同济大学
TONGJI UNIVERSITY

二元离散选择模型理论

同济大学
交通运输工程学院
叶 昕 教授

同济大学交通运输工程学院
SCHOOL OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

学习目的

- 理解离散选择模型的用途
- 理解选择过程中的基本要素
- 理解随机效用函数的概念
- 了解二元选择模型的推导
- 掌握二元Logit模型的概率计算公式

什么是离散选择？

- 线性回归模型中的连续因变量 y,

  $y = x\beta + \varepsilon$ 或者 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$

- 影响交通需求的个体选择
  - 住在郊区还是市区？
  - 上班是开车去还是乘地铁去？
  - 高峰时段外出购物还是避开高峰时段外出购物？
- 无法用连续变量来表示这类离散变量

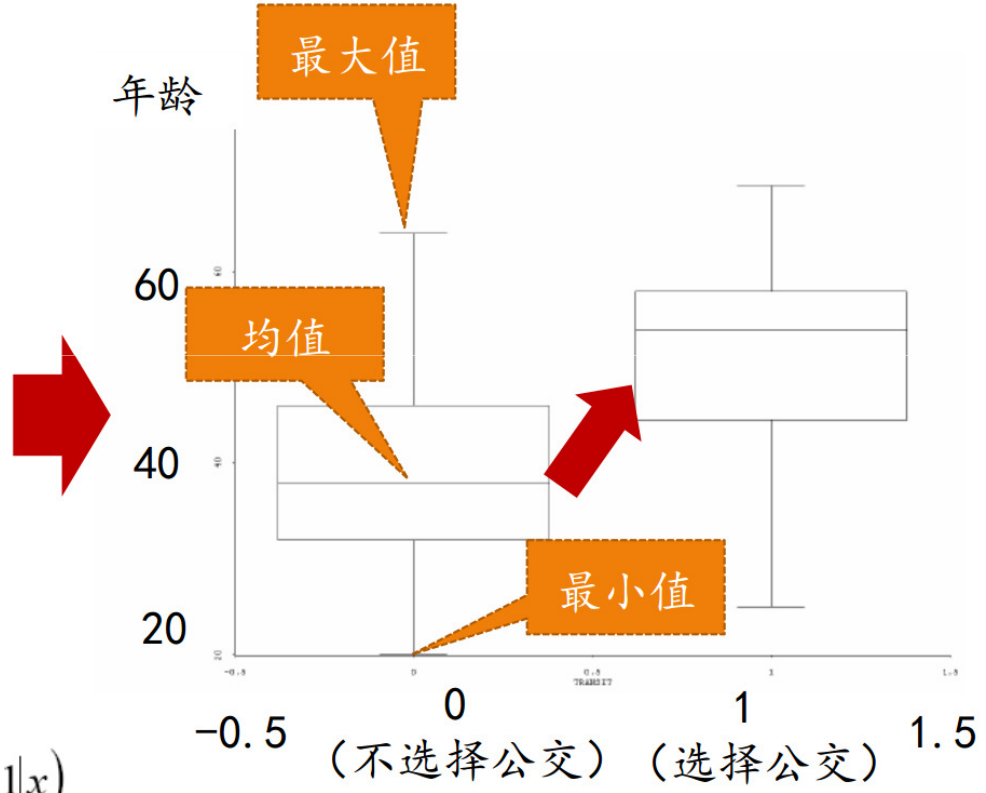
能否使用线性回归模型分析离散选择? [1]

- 若将选择变量 y 设定为0-1变量
- 1 表示选择某选项，0 表示没有选择
- 例如，如果 y = 1，表示出行者选择公交出行方式；当 y = 0， 表示出行者没有选择公交出行方式。
- 能否对于 y 变量进行线性回归分析呢？

能否使用线性回归模型分析离散选择？[2]

