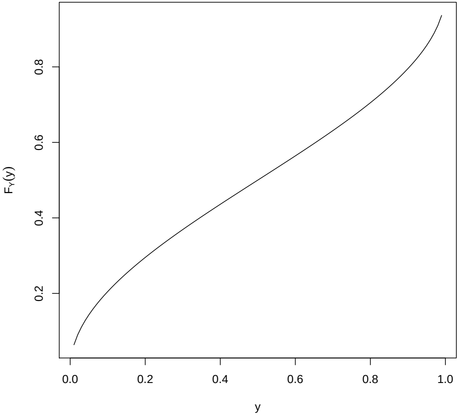
Bounded: $\Xi = (0, 1)$

Positive: $\Xi = (0, \infty)$

Real: $\Xi = \mathbb{R}$

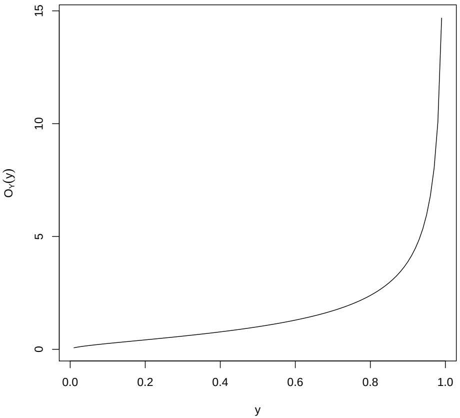
Bounded Continuous Variables

Cumulative distribution function: $F_Y : (0,1) \to [0,1]$



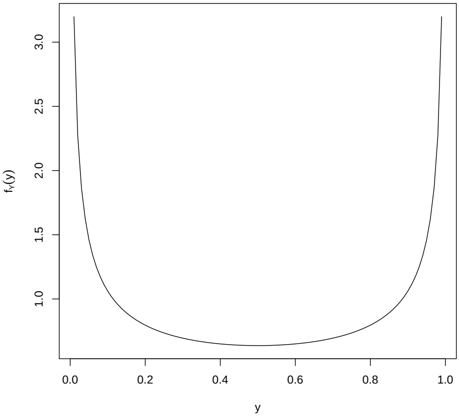
Bounded Continuous Variables

Odds function: $O_Y(y) = \frac{F_Y(y)}{1-F_Y(y)}$



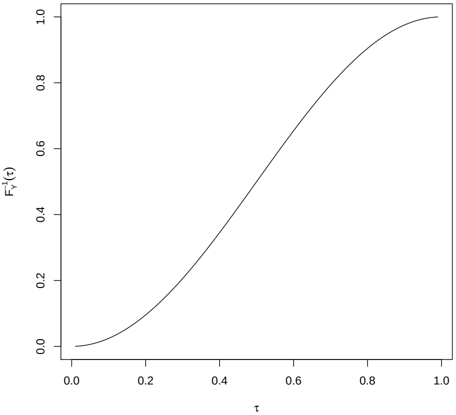
Bounded Continuous Variables

Density function: $f_Y(y) = F'_Y(y)$



Bounded Continuous Variables

Quantile function: $F_Y^{-1}(\tau)$ with $F_Y^{-1} : (0, 1) \rightarrow \Xi$



NB: Densities

Density wrt dominating measure μ: $\mathbb{P}_Y = f_Y \odot \mu$

$$F_Y(y) = \int \mathbb{1}(u \leq y) f_Y(u)\, d\mu(u) \stackrel{y=\infty}{=} 1$$

Discrete (wrt counting measure):

$$F_Y(y) = \sum_{u \in \Xi} \mathbb{1}(u \leq y) f_Y(u)$$

$f_Y(y)$ is a probability

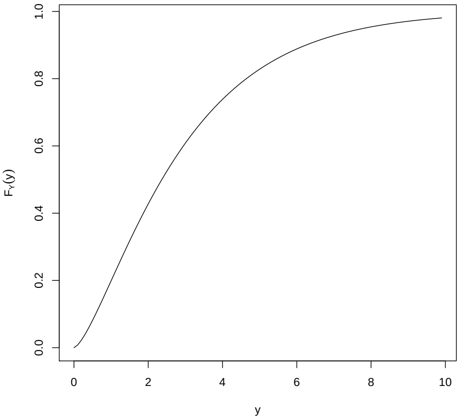
Continuous (wrt Lebesque measure):

$$F_Y(y) = \int \mathbb{1}(u \leq y) f_Y(u)\, du$$

$f_Y(y)$ is NOT a probability (but risk or intensity)

