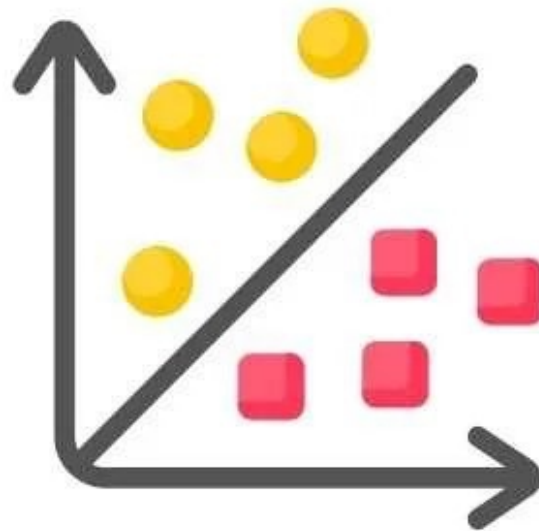




Regression



Classification



Logistic回归

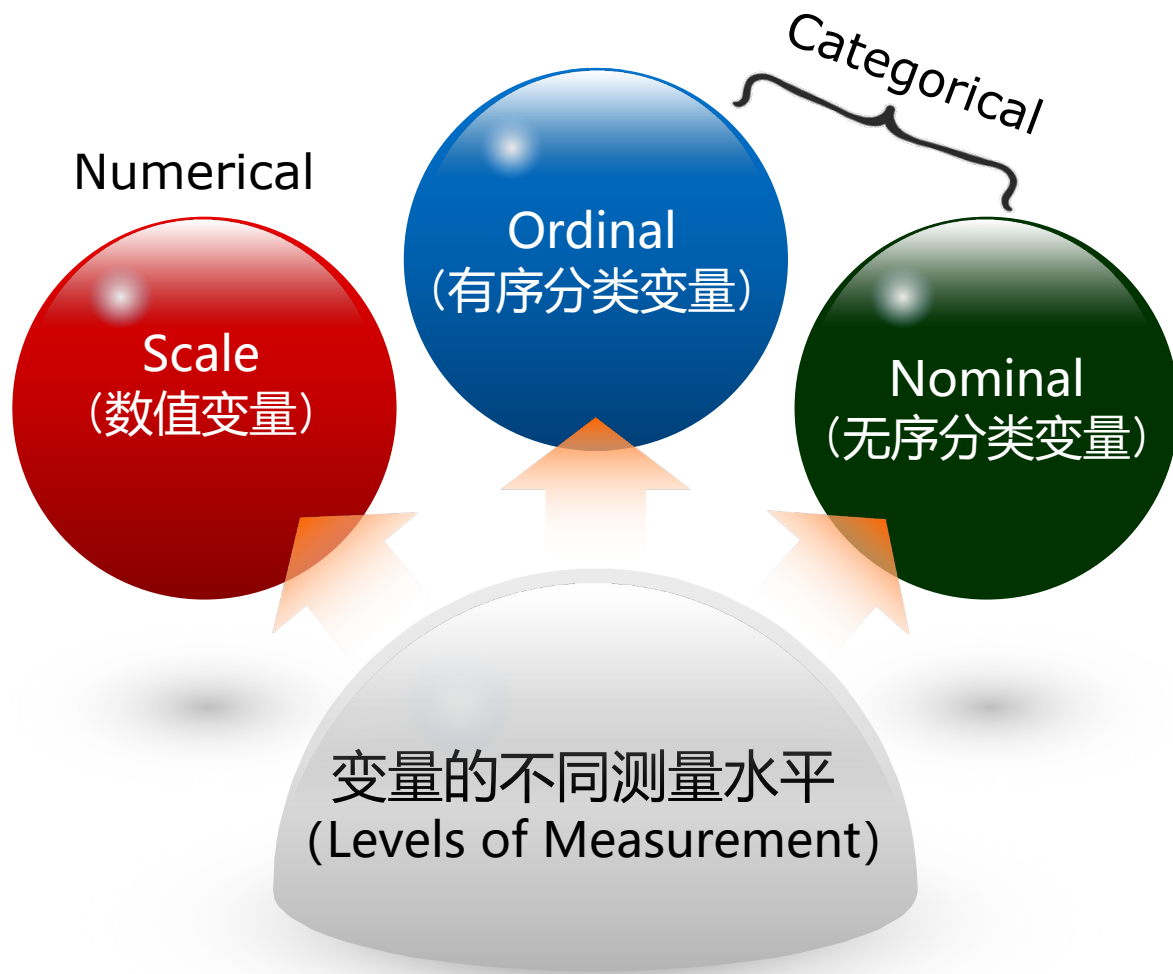
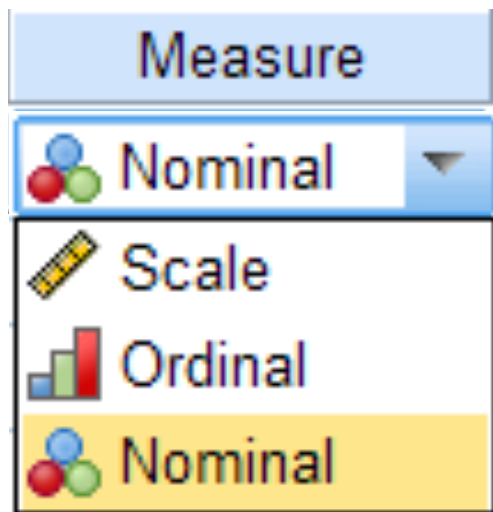
城市分析方法系列课程