- 例子
  - 问题：分析年龄对于选择公交出行方式的影响；
  - 数据：

	被说明变量（Y）	说明变量（X）
居民	公交	年龄
1	1	35
2	1	65
3	0	25
…	…	…
n	0	45

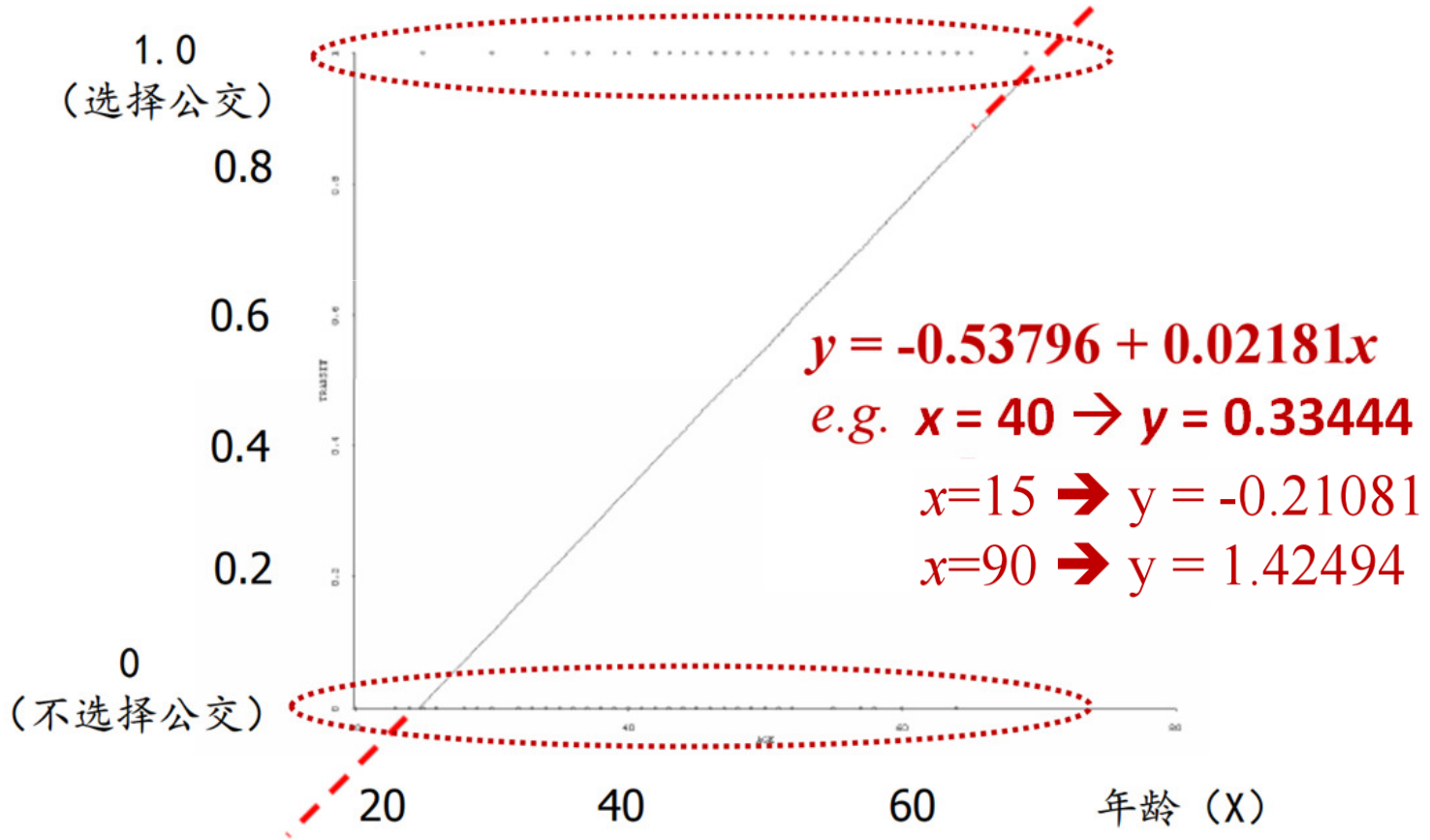


  - 线性回归模型 $E(y|x) = P(y=1|x)$

$$= \beta_0 + \beta_1 age$$

能否使用线性回归模型分析离散选择？[3]

- 例子
  - 线性回归模型的拟合曲线



能否使用线性回归模型分析离散选择？[4]

- 在线性回归$y = x\beta + \varepsilon$中对于随机干扰项 ε 做了什么分布假设？
  - 正态分布假设
- 为什么需要这个假设？
  - 为了对于系数估计值进行统计推断
- 如果y是0-1变量，这个假设还能满足吗？
  - $\varepsilon = y - x\beta$
  - 基本不可能满足正态分布假设,无法进行可靠的统计推断

分析离散因变量的离散选择模型

- 怎么把离散因变量 y 和 $x\beta + \varepsilon$ 联系起来？
- 2000年度诺贝尔经济学奖得主：丹尼尔·麦克法登教授的获奖成果
- 源于交通的路线选择问题

Prof. Daniel McFadden (1937 -)

选择过程中的基本要素

- 决策者 （Decision Maker）
  - 决策者可能具有不同的价值取向 (Value)
- 选项 （Alternatives）
  - 普遍选择集（Universal Set）和选择集（Choice Set）
- 选项属性 （ Attributes of Alternatives）
  - 一般属性 (Generic Attribute)
  - 选项特有属性 (Alternative-specific Attribute)
- 决策规则（Decision Rule）
  - 基于满足与否的决策规则（Satisfaction-based）
  - 基于效用的综合决策规则（Utility-based ）