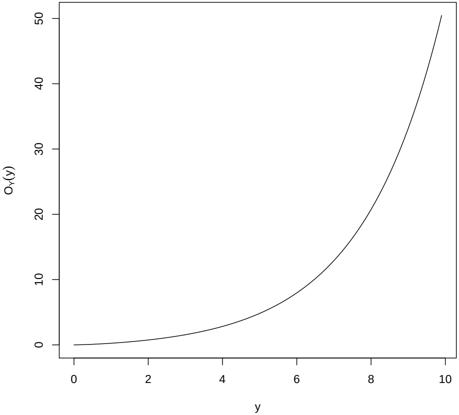
Positive Continuous Variables

Cumulative distribution function: $F_Y : (0, \infty) \to [0, 1]$



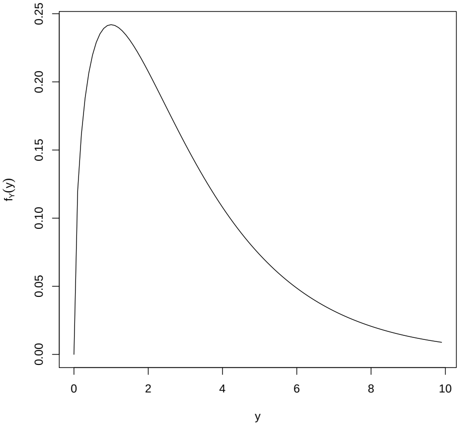
Positive Continuous Variables

Odds function: $O_Y(y) = \frac{F_Y(y)}{1-F_Y(y)}$



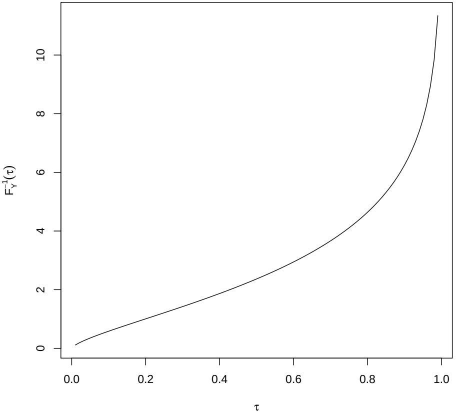
Positive Continuous Variables

Density function: $f_Y(y) = F'_Y(y)$



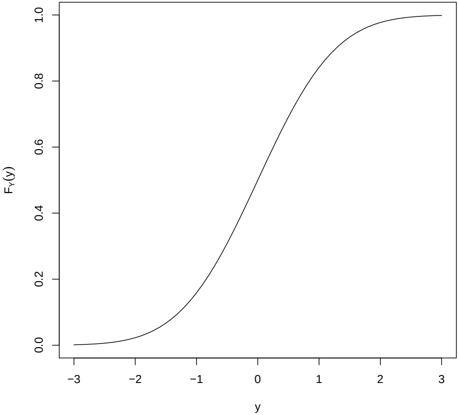
Positive Continuous Variables

Quantile function: $F^{-1}_{Y|\boldsymbol{X}=\boldsymbol{x}}(\tau)$



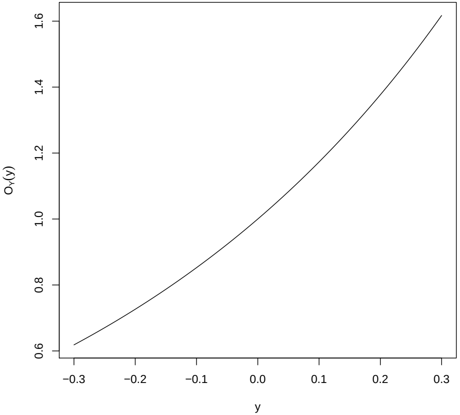
Real Continuous Variables

Cumulative distribution function: $F_Y : \mathbb{R} \to [0, 1]$



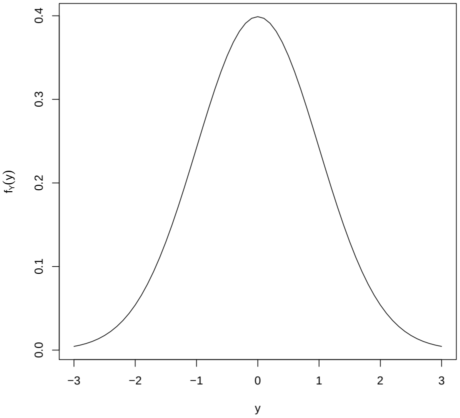
Real Continuous Variables

Odds function: $O_Y(y) = \frac{F_Y(y)}{1-F_Y(y)}$



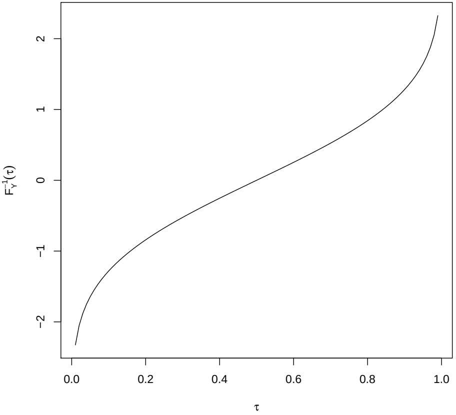
Real Continuous Variables

Density function: $f_Y(y) = F'_Y(y)$



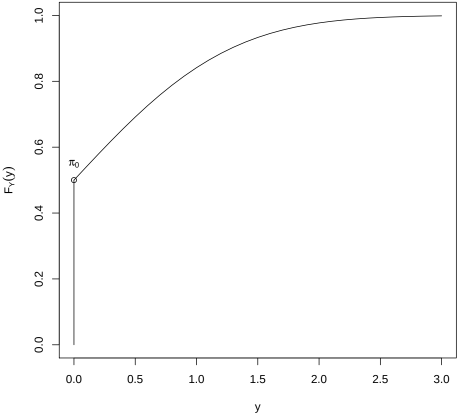
Real Continuous Variables

Quantile function: $F^{-1}_{Y|\boldsymbol{X}=\boldsymbol{x}}(\tau)$



Mixed Discrete / Continuous Variables

$\Xi = [0, \infty)$, $F_Y(0) = \pi_0$



Parameterisations