苏州大学 王灿

大纲

- 不同类型的数据
- Logistic回归的基本原理
- 二元Logistic回归
- 多元Logistic回归

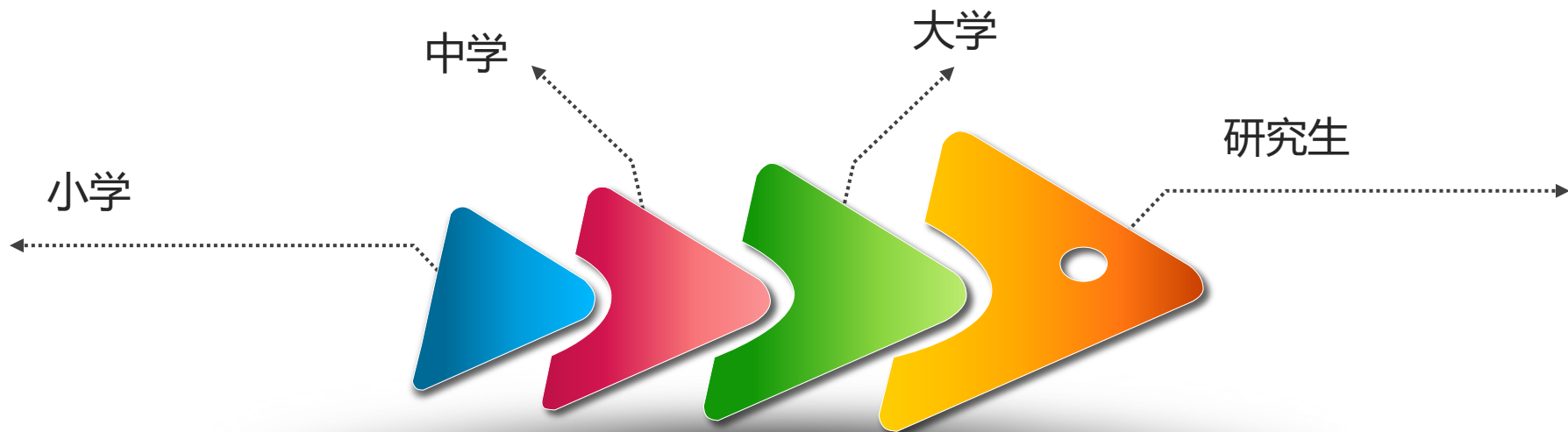
不同类型的变量



不同类型的变量

分类变量 Categorical

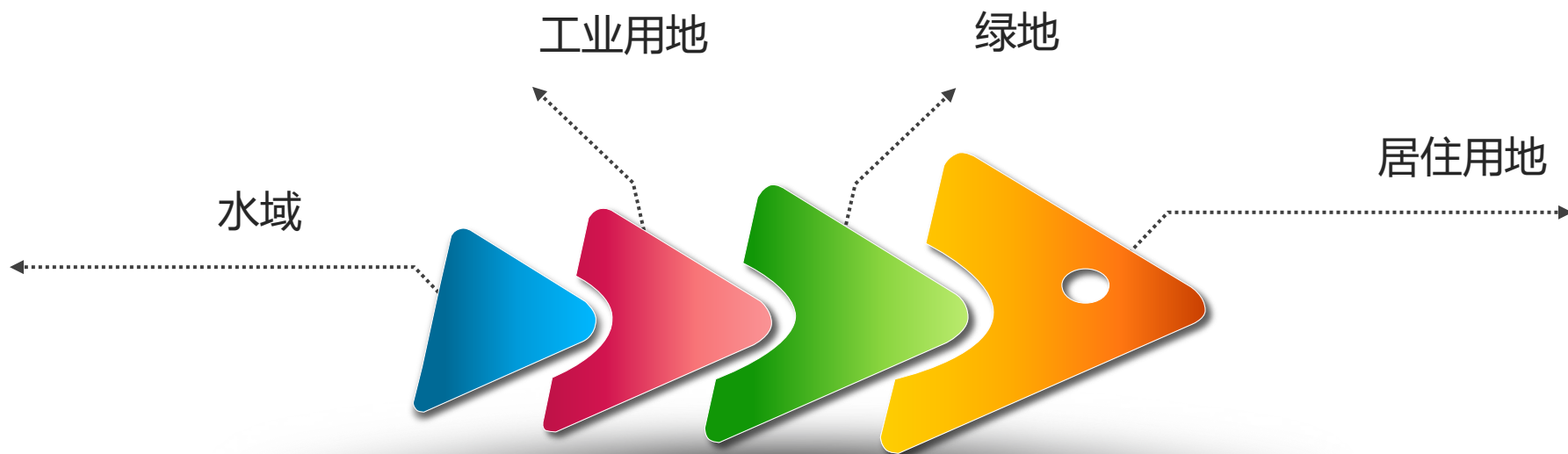
- 受教育程度是一个有序分类变量 (ordinal)



不同类型的变量

分类变量 Categorical

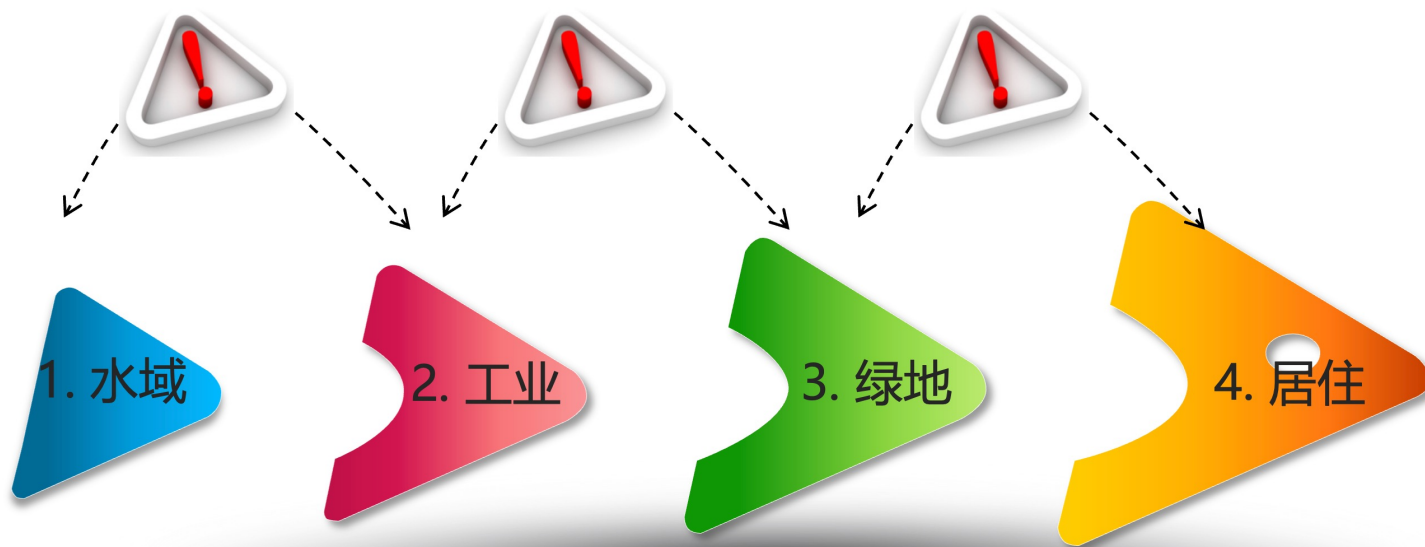
- 受教育程度是一个有序分类变量 (ordinal)
- 用地性质是一个无序分类变量 (nominal)



不同类型的变量

分类变量 Categorical

- 数字编码没有数值意义：1→2, 2→3, 3→4的效应可以完全不同。



分类变量作为**自变量**，怎么办？

- 把分类变转化为0/1虚拟变量

分类变量作为**因变量**，怎么办？

- Logistic回归
- 判别分析
- 对数线性模型
- 分类机器学习

大纲

- 不同类型的数据
- **Logistic回归的基本原理**
- 二元Logistic回归
- 多元Logistic回归

Logistic回归的基本原理

What?

- **逻辑回归**（又称**逻辑斯蒂回归**，**Logistic Regression**）用于被解释的因变量为二分类或多分类的问题，可视为普通线性回归在分类因变量上的推广。
- **二元逻辑回归**（Binary Logistic Regression）：二分类因变量
 - 城市化进程中，农民工是否落户？
 - 城市非建设用地是否转化为建设用地？
- **多元逻辑回归**（Multinomial Logistic Regression）：多分类因变量
 - 企业选址在哪一类区位（市区/近郊/远郊）？
 - 地块更新为哪一种用地性质（R类/A类/B类/G类/.....）？

Logistic回归的基本原理

How?

- 被解释变量：分类变量的**类别概率** (probability, P)
- 概率P与解释变量的关系：

- 可以直接线性回归吗?

$$P = a + \sum bx$$

Logistic回归的基本原理

How?

- 被解释变量：分类变量的**类别概率** (probability, P)

- 概率 P 与解释变量的关系：

- 可以直接线性回归吗？



$$P = a + \sum bx$$

- 不可以：因为可能得到无意义的结果。

$$0 \leq P \leq 1$$

Logistic回归的基本原理

How?

- 被解释变量：分类变量的**类别概率** (probability, P)
- 概率 P 与解释变量的关系：

- 可以直接线性回归吗?



$$P = a + \sum bx$$

- 不可以：因为可能得到无意义的结果。

$$0 \leq P \leq 1$$

$$\frac{P}{1-P} : [0, +\infty)$$

Logistic回归的基本原理

How?

- 被解释变量：分类变量的**类别概率** (probability, P)
- 概率 P 与解释变量的关系：

- 可以直接线性回归吗？



$$P = a + \sum bx$$

- 不可以：因为可能得到无意义的结果。

$$0 \leq P \leq 1$$

$$\ln\left(\frac{P}{1-P}\right) : (-\infty, +\infty)$$

Logistic回归的基本原理

How?

- 被解释变量：分类变量的**类别概率** (probability, P)
- 概率P与解释变量的关系：

- 可以直接线性回归吗？



$$P = a + \sum bx$$

- 不可以：因为可能得到无意义的结果。

$$0 \leq P \leq 1$$

$$\ln\left(\frac{P}{1-P}\right) = a + \sum bx$$

Logistic回归的基本原理

How?