基于效用的决策规则举例 [1]

- 如果选择集当中只有两种通勤方式：开车和乘坐地铁

属性	开车	地铁
步行时间 （分种）	5	15
等候时间 （分种）	--	5
车内时间 （分种）	35	50
汽油费 （元）	10	--
停车费 （元）	5	--
票价 （元）	--	5

- 你会选择哪种通勤方式？

基于效用的决策规则举例 [2]

- 金钱与时间属性具有可比较性和可交换性
- 如何综合评价两种出行方式？
- 对于不同的属性赋予不同的权重系数，计算综合评价指标
- 这个综合评价指标可以理解为效用（Utility）
- 效用最大化原理：理性的决策者会从选择集中选择效用最大的选项

基于效用的决策规则举例 [3]

- 如果某个决策者如下表所示，对于每项属性都给出了不同的权重系数
- 为什么权重系数是负值？

属性	权重	开车	地铁
步行时间 （分种）	-2	5	15
等候时间 （分种）	-2	--	5
车内时间 （分种）	-1	35	50
汽油费 （元）	-5	10	--
停车费 （元）	-5	5	--
票价 （元）	-5	--	5

基于效用的决策规则举例 [4]

属性	权重	开车	开车效用值	地铁	地铁效用值
步行时间 （分种）	-2	5	-10	15	-30
等候时间 （分种）	-2	--	0	5	-10
车内时间 （分种）	-1	35	-35	50	-50
汽油费 （元）	-5	10	-50	--	0
停车费 （元）	-5	5	-25	--	0
票价 （元）	-5	--	0	5	-25
总和			-120		-115

- 地铁的效用值高于开车的效用值，因此决策者将选择地铁作为通勤方式

确定性的效用函数对应确定性的选择

- 确定性的效用函数 $U = X\beta$
- 如果选择集中仅有2个选项，分别计算效用值：
  - $U_1 = X\beta = \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_n X_n$
  - $U_2 = Z\beta = \beta_1 Z_1 + \beta_2 Z_2 \ldots + \beta_n Z_n$
- 如果$U_1 > U_2$，选择选项1；如果$U_1 < U_2$，选择选项2
- 存在什么问题？
- 如何解决？

引进随机干扰项，形成随机效用函数

- 类比于线性回归模型，引进随机干扰项 ε 和常数项 β_0，形成随机效用函数 (Random Utility Function)
- 随机效用函数 $U = \beta_0 + X\beta + \varepsilon$
- ε代表什么？
  - ε用来替代一切被排除在向量X外的各种因素的影响总和
  - 建模者考虑的选项属性和决策者真正考虑的选项属性未必是一样的
- β_0代表什么？
  - 如果假设随机干扰项 ε的期望值为0，β_0则代表被排除在外的影响总和的期望值

随机效用函数的常用术语

- $U = \beta_0 + X\beta + \varepsilon$
- $V = \beta_0 + X\beta$ (叫做系统部分，Systematic Component)
- ε叫做随机部分 (Random Component)
- $U = V + \varepsilon$

比较两个选项的随机效用值，做出选择

- $U_1 = V_1 + \varepsilon_1$, $U_2 = V_2 + \varepsilon_2$
- 根据效用最大化原理：理性的决策者会从选择集中选择效用最大的选项
- 如果$U_1 > U_2$, 选择选项1；如果$U_1 < U_2$, 选择选项2
- 由于U中含有随机部分，事件$U_1 > U_2$不再是确定事件，而是随机事件
- 事件$U_1 > U_2$的发生概率用 $P(U_1 > U_2)$来表示
- 如果y是0-1变量，$y = 1$代表选项1被选择；$y = 0$代表选项2被选择。
- 如何计算$P(y = 1)$和$P(y = 0)$?

计算选择概率 [1]

- 如果$U_1 > U_2$, 选择选项1
- 因此，这两个随机事件应该具有相同的发生概率
- $P(U_1 > U_2) = P(y = 1)$
- 用$V + \varepsilon$来替换U，可得：$P(y = 1) = P(V_1 + \varepsilon_1 > V_2 + \varepsilon_2) = P(\varepsilon_2 - \varepsilon_1 < V_1 - V_2) = P(\varepsilon_{21} < V_{12})$
- 为了计算概率值，需要对 ε_1 和 ε_2做分布假设
- 正态是常用的分布假设，这个假设在这里容易满足吗？

计算选择概率 [2]