Dicho/Polytomous: $\pi_1, \dots, \pi_{K-1}$ (binomial or multinomial distribution); fully parameterised

Sparse parameterisations ("shape" given, few parameters $\boldsymbol{\vartheta}$):
Count: Poisson, Negative-binomial, ...

Bounded continuous: Beta

Positive continuous: χ^2, F, Weibull, log-normal, ...

Real: Normal, Logistic, t, ...

write $F_Y(Y \mid \boldsymbol{\vartheta})$

Estimation (1: "nonparametric")

$Y_1, \ldots, Y_N$ iid $Y_i \sim \mathbb{P}_Y$

$$\hat{F}_{Y,N}(y) = N^{-1} \sum_{i=1}^{N} \mathbb{1}(Y_i \leq y)$$

Empirical cumulative distribution function;
$| F_Y(y) - \hat{F}_{Y,N}(y) | \rightarrow 0$ a.s. when $N \rightarrow \infty$ (Glivenko-Cantelli)

Estimation (2: “parametric”)

$Y_1, \dots, Y_N$ iid $Y_i \sim \mathbb{P}_Y$ with $F_Y(y \mid \boldsymbol{\vartheta})$, unknown parameter(s) $\boldsymbol{\vartheta} \in \Theta$ (parameter space)

$$\hat{\boldsymbol{\vartheta}}_N = \operatorname*{arg\,max}_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} \log(f_Y(Y_i \mid \boldsymbol{\vartheta}))$$

Maximum-likelihood estimation (Fisher, 1922); very general and some nice properties

Things are a bit more complex, but for the moment we’ve got everything we need.