- 被解释变量：分类变量的**类别概率** (probability, P)

- 概率P与解释变量的关系：

- 可以直接线性回归吗？



$$P = a + \sum bx$$

- 不可以：因为可能得到无意义的结果。

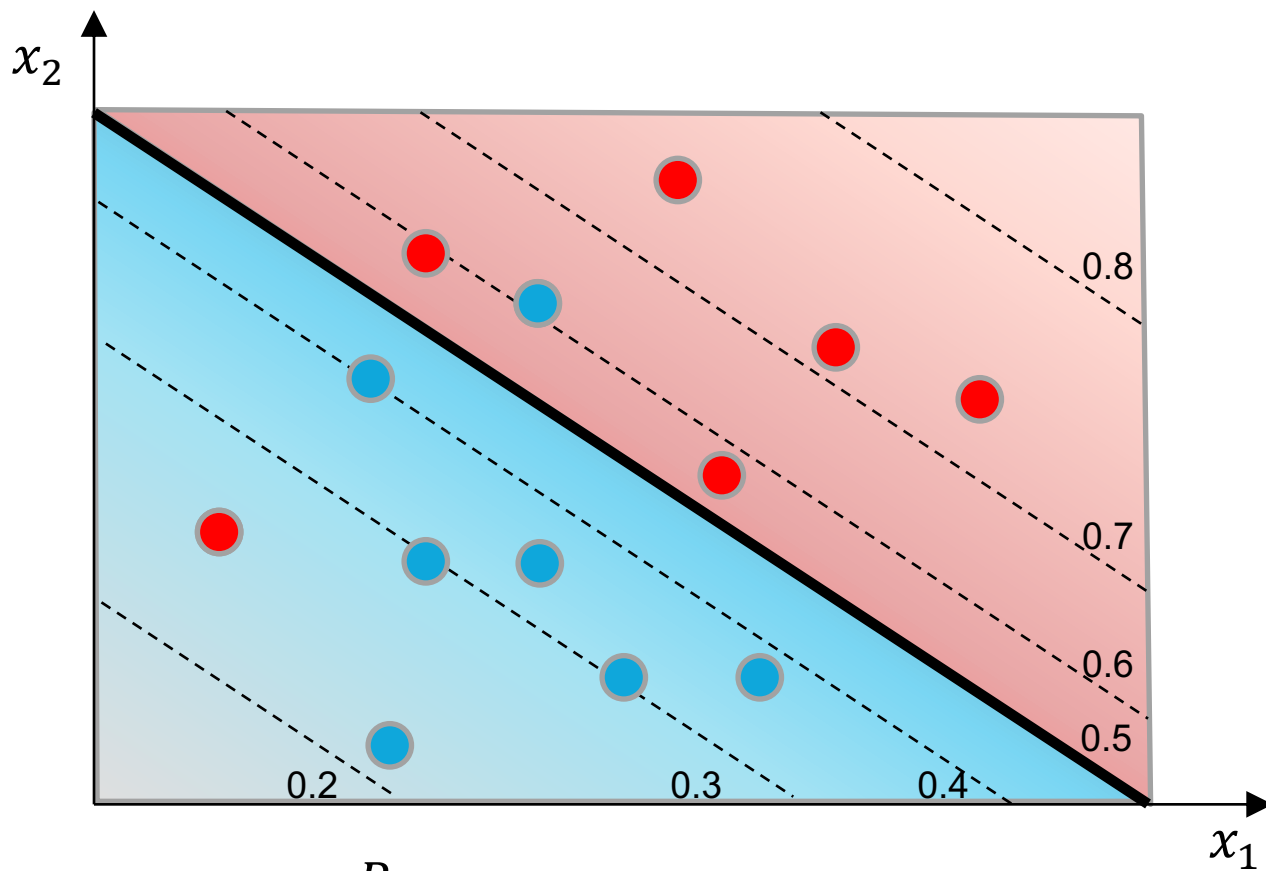
$$0 \leq P \leq 1$$

- Logit变换后，再线性回归。



$$\text{Logit}(P) = \ln\left(\frac{P}{1-P}\right) = a + \sum bx$$

Logistic回归的基本原理



$$\ln \frac{P}{1-P} = y = a + b_1 x_1 + b_2 x_2$$

Logistic回归的基本原理

How?

互逆

• **Logit变换：建模阶段**

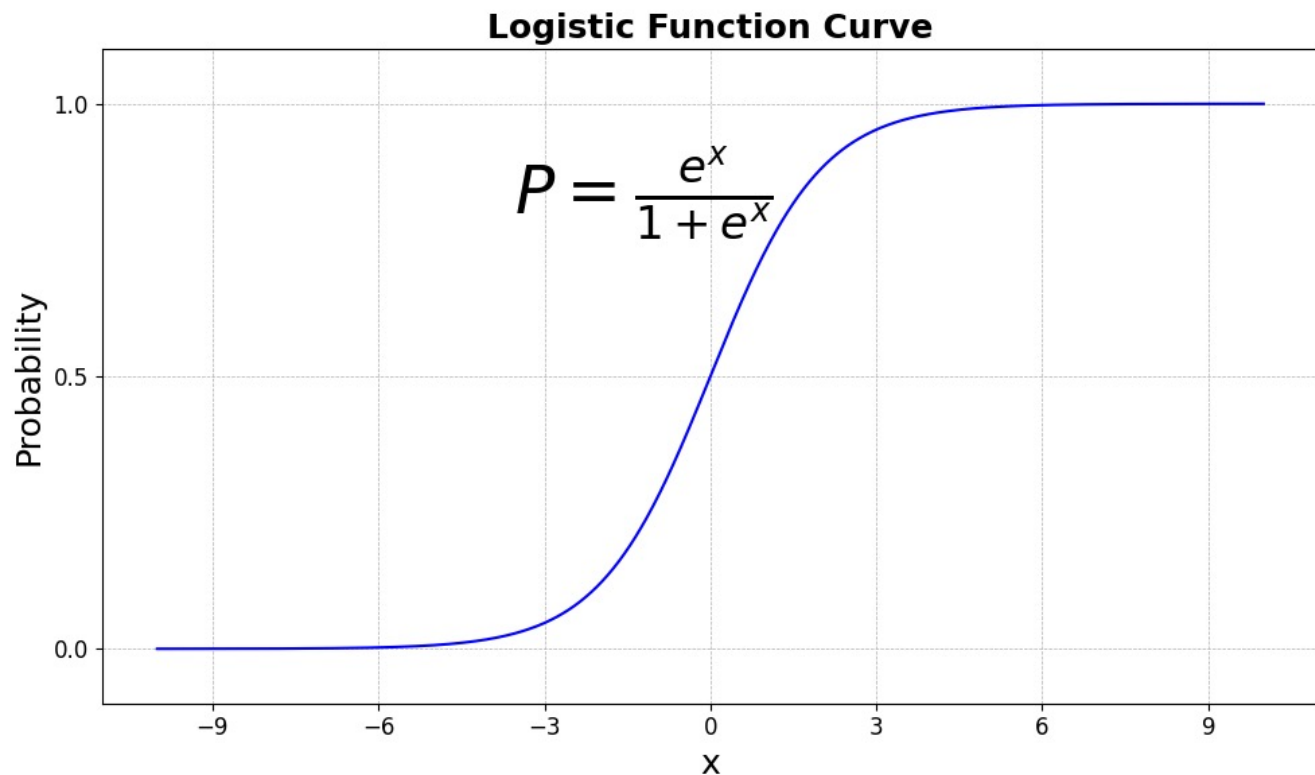
$$\text{Logit}(P) = \ln\left(\frac{P}{1-P}\right) = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

• **Logistic变换：预测阶段**

$$P = \frac{e^{\text{Logit}(P)}}{1 + e^{\text{Logit}(P)}} = \frac{e^{a+b_1x_1+b_2x_2+\cdots+b_kx_k}}{1 + e^{a+b_1x_1+b_2x_2+\cdots+b_kx_k}}$$

Logistic回归的基本原理

概率与自变量的非线性关系



S形曲线

中段附近变化最快

当事件的发生与否已经很确定时，概率对影响要素的变化不敏感。

在两端变化最慢

当不确定性很强时，概率对影响要素的变化很敏感。

Logistic回归的基本原理

Odds

- odds=发生与不发生的概率之比，常译为优势、比值

$$\text{Logit}(P) = \ln\left(\frac{P}{1-P}\right) = \ln(\text{odds})$$

- odds=一种结果的概率相对于另一种结果的概率之比

$$\ln(\text{odds}^{1,Ref}) = \ln\left(\frac{\text{Prob}(y=1)}{\text{Prob}(y^{Ref}=0)}\right) \quad \text{二元Logistic回归}$$

$$\ln(\text{odds}^{3,Ref}) = \ln\left(\frac{\text{Prob}(y=3)}{\text{Prob}(y^{Ref}=4)}\right) \quad \begin{array}{l} \text{多项Logistic回归} \\ \text{(以} y=4 \text{为参照)} \end{array}$$