- 如果 $\varepsilon_1 \sim N(0, \sigma_1^2)$, $\varepsilon_2 \sim N(0, \sigma_2^2)$, ε_1和ε_2相互独立
- 如果$\varepsilon_{21} = \varepsilon_2 - \varepsilon_1$， ε_{21} 符合什么分布？
- $\varepsilon_{21} \sim N(0, \sigma_1^2 + \sigma_2^2)$
- 令$\sigma^2 = \sigma_1^2 + \sigma_2^2$， $\varepsilon_{21} \sim N(0, \sigma^2)$
- $P(y = 1) = P(\varepsilon_{21} < V_{12}) = P(\sigma \cdot z < V_{12})$
- z符合标准正态分布N(0, 1)
- $P(y = 1) = P(z < V_{12}/\sigma)$
- σ按比例调节V_1和V_2的值，因此可以标准化为1

二元probit模型

- 那么$P(y = 1) = P(z < V_{12}) = \Phi(V_{12})$
- $P(y = 0) = 1 - P(y = 1) = 1 - \Phi(V_{12})$

标准正态分布的密度函数：$\phi(u) = \dfrac{\exp(-u^2/2)}{\sqrt{2\pi}}$

标准正态分布的分布函数：$P(z < x) = \Phi(x) = \int_{-\infty}^{x} \phi(u)du$

- 这类模型叫做二元probit模型(Binary Probit Model)
- “Probit”指 Probability Unit

具有收敛概率表达式的二元选择模型

- 二元Probit模型的选择概率值没有收敛的解析表达式
- 能否获得具有收敛表达式的二元选择模型呢？
- 用Gumbel分布取代正态分布，即可获得收敛的概率表达式
- Gumbel分布又叫极值分布，用于描述在一段时间内出现的极大或极小值的分布规律

Gumbel分布简介

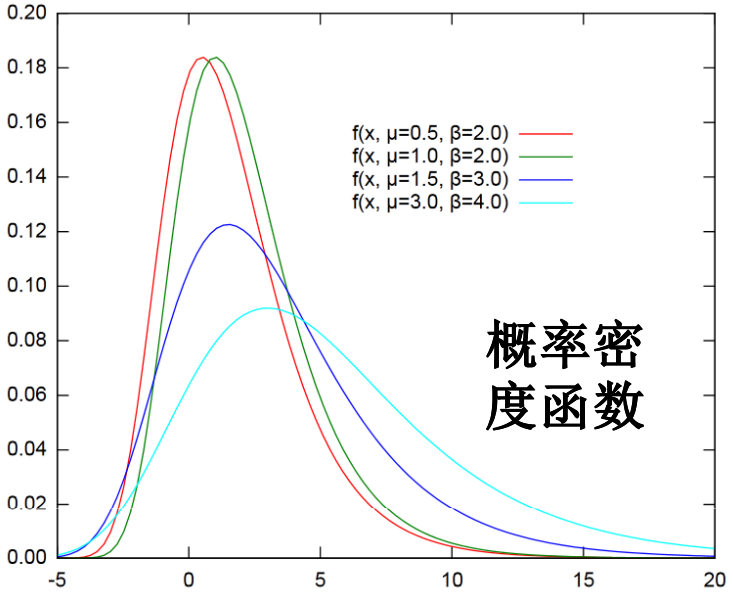
- $x \sim G(\mu, \beta)$
  - μ: 位置参数
  - β: 尺度参数 ($\beta > 0$)
- 分布函数:

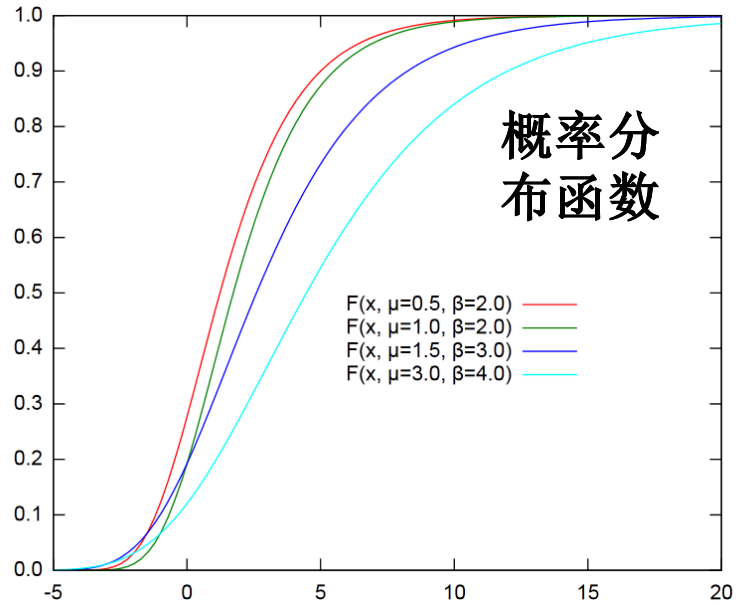
$$F(x) = \exp\left[-\exp\frac{-(x-\mu)}{\beta}\right]$$

- 密度函数:

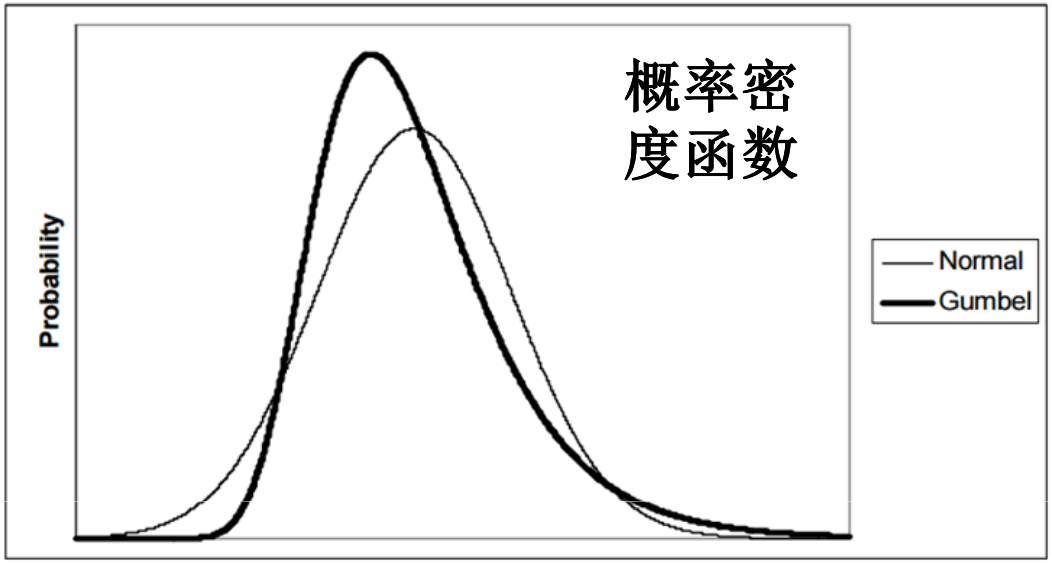
$$f(x) = \frac{1}{\beta}\exp\left\{-\exp\left[\frac{-(x-\mu)}{\beta}\right] - \frac{(x-\mu)}{\beta}\right\}$$

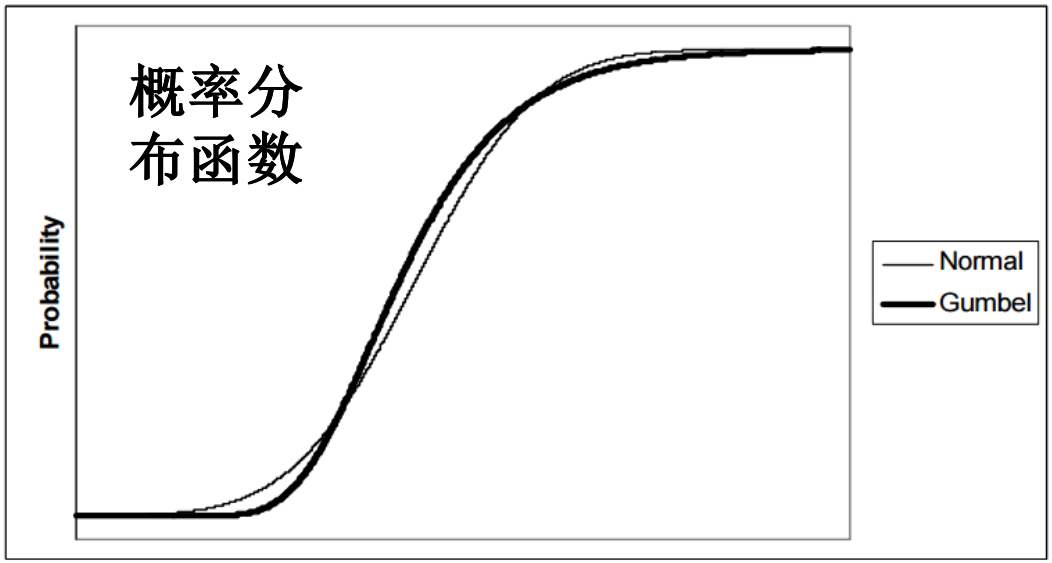
- $E(x) = \mu + \gamma\beta$ ($\gamma \approx 0.577$, 欧拉常数)
- $V(x) = \pi^2\beta^2/6$, $Mode(x) = \mu$