Logistic回归的基本原理

回归系数的解释

- **线性回归**: $y = a + bx$

x 每增加1个单位, y 将增加 b 个单位。

- **Logistic回归**: $\ln(\text{odds}) = a + bx$

x 每增加1个单位, $\ln(\text{odds})$ 将增加 b 个单位。

x 每增加1个单位, odds 将变为原来的 e^b 倍。

$$b = \ln[\text{odds}(x + 1)] - \ln[\text{odds}(x)] \quad \Rightarrow \quad e^b = \frac{\text{odds}(x + 1)}{\text{odds}(x)} = \text{OR}$$

Logistic回归的基本原理

回归系数的解释

- **Odds ratio (OR)**: 优势比, 比值比, Logistic回归常用的系数解释方式

$$OR = e^b = \frac{\text{odds}(x + 1)}{\text{odds}(x)}$$

线性回归: **b**

Logistic回归: **exp(b)**

判断

b=0

exp(b)=1

无效应

b<0

exp(b)<1

负效应

exp(b)=0.4

odds降至原来的40%

b>0

exp(b)>1

正效应

exp(b)=2

odds升至原来的2倍

大纲

- 不同类型的数据
- Logistic回归的基本原理
- **二元Logistic回归**
(案例：毕业生留苏意愿)
- 多元Logistic回归

二元Logistic回归

数据:

苏州市某高校30位大学毕业生的**性别**、**学习成绩**、**户籍地**以及**毕业去向**(是否愿意留在苏州)。

问题:

通过Logistic回归模型定量分析各解释变量对毕业去向的影响。

二元Logistic回归

样本编号	Gender	Record	Region	Stay
(1)	(2)	(3)	(4)	(5)
1	女-0	较好-1	东北-1	否-0
2	女-0	较好-1	东北-1	否-0
3	女-0	一般-0	西部-2	否-0
4	男-1	一般-0	中部-4	否-0
5	男-1	一般-0	西部-2	否-0
6	男-1	一般-0	西部-2	否-0
7	女-0	一般-0	西部-2	是-1
8	女-0	较好-1	东部-3	是-1
⋮	⋮	⋮	⋮	⋮
26	男-1	一般-0	中部-4	否-0
27	女-0	较好-1	东北-1	是-1
28	女-0	较好-1	中部-4	否-0
29	男-1	较好-1	东部-3	是-1
30	女-0	一般-0	东北-1	否-0

二元Logistic回归

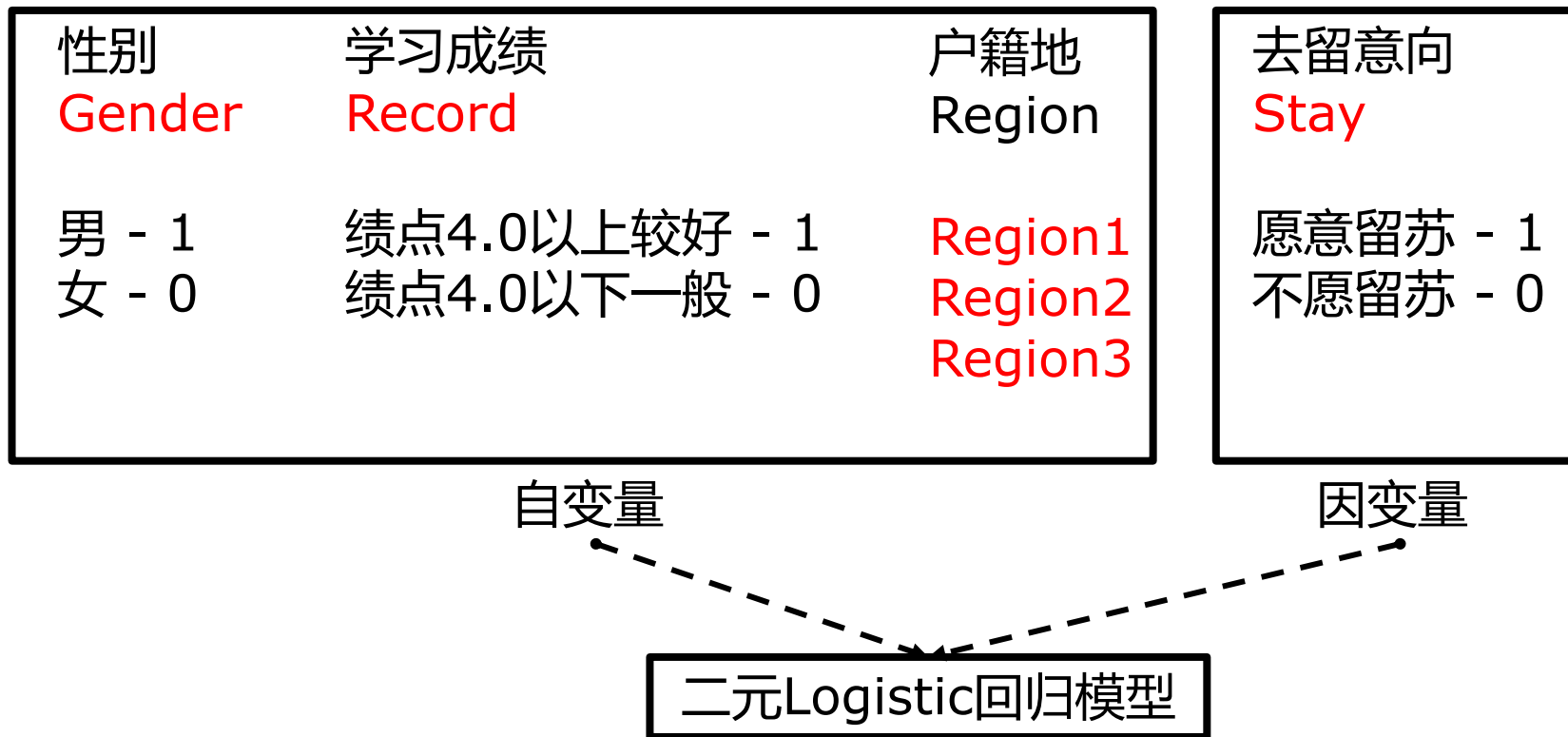
性别 Gender	学习成绩 Record	户籍地 Region	去留意向 Stay
男 - 1 女 - 0	绩点4.0以上较好 - 1 绩点4.0以下一般 - 0	东北 - 1 西部 - 2 东部 - 3 中部 - 4	愿意留苏 - 1 不愿留苏 - 0

取值不具有数值意义
应设虚拟变量

以户籍地在 “中部” 为参照设3个虚拟变量

		人数	虚拟变量设置		
			Region1	Region2	Region3
户籍地 (Region)	东北	6	1	0	0
	西部	8	0	1	0
	东部	7	0	0	1
	中部	9	0	0	0

二元Logistic回归



二元Logistic回归

$$\begin{aligned}\text{Logit}(P) &= \ln(\text{odds}) = \ln\left(\frac{P}{1-P}\right) \\ &= a + b_{\text{Gender}}\text{Gender} + b_{\text{Record}}\text{Record} + b_{\text{Region1}}\text{Region1} \\ &\quad + b_{\text{Region2}}\text{Region2} + b_{\text{Region3}}\text{Region3}\end{aligned}$$

式中, $P = \text{Prob}(\text{Stay} = 1)$, 是毕业生愿意留苏的概率。

二元Logistic回归

样本量(N): 30
对数似然值(LL): -16.073
Nagelkerke r^2 : 0.356

变量个数(K): 6
Cox & Snell r^2 : 0.267
McFadden r^2 : 0.225

变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
Gender	1.466	1.017	2.075	1	.150	4.330
Record	2.067	1.123	3.387	1	.066	7.902
Region1	.393	1.382	.081	1	.776	1.481
Region2	1.072	1.167	.844	1	.358	2.921
Region3	2.323	1.278	3.304	1	.069	10.208
常数项	-2.663	1.176	5.129	1	.024	.070

- 线性回归的 R^2 : 因变量的变异有百分之多少被解释?
- Logistic回归的 R^2 : 所估计的模型相比于“零模型”, 表现提升了多少?
 - 没有绝对的可接受标准, 通常比线性回归低得多, 大概在0.1 – 0.5之间。