等期望和方差的Gumbel分布与正态分布比较





Gumbel分布的线性变换

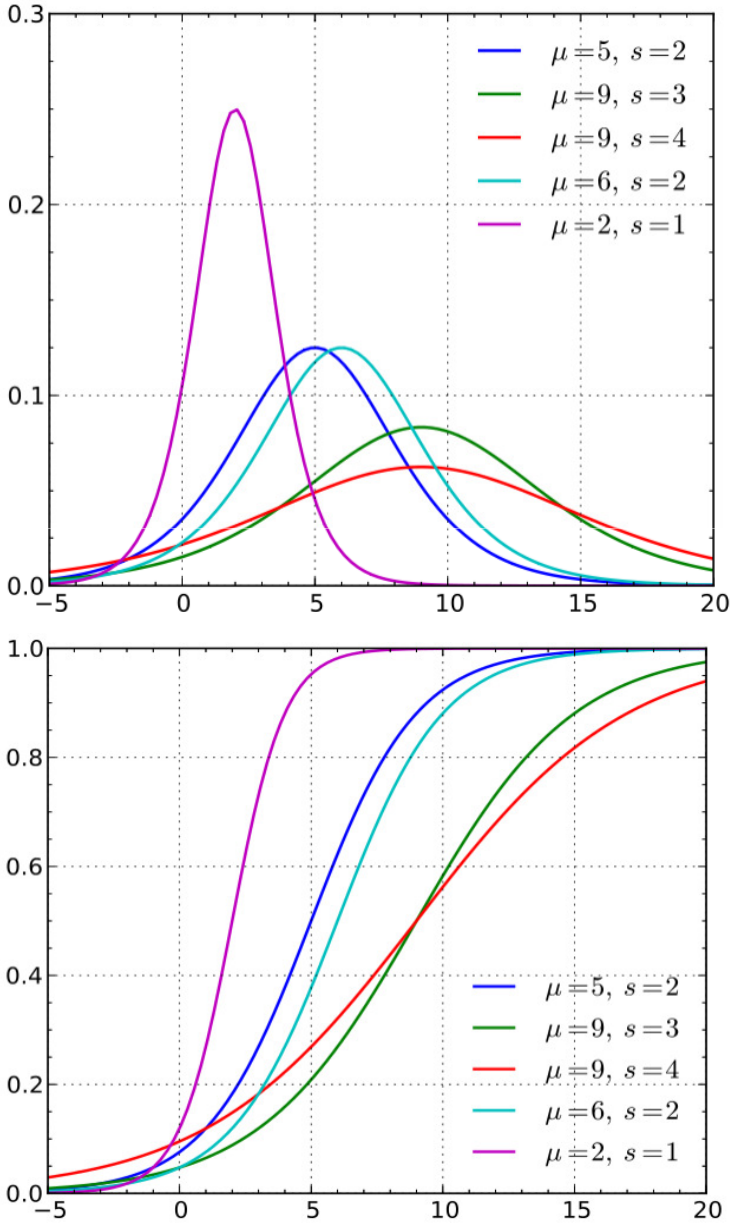
- 如果$x \sim N(\mu, \sigma^2)$, $\alpha x + \gamma$ 符合什么分布？
- $\alpha x + \gamma \sim N(\alpha\mu + \gamma, \alpha^2\sigma^2)$ （α和γ为常数）
- 如果$x \sim G(\mu, \beta)$, $\alpha x + \gamma$ 符合什么分布？
- 结论: $\alpha x + \gamma \sim G(\alpha\mu + \gamma, \alpha\beta)$
- 证明：令$z = \alpha x + \gamma$,

$$F_Z(z) = P(\alpha x + \gamma < z) = P\left(x < \frac{z-\gamma}{\alpha}\right)$$

$$= \exp\left[-\exp\frac{-\left(\frac{z-\gamma}{\alpha}-\mu\right)}{\beta}\right] = \exp\left\{-\exp\frac{-[z-(\gamma+\alpha\mu)]}{\alpha\beta}\right\}$$

Logistic分布简介

- **x ~ Logistic(μ, β)**
  - **μ: 位置参数**
  - **β: 尺度参数 (β > 0)**
- **分布函数:** $$F(x) = \frac{1}{1+\exp(-\frac{x-\mu}{\beta})}$$
- **密度函数:** $$f(x) = \frac{\exp(-\frac{x-\mu}{\beta})}{\beta \cdot \left[1+\exp(-\frac{x-\mu}{\beta})\right]^2}$$
- **E(x) = μ, V(x) = $\pi^2\beta^2/3$**
- **Mode(x) = μ**



Gumbel分布的重要属性之一

- 两个相互独立的服从Gumbel分布且等方差的随机变量之差服从Logistic分布
- 如果 $x \sim G(\mu_1, \beta)$, $y \sim G(\mu_2, \beta)$, 且x 与 y相互独立, $z = x - y$, 那么$z \sim Logistic(\mu_1 - \mu_2, \beta)$

证明过程

$$F_Z(z) = P(x - y < z) = \int_{-\infty}^{+\infty} dy \int_{-\infty}^{y+z} f_{x,y}(x,y)dx = \int_{-\infty}^{+\infty} f_y(y)dy \int_{-\infty}^{y+z} f_x(x)dx$$

$$= \int_{-\infty}^{+\infty} F_x(y+z) f_y(y)dy = \int_{-\infty}^{+\infty} \exp\left[-\exp\left(\frac{\mu_1 - y - z}{\beta}\right)\right] \frac{1}{\beta} \exp\left[-\exp\left(\frac{\mu_2 - y}{\beta}\right) + \left(\frac{\mu_2 - y}{\beta}\right)\right] dy$$

$$= \int_{-\infty}^{+\infty} \exp\left[-\exp\left(\frac{\mu_2 - y}{\beta}\right)\exp\left(\frac{\mu_1 - \mu_2 - z}{\beta}\right) - \exp\left(\frac{\mu_2 - y}{\beta}\right)\right] \frac{1}{\beta} \exp\left[\frac{\mu_2 - y}{\beta}\right] dy$$

Let $w = \exp\left(\frac{\mu_2 - y}{\beta}\right)$ then $y = \mu_2 - \beta \ln(w)$ and $dy = -\frac{\beta}{w} dw$.

Since y starts from -∞ to +∞, w starts from +∞ to 0. Let $k = \exp\left(\frac{\mu_1 - \mu_2 - z}{\beta}\right)$.

$$F_Z(z) = \int_{+\infty}^{0} \exp[-wk - w] \frac{w}{\beta} \frac{-\beta}{w} dw$$

$$= \int_{+\infty}^{0} -\exp[-wk - w] dw = \left\{\frac{\exp[(-k-1)w]}{1+k}\right\}_{+\infty}^{0}$$

$$= \frac{1}{1+k} - 0 = \frac{1}{1+\exp\left(\frac{\mu_1 - \mu_2 - z}{\beta}\right)} = \frac{1}{1+\exp\left[-\frac{z-(\mu_1 - \mu_2)}{\beta}\right]}$$

z is logistically distributed: $z \sim \text{Logistic}[(\mu_1 - \mu_2), \beta]$

$$Logistic(\mu, \beta): P(Z < z) = \frac{1}{1 + \exp\left(-\frac{z - \mu}{\beta}\right)}$$

基于Gumbel分布的二元选择模型

- $U = V + \varepsilon$
- $\varepsilon \sim G(0, \beta)$
- $P(y = 1) = P(U_1 > U_2) = P(V_1 + \varepsilon_1 > V_2 + \varepsilon_2)$

  $= P(\varepsilon_2 - \varepsilon_1 < V_{12}) = P(\varepsilon_{21} < V_{12})$
- 两个等方差且独立的Gumbel随机变量差ε_{21} ~ Logistic(0, β)
- $P(y = 1) = P(\varepsilon_{21} < V_{12}) = \dfrac{1}{1+\exp\left(\dfrac{-V_{12}}{\beta}\right)}$
- β按比例调节V_1和V_2的值，因此可以标准化为1

二元logit模型

- $P(y = 1) = \frac{1}{1+\exp(-V_{12})} = \frac{1}{1+\exp(V_2 - V_1)} = \frac{\exp(V_1)}{\exp(V_1)+\exp(V_2)}$
- $P(y = 0) = 1 - P(y = 1) = \frac{\exp(V_2)}{\exp(V_1)+\exp(V_2)}$
- 这类模型叫做二元logit模型(Binary Logit Model)

效用函数的参数化

- 解释变量的线性组合1：$U_1 = \alpha_0 + X\beta + \varepsilon_1$

$$U_2 = \theta_0 + Z\theta + \varepsilon_2$$

$$P(U_1 > U_2) = P(\beta_0 + X\beta + \varepsilon_1 > \theta_0 + Z\theta + \varepsilon_2) = P[\varepsilon_{21} < (\alpha_0 - \theta_0) + X\beta - Z\theta]$$

- 解释变量的线性组合2：$U_1 = \alpha_0 - \theta_0 + X\beta - Z\theta + \varepsilon_1 - \varepsilon_2$

$$U_2 = 0$$

$$P(U_1 > U_2) = P[\varepsilon_{21} < (\alpha_0 - \theta_0) + X\beta - Z\theta]$$

- 线性组合1和2等效

二元选择模型效用函数的标准化

- 根据线性组合2，令$\beta_0 = \alpha_0 - \theta_0$，$\varepsilon = \varepsilon_1 - \varepsilon_2$，向量Z中的解释变量看作是X向量的一部分
- $U_1 = \beta_0 + X\beta + \varepsilon = V + \varepsilon$，$U_2 = 0$
- $P(y = 1) = P(U_1 > 0) = P(\varepsilon > -V) = 1 - P(\varepsilon < -V)$
- 如果$\varepsilon \sim N(0, 1)$, $P(y = 1) = 1 - \Phi(-V) = \Phi(V)$;