二元Logistic回归

样本量(N): 30
对数似然值(LL): -16.073
Nagelkerke r^2 : 0.356

变量个数(K): 6
Cox & Snell r^2 : 0.267
McFadden r^2 : 0.225

变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
Gender	1.466	1.017	2.075	1	.150	4.330
Record	2.067	1.123	3.387	1	.066	7.902
Region1	.393	1.382	.081	1	.776	1.481
Region2	1.072	1.167	.844	1	.358	2.921
Region3	2.323	1.278	3.304	1	.069	10.208
常数项	-2.663	1.176	5.129	1	.024	.070

建立模型:

$$\text{Logit}(P) = -2.663 + 1.466\text{Gender} + 2.067\text{Record} \\ + 0.393\text{Region1} + 1.072\text{Region2} + 2.323\text{Region3}$$

二元Logistic回归

预测

愿意留苏的概率:

$$P = \frac{e^{\text{Logit}(P)}}{1 + e^{\text{Logit}(P)}} = \frac{e^{a+b_1x_1+b_2x_2+\cdots+b_kx_k}}{1 + e^{a+b_1x_1+b_2x_2+\cdots+b_kx_k}}$$

二元Logistic回归

预测

愿意留苏的概率：

$$P = \frac{e^{-2.663 + 1.466Gender + 2.067Record + 0.393Region1 + 1.072Region2 + 2.323Region3}}{1 + e^{-2.663 + 1.466Gender + 2.067Record + 0.393Region1 + 1.072Region2 + 2.323Region3}}$$

“学习成绩较好、户籍地为东部地区的男生” 愿意留苏的概率？

Record=1

Region1=0

Gender=1

Region2=0

Region3=1

$$P = \frac{e^{-2.663 + 1.466 + 2.067 + 2.323}}{1 + e^{-2.663 + 1.466 + 2.067 + 2.323}} = 96.1\% \quad \text{愿意留苏的概率为96.1\%}$$

二元Logistic回归

解释

变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
Gender	1.466	1.017	2.075	1	.150	4.330
Record	2.067	1.123	3.387	1	.066	7.902
Region1	.393	1.382	.081	1	.776	1.481
Region2	1.072	1.167	.844	1	.358	2.921
Region3	2.323	1.278	3.304	1	.069	10.208
常数项	-2.663	1.176	5.129	1	.024	.070

男生与女生之间愿意留苏的概率之比为4.33？

$$\frac{\text{Prob}(\text{Stay} = 1 | \text{Gender} = 1)}{\text{Prob}(\text{Stay} = 1 | \text{Gender} = 0)} = 4.330$$

二元Logistic回归

解释

变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
Gender	1.466	1.017	2.075	1	.150	4.330
Record	2.067	1.123	3.387	1	.066	7.902
Region1	.393	1.382	.081	1	.776	1.481
Region2	1.072	1.167	.844	1	.358	2.921
Region3	2.323	1.278	3.304	1	.069	10.208
常数项	-2.663	1.176	5.129	1	.024	.070

男生与女生之间愿意留苏的概率之比为4.33?

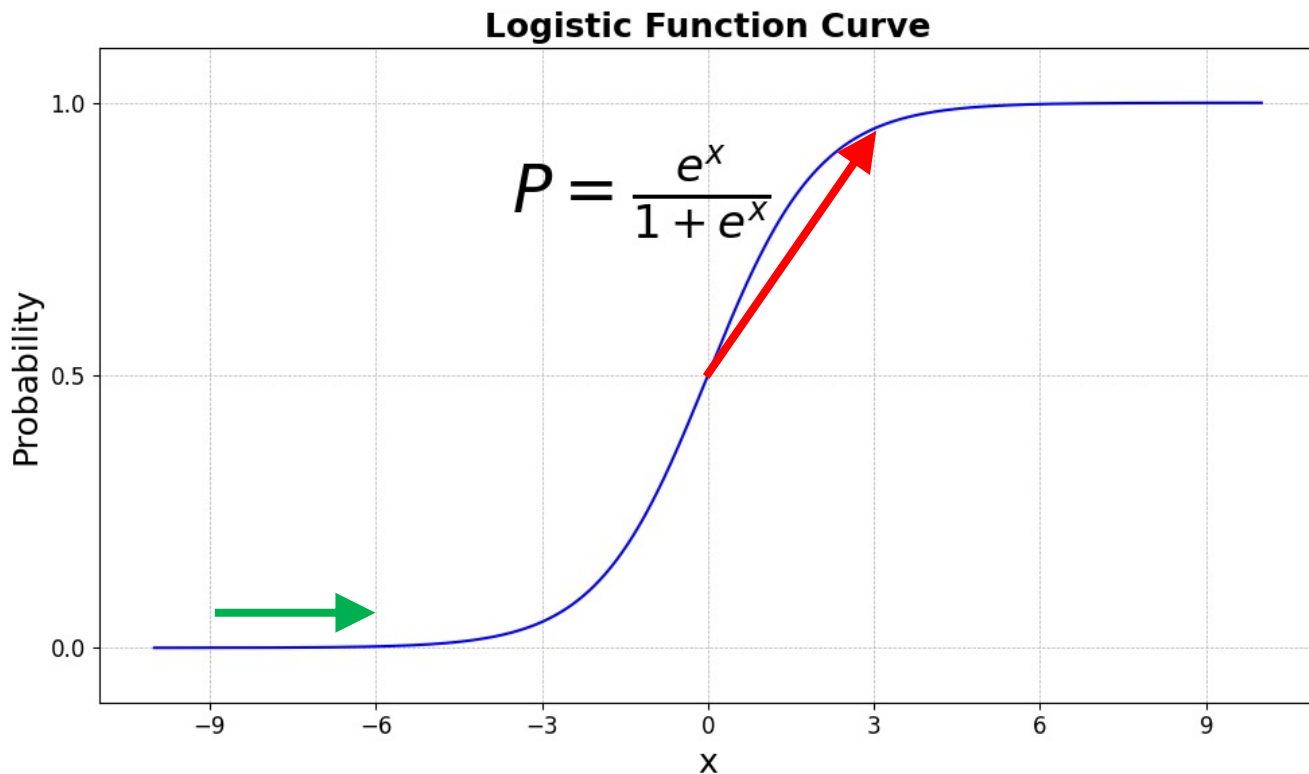


$$\frac{\text{Prob}(\text{Stay} = 1 | \text{Gender} = 1)}{\text{Prob}(\text{Stay} = 1 | \text{Gender} = 0)} \neq 4.330$$

这种看似直白的解释
是初学者的常见错误!

二元Logistic回归

解释：为什么不对？

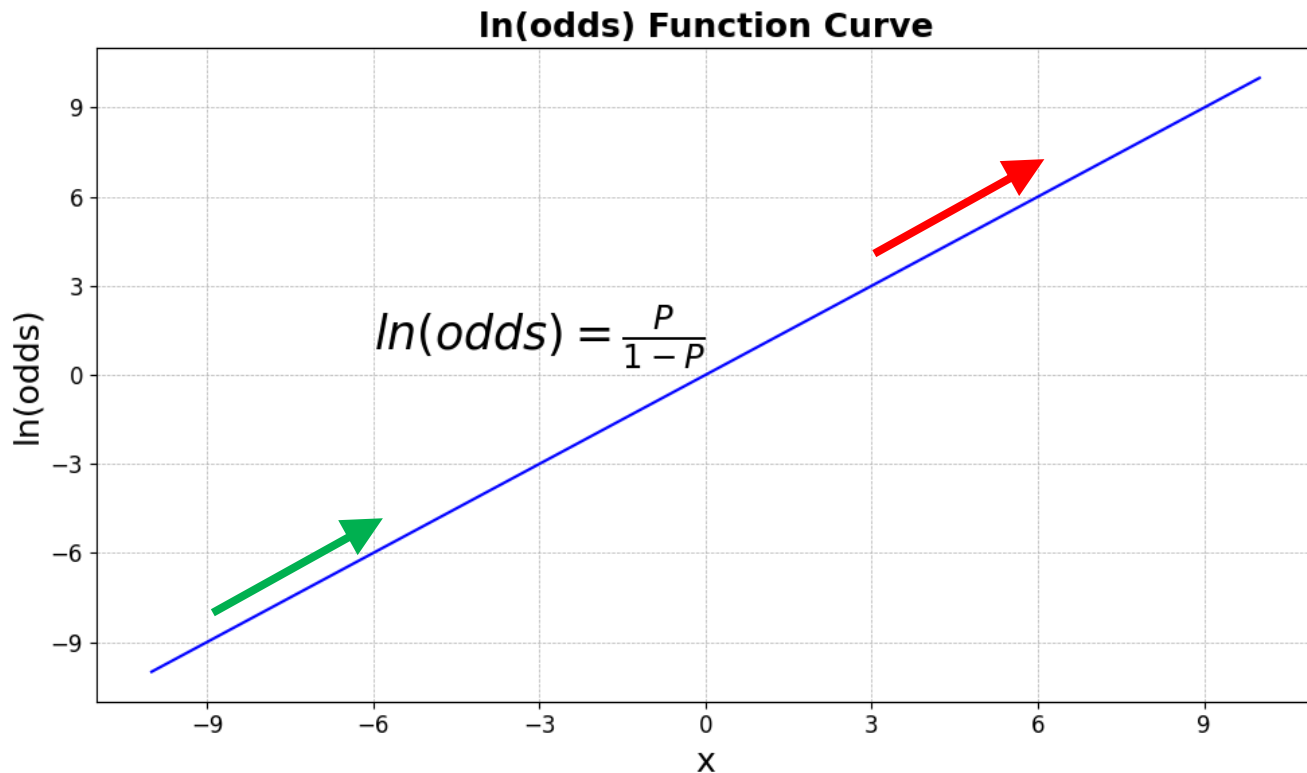


自变量发生变化时

- 事件发生的概率会根据当前水平的不同而发生不同比例的变化。

二元Logistic回归

解释：为什么不对？



自变量发生变化时

- 事件发生的概率会根据当前水平的不同而发生不同比例的变化。
- 但odds的变化比例永远是固定的。

二元Logistic回归

解释

$$\begin{aligned} OR &= \frac{\text{odds}(Gender = 1)}{\text{odds}(Gender = 0)} \\ &= \frac{\text{Prob}(Stay = 1|Gender = 1)/\text{Prob}(Stay = 0|Gender = 1)}{\text{Prob}(Stay = 1|Gender = 0)/\text{Prob}(Stay = 0|Gender = 0)} \\ &= 4.330 \end{aligned}$$

- Gender的exp(B)为4.330, 由于 $4.330 > 1$, 所以男生 (Gender=1) 与女生 (Gender=0) 相比具有更强的留苏意愿, 前者愿意留苏的odds是后者的4.33倍。

二元Logistic回归

解释

变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
Gender	1.466	1.017	2.075	1	.150	4.330
Record	2.067	1.123	3.387	1	.066	7.902
Region1	.393	1.382	.081	1	.776	1.481
Region2	1.072	1.167	.844	1	.358	2.921
Region3	2.323	1.278	3.304	1	.069	10.208
常数项	-2.663	1.176	5.129	1	.024	.070

二元Logistic回归

解释

- Record的 $\exp(B)$ 为7.902, 由于 $7.902 > 1$, 所以成绩较好 (Record=1) 的毕业生相比于成绩一般 (Record=0) 的毕业生, 具有更强的留苏意愿, 前者留苏的odds是后者的7.902倍。
- Region1、Region2、Region3的 $\exp(B)$ 分别为1.481、2.921、10.208, 均大于1, 因此与户籍地在中部地区 (参照水平) 的毕业生相比, 户籍地在东北、西部、东部地区的毕业生均具有更强的留苏意愿, 其留苏的odds分别是中部地区毕业生的1.481、2.921、10.208倍。

二元Logistic回归

解释：效应大小比较

变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
Gender	1.466	1.017	2.075	1	.150	4.330
Record	2.067	1.123	3.387	1	.066	7.902
Region1	.393	1.382	.081	1	.776	1.481
Region2	1.072	1.167	.844	1	.358	2.921
Region3	2.323	1.278	3.304	1	.069	10.208
常数项	-2.663	1.176	5.129	1	.024	.070

- 由于所有的自变量都是0/1变量，因此它们的系数可以直接比较。各因素中，户籍地是否是东部地区对留苏意愿的影响最大，其次为学习成绩。

二元Logistic回归

解释：效应的显著性

变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
Gender	1.466	1.017	2.075	1	.150	4.330
Record	2.067	1.123	3.387	1	.066	7.902
Region1	.393	1.382	.081	1	.776	1.481
Region2	1.072	1.167	.844	1	.358	2.921
Region3	2.323	1.278	3.304	1	.069	10.208
常数项	-2.663	1.176	5.129	1	.024	.070

估计值 (效应大小)

二元Logistic回归

解释：效应的显著性

变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
Gender	1.466	1.017	2.075	1	.150	4.330
Record	2.067	1.123	3.387	1	.066	7.902
Region1	.393	1.382	.081	1	.776	1.481
Region2	1.072	1.167	.844	1	.358	2.921
Region3	2.323	1.278	3.304	1	.069	10.208
常数项	-2.663	1.176	5.129	1	.024	.070

标准误（估计误差，不确定性）

二元Logistic回归

解释：效应的显著性

变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
Gender	1.466	1.017	2.075	1	.150	4.330
Record	2.067	1.123	3.387	1	.066	7.902
Region1	.393	1.382	.081	1	.776	1.481
Region2	1.072	1.167	.844	1	.358	2.921
Region3	2.323	1.278	3.304	1	.069	10.208
常数项	-2.663	1.176	5.129	1	.024	.070

Wald值 $(= (B/S.E.)^2$, 显著性)

P值 (显著性)

二元Logistic回归

解释：效应的显著性

变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
Gender	1.466	1.017	2.075	1	.150	4.330
Record	2.067	1.123	3.387	1	.066	7.902
Region1	.393	1.382	.081	1	.776	1.481
Region2	1.072	1.167	.844	1	.358	2.921
Region3	2.323	1.278	3.304	1	.069	10.208
常数项	-2.663	1.176	5.129	1	.024	.070

- 本案例的样本量偏小，在 $\alpha = 0.1$ 的相对宽松的显著性水平下，只有学习成绩（Record）和户籍地是否属于东部地区（Region3）两个变量显著。

大纲

- 不同类型的数据
- Logistic回归的基本原理
- 二元Logistic回归
- **多元Logistic回归**
(案例：工业用地的更新)

多元Logistic回归

数据：

收集了上海市虹口区 113 块工业用地更新的数据，包括**企业名称**、更新后的**用地功能**、**用地面积**、**建筑面积**、用地的**不规则系数**、**区位**、到虹口区中心的**直线距离**、地块**相邻道路的等级**、**是否为国有企业**。

问题：

- 1) 分析工业用地的功能更新受到了哪些要素的影响？
- 2) 什么样的条件使之更新为特定功能的可能性更大？

多元Logistic回归

[illegible]

多元Logistic回归

企业名称
Name

更新后的用地功能
New Function

创意产业园 - 1
商务办公 - 2
商业服务 - 3
生产物流 - 4

因变量

用地面积
Lot Area

(单位: ha)

建筑面积
Floor Area

(单位: 万m²)

自变量

用地的不规则系数
Irregular Coefficient

$$(\text{= } \frac{\text{用地的周长}}{\sqrt{\text{用地的面积}}})$$

IC 值越小, 则用地越接近正圆形, 可认为用地形状越规则

区位
Location

内环内 - 1
内中环间 - 2
中外环间 - 3

到虹口区3个中心的距离
Distance

(单位: km)

地块相邻道路的等级
Road

干路 - 1
支路 - 0

是否为国有企业
State Owned

国有企业 - 1
非国有企业 - 0

多元Logistic回归

更新后的用地功能

New Function

创意产业园 - 1

商务办公 - 2

商业服务 - 3

生产物流 - 4 作为参照水平

四分类变量

多项 Logistic 回归需要从因变量的

多个类别中选择一个作为参照水平

即, $NF^{Ref} = 4$

区位

Location

三分类变量

内环内 - 1

内中环间 - 2

中外环间 - 3 作为参照水平

设置虚拟变量



Loc1 是否在内环内

Loc2 是否在内中环间

	Loc1	Loc2
内环内	1	0
内中环间	0	1
中外环间	0	0

多元Logistic回归

企业名称
Name

更新后的用地功能
New Function

创意产业园 - 1
商务办公 - 2
商业服务 - 3
生产物流 - 4

用地面积
Lot Area

(单位: ha)

建筑面积
Floor Area

(单位: 万m²)

用地的不规则系数
Irregular Coefficient

$$(\text{= } \frac{\text{用地的周长}}{\sqrt{\text{用地的面积}}})$$

IC 值越小, 则用地越接近正圆形, 可认为用地形状越规则

因变量

自变量

多元Logistic回归模型

区位
Location

Loc1
Loc2

到虹口区3个中心的距离
Distance

(单位: km)

地块相邻道路的等级
Road

干路 - 1
支路 - 0

是否为国有企业
State Owned

国有企业 - 1
非国有企业 - 0

Logistic回归的基本原理

Odds

- odds=发生与不发生的概率之比，常译为优势、比值

$$\text{Logit}(P) = \ln\left(\frac{P}{1-P}\right) = \ln(\text{odds})$$

- odds=一种结果的概率相对于另一种结果的概率之比

$$\ln(\text{odds}^{1,Ref}) = \ln\left(\frac{\text{Prob}(y=1)}{\text{Prob}(y^{Ref}=0)}\right) \quad \text{二元Logistic回归}$$

$$\ln(\text{odds}^{3,Ref}) = \ln\left(\frac{\text{Prob}(y=3)}{\text{Prob}(y^{Ref}=4)}\right) \quad \begin{array}{l} \text{多项Logistic回归} \\ \text{(以} y=4 \text{为参照)} \end{array}$$

多元Logistic回归

$$\begin{aligned}\ln(\text{odds}^{k,Ref}) &= \ln\left(\frac{\Pr(NF = k)}{\Pr(NF^{Ref} = 4)}\right) \\ &= a_k + b_{k,LA}LA + b_{k,FA}FA + b_{k,IC}IC + b_{k,Loc1}Loc1 + b_{k,Loc2}Loc2 + b_{k,Dist}Dist \\ &\quad + b_{k,Road}Road + b_{k,SO}SO \\ &\quad k = 1, 2, 3\end{aligned}$$

式中， $\Pr(NF = k)$ 表示更新为第 k 类功能的概率。

$\text{odds}^{k,Ref}$ 是 “更新为第 k 类功能” 相对于 “更新为参照水平（生产物流）” 的，是两个概率之比。

由于自变量较多，在模型估计中采用了**逐步回归**的技术。

在逐步回归以及相应的自变量选择方面，Logistic回归与普通的线性回归相同。

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
创意 产业园 (NF = 1)	常数项	-1.216	1.536	0.627	1	0.429	
	LA	-0.964	1.575	0.375	1	0.540	0.381
	FA	2.931	1.148	6.515	1	0.011	18.743
	Loc1	0.757	1.353	0.313	1	0.576	2.133
	Loc2	-0.012	1.356	<0.001	1	0.993	0.988
	Dist	-2.954	0.929	10.107	1	0.001	0.052
	Road	3.787	1.197	10.019	1	0.002	44.146
商务 办公 (NF = 2)	常数项	1.155	1.331	0.753	1	0.385	
	LA	-5.220	2.250	5.382	1	0.020	0.005
	FA	2.977	1.322	5.067	1	0.024	19.620
	Loc1	0.983	1.030	0.910	1	0.340	2.671
	Loc2	0.216	1.082	0.040	1	0.842	1.241
	Dist	-1.759	0.622	8.004	1	0.005	0.172
	Road	2.703	1.034	6.838	1	0.009	14.930
商业 服务 (NF = 3)	常数项	-0.861	1.224	0.495	1	0.482	
	LA	0.835	1.464	0.326	1	0.568	2.306
	FA	-0.208	1.141	0.033	1	0.855	0.812
	Loc1	1.985	0.937	4.486	1	0.034	7.282
	Loc2	-1.446	1.276	1.284	1	0.257	0.236
	Dist	-0.651	0.503	1.679	1	0.195	0.521
	Road	2.621	0.957	7.496	1	0.006	13.743

因变量参照水平：生产物流(NF = 4)

样本量(N): 113
变量个数(K): 21
对数似然值(LL): -102.739
Nagelkerke r^2 : 0.608
Cox & Snell r^2 : 0.565
McFadden r^2 : 0.314

变量IC、SO 没有进入模型，表明用地不规则系数以及企业是否为国有对于工业用地更新为何种功能的影响没有通过统计检验。

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
创意 产业园 (NF = 1)	常数项	-1.216	1.536	0.627	1	0.429	
	LA	-0.964	1.575	0.375	1	0.540	0.381
	FA	2.931	1.148	6.515	1	0.011	18.743
	Loc1	0.757	1.353	0.313	1	0.576	2.133
	Loc2	-0.012	1.356	<0.001	1	0.993	0.988
	Dist	-2.954	0.929	10.107	1	0.001	0.052
	Road	3.787	1.197	10.019	1	0.002	44.146
商务 办公 (NF = 2)	常数项	1.155	1.331	0.753	1	0.385	
	LA	-5.220	2.250	5.382	1	0.020	0.005
	FA	2.977	1.322	5.067	1	0.024	19.620
	Loc1	0.983	1.030	0.910	1	0.340	2.671
	Loc2	0.216	1.082	0.040	1	0.842	1.241
	Dist	-1.759	0.622	8.004	1	0.005	0.172
	Road	2.703	1.034	6.838	1	0.009	14.930
商业 服务 (NF = 3)	常数项	-0.861	1.224	0.495	1	0.482	
	LA	0.835	1.464	0.326	1	0.568	2.306
	FA	-0.208	1.141	0.033	1	0.855	0.812
	Loc1	1.985	0.937	4.486	1	0.034	7.282
	Loc2	-1.446	1.276	1.284	1	0.257	0.236
	Dist	-0.651	0.503	1.679	1	0.195	0.521
	Road	2.621	0.957	7.496	1	0.006	13.743

因变量参照水平：生产物流(NF = 4)

样本量(N): 113
变量个数(K): 21
对数似然值(LL): -102.739
Nagelkerke r^2 : 0.608
Cox & Snell r^2 : 0.565
McFadden r^2 : 0.314

根据R²值，该模型拟合效果较为理想。

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
创意 产业园 (NF = 1)	常数项	-1.216	1.536	0.627	1	0.429	
	LA	-0.964	1.575	0.375	1	0.540	0.381
	FA	2.931	1.148	6.515	1	0.011	18.743
	Loc1	0.757	1.353	0.313	1	0.576	2.133
	Loc2	-0.012	1.356	<0.001	1	0.993	0.988
	Dist	-2.954	0.929	10.107	1	0.001	0.052
	Road	3.787	1.197	10.019	1	0.002	44.146
商务 办公 (NF = 2)	常数项	1.155	1.331	0.753	1	0.385	
	LA	-5.220	2.250	5.382	1	0.020	0.005
	FA	2.977	1.322	5.067	1	0.024	19.620
	Loc1	0.983	1.030	0.910	1	0.340	2.671
	Loc2	0.216	1.082	0.040	1	0.842	1.241
	Dist	-1.759	0.622	8.004	1	0.005	0.172
	Road	2.703	1.034	6.838	1	0.009	14.930
商业 服务 (NF = 3)	常数项	-0.861	1.224	0.495	1	0.482	
	LA	0.835	1.464	0.326	1	0.568	2.306
	FA	-0.208	1.141	0.033	1	0.855	0.812
	Loc1	1.985	0.937	4.486	1	0.034	7.282
	Loc2	-1.446	1.276	1.284	1	0.257	0.236
	Dist	-0.651	0.503	1.679	1	0.195	0.521
	Road	2.621	0.957	7.496	1	0.006	13.743

因变量参照水平：生产物流(NF = 4)

样本量(N): 113
变量个数(K): 21
对数似然值(LL): -102.739
Nagelkerke r^2 : 0.608
Cox & Snell r^2 : 0.565
McFadden r^2 : 0.314

三组参数，分别是工业用地
更新为
创意产业园(NF = 1)、

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
创意 产业园 (NF = 1)	常数项	-1.216	1.536	0.627	1	0.429	
	LA	-0.964	1.575	0.375	1	0.540	0.381
	FA	2.931	1.148	6.515	1	0.011	18.743
	Loc1	0.757	1.353	0.313	1	0.576	2.133
	Loc2	-0.012	1.356	<0.001	1	0.993	0.988
	Dist	-2.954	0.929	10.107	1	0.001	0.052
	Road	3.787	1.197	10.019	1	0.002	44.146
商务 办公 (NF = 2)	常数项	1.155	1.331	0.753	1	0.385	
	LA	-5.220	2.250	5.382	1	0.020	0.005
	FA	2.977	1.322	5.067	1	0.024	19.620
	Loc1	0.983	1.030	0.910	1	0.340	2.671
	Loc2	0.216	1.082	0.040	1	0.842	1.241
	Dist	-1.759	0.622	8.004	1	0.005	0.172
	Road	2.703	1.034	6.838	1	0.009	14.930
商业 服务 (NF = 3)	常数项	-0.861	1.224	0.495	1	0.482	
	LA	0.835	1.464	0.326	1	0.568	2.306
	FA	-0.208	1.141	0.033	1	0.855	0.812
	Loc1	1.985	0.937	4.486	1	0.034	7.282
	Loc2	-1.446	1.276	1.284	1	0.257	0.236
	Dist	-0.651	0.503	1.679	1	0.195	0.521
	Road	2.621	0.957	7.496	1	0.006	13.743

因变量参照水平：生产物流(NF = 4)

样本量(N): 113
变量个数(K): 21
对数似然值(LL): -102.739
Nagelkerke r^2 : 0.608
Cox & Snell r^2 : 0.565
McFadden r^2 : 0.314

三组参数，分别是工业用地更新为
创意产业园(NF = 1)、
商务办公(NF = 2)、

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
创意 产业园 (NF = 1)	常数项	-1.216	1.536	0.627	1	0.429	
	LA	-0.964	1.575	0.375	1	0.540	0.381
	FA	2.931	1.148	6.515	1	0.011	18.743
	Loc1	0.757	1.353	0.313	1	0.576	2.133
	Loc2	-0.012	1.356	<0.001	1	0.993	0.988
	Dist	-2.954	0.929	10.107	1	0.001	0.052
	Road	3.787	1.197	10.019	1	0.002	44.146
商务 办公 (NF = 2)	常数项	1.155	1.331	0.753	1	0.385	
	LA	-5.220	2.250	5.382	1	0.020	0.005
	FA	2.977	1.322	5.067	1	0.024	19.620
	Loc1	0.983	1.030	0.910	1	0.340	2.671
	Loc2	0.216	1.082	0.040	1	0.842	1.241
	Dist	-1.759	0.622	8.004	1	0.005	0.172
	Road	2.703	1.034	6.838	1	0.009	14.930
商业 服务 (NF = 3)	常数项	-0.861	1.224	0.495	1	0.482	
	LA	0.835	1.464	0.326	1	0.568	2.306
	FA	-0.208	1.141	0.033	1	0.855	0.812
	Loc1	1.985	0.937	4.486	1	0.034	7.282
	Loc2	-1.446	1.276	1.284	1	0.257	0.236
	Dist	-0.651	0.503	1.679	1	0.195	0.521
	Road	2.621	0.957	7.496	1	0.006	13.743

因变量参照水平：生产物流(NF = 4)

样本量(N): 113
变量个数(K): 21
对数似然值(LL): -102.739
Nagelkerke r^2 : 0.608
Cox & Snell r^2 : 0.565
McFadden r^2 : 0.314

三组参数，分别是工业用地更新为
创意产业园(NF = 1)、
商务办公(NF = 2)、
商业服务(NF = 3)
与参照水平——生产物流
(NF^{Ref} = 4)的比较。

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
(NF = 1)	常数项	-1.216	1.536	0.627	1	0.429	
	LA	-0.964	1.575	0.375	1	0.540	0.381
	创意 FA	2.931	1.148	6.515	1	0.011	18.743
	产业园 Loc1	0.757	1.353	0.313	1	0.576	2.133
	Loc2	-0.012	1.356	<0.001	1	0.993	0.988
	Dist	-2.954	0.929	10.107	1	0.001	0.052
	Road	3.787	1.197	10.019	1	0.002	44.146

更新为创新产业园($NF = 1$)的模型:

$$\ln(\text{odds}^{1,Ref}) = \ln\left(\frac{\Pr(NF = 1)}{\Pr(NF^{Ref} = 4)}\right)$$

$$= -1.216 - 0.964LA + 2.931FA + 0.757Loc1 - 0.012Loc2 - 2.954Dist + 3.787Road$$

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
创意 产业园 (NF = 1)	常数项	-1.216	1.536	0.627	1	0.429	
	LA	-0.964	1.575	0.375	1	0.540	0.381
	FA	2.931	1.148	6.515	1	0.011	<0.053
	Loc1	0.757	1.353	0.313	1	0.576	2.133
	Loc2	-0.012	1.356	<0.001	1	0.993	0.988
	Dist	-2.954	0.929	10.107	1	0.001	0.052
	Road	3.787	1.197	10.019	1	0.002	<0.05 44.146

显著性:

- 在 $\alpha = 0.05$ 的水平下, 影响显著的变量有 FA、Dist、Road。

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
(NF = 1)	常数项	-1.216	1.536	0.627	1	0.429	
	LA	-0.964	1.575	0.375	1	0.540	0.381
	创意 产业园	FA	2.931	1.148	6.515	0.011	18.743
		Loc1	0.757	1.353	0.313	1	2.133
		Loc2	-0.012	1.356	<0.001	1	0.988
		Dist	-2.954	0.929	10.107	1	0.052
		Road	3.787	1.197	10.019	1	44.146

odds ratio (优势比) :

FA 的 $\text{Exp}(B) = 18.743$, 大于1, 所以:

1. 建筑面积的增加, 将提高工业用地更新为创意产业园的可能性(相对于生产物流)
2. 其他条件不变, 建筑面积每增加 10000m^2 , 更新为创意产业园的概率与更新为生产物流的概率之比($\text{odds}^{1, \text{Ref}}$), 变为原来的 18.743倍。

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
创意 产业园 (NF = 1)	常数项	-1.216	1.536	0.627	1	0.429	
	LA	-0.964	1.575	0.375	1	0.540	0.381
	FA	2.931	1.148	6.515	1	0.011	18.743
	Loc1	0.757	1.353	0.313	1	0.576	2.133
	Loc2	-0.012	1.356	<0.001	1	0.993	0.988
	Dist	-2.954	0.929	10.107	1	0.001	0.052
	Road	3.787	1.197	10.019	1	0.002	44.146

odds ratio (优势比) :

Dist 的 $\text{Exp}(B) = 0.052$, 小于1, 所以:

1. 到区中心距离的增加, 将减少工业用地更新为创意产业园的可能性(相对于生产物流);
2. 其他条件不变, 距离每增加 1km, 更新为创意产业园的概率与更新为生产物流的概率之比($\text{odds}^{1, \text{Ref}}$), 变为原来的 5.2%。

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
(NF = 1)	常数项	-1.216	1.536	0.627	1	0.429	
	LA	-0.964	1.575	0.375	1	0.540	0.381
	创意 产业园	FA	2.931	1.148	6.515	0.011	18.743
	Loc1	0.757	1.353	0.313	1	0.576	2.133
	Loc2	-0.012	1.356	<0.001	1	0.993	0.988
	Dist	-2.954	0.929	10.107	1	0.001	0.052
	Road	3.787	1.197	10.019	1	0.002	44.146

odds ratio (优势比) :

Road 的 $\text{Exp}(B) = 44.146$, 大于1, 所以:

- “与干路相邻 (Road=1) 的工业用地” 相比于 “与支路相邻 (Road=0) 的工业用地” 更新为创意产业园的可能性(相对于生产物流)更高;
- 其他条件不变, 工业用地相邻道路由支路变为干路时, 更新为创意产业园的概率与更新为生产物流的概率之比($\text{odds}^{1, \text{Ref}}$), 变为原来的 44.146倍。

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)	
(NF = 2)	常数项	1.155	1.331	0.753	1	0.385		
	LA	-5.220	2.250	5.382	1	0.020	0.005	
	商务 办公	FA	2.977	1.322	5.067	1	0.024	19.620
	Loc1	0.983	1.030	0.910	1	0.340	2.671	
	Loc2	0.216	1.082	0.040	1	0.842	1.241	
	Dist	-1.759	0.622	8.004	1	0.005	0.172	
	Road	2.703	1.034	6.838	1	0.009	14.930	

更新为商务办公($NF = 2$)的模型:

$$\ln(\text{odds}^{2,Ref}) = \ln\left(\frac{\Pr(NF = 2)}{\Pr(NF^{Ref} = 4)}\right)$$

$$= 1.155 - 5.220LA + 2.977FA + 