  $P(y = 0) = \Phi(-V) = 1 - \Phi(V)$
- 如果 $\varepsilon \sim$ Logistic(0, 1), $P(y=1) = \dfrac{\exp(V)}{1+\exp(V)}$

$$P(y=0) = \frac{1}{1+\exp(V)}$$

Logistic回归模型

- $U = \beta_0 + X\beta + \varepsilon = V + \varepsilon$
- 如果$U > 0$, $y = 1$; 否则，$y = 0$
- 假设 $\varepsilon \sim Logistic(0, 1)$, $P(\varepsilon < x) = 1/[1 + \exp(-x)]$
- $P(y = 1) = P(U > 0) = P(V + \varepsilon > 0) = P(\varepsilon > -V) = 1 - P(\varepsilon < -V)$

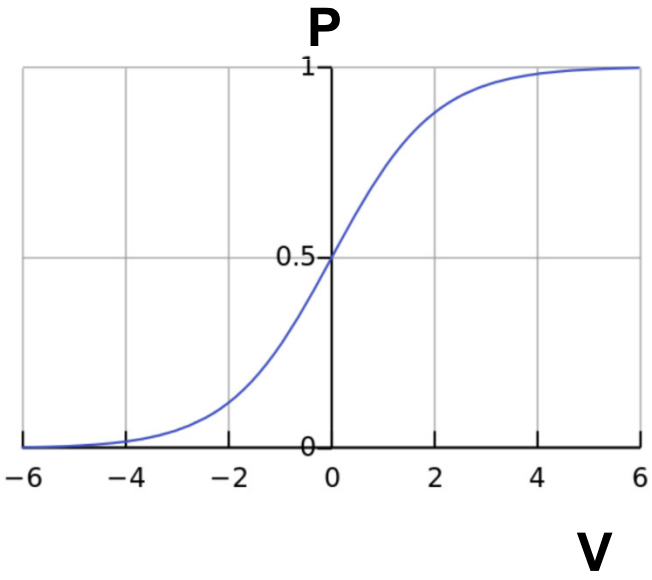
$$P(y=1)=\frac{\exp(V)}{1+\exp(V)} \qquad P(y=0)=\frac{1}{1+\exp(V)}$$

- 这里$V = \beta_0 + X\beta$，类似于线性回归模型，因此这种模型又叫logistic回归模型，X向量中的变量也可以叫做解释变量

Logistic回归模型中的概率函数特性

- 当V值倾近于+∞, P(y = 1) = 1; V远大于参考值0，选项1优势明显，必然被选择
- 当V值倾近于-∞, P(y = 1) = 0；V远小于参考值0，选项1劣势明显，必然不被选择
- 当V值恰好等于0, P(y = 1) = 0.5；V等于参考值0，选项1和2具有相同优势，被选中的概率均为0.5

$$P(y=1)=\frac{\exp(V)}{1+\exp(V)}$$



二元logit模型的数值例子 [1]

- 如果选择集当中有两种通勤方式：开车和乘坐地铁
- 选项的属性和相应的权重系数如下表所示
- 根据二元logit模型计算选择开车和乘坐地铁的概率

属性	权重	开车	地铁
步行时间 （分种）	-0.02	5	15
等候时间 （分种）	-0.02	--	5
车内时间 （分种）	-0.01	35	50
汽油费 （元）	-0.05	10	--
停车费 （元）	-0.05	5	--
票价 （元）	-0.05	--	5

二元logit模型的数值例子 [2]

属性	权重	开车	开车效用值	地铁	地铁效用值
步行时间 （分种）	-0.02	5	-0.10	15	-0.30
等候时间 （分种）	-0.02	--	0	5	-0.10
车内时间 （分种）	-0.01	35	-0.35	50	-0.50
汽油费 （元）	-0.05	10	-0.50	--	0
停车费 （元）	-0.05	5	-0.25	--	0
票价 （元）	-0.05	--	0	5	-0.25
总和			-1.2		-1.15

如果y = 1为开车， y = 0为乘坐地铁

$$P(y=1)=\frac{\exp(V_1)}{\exp(V_1)+\exp(V_2)}=\frac{\exp(-1.2)}{\exp(-1.2)+\exp(-1.15)}=0.4875$$

$$P(y=0)=\frac{\exp(V_2)}{\exp(V_1)+\exp(V_2)}=\frac{\exp(-1.15)}{\exp(-1.2)+\exp(-1.15)}=0.5125$$

同濟大學
TONGJI UNIVERSITY

谢谢大家！

同济大学交通运输工程学院
SCHOOL OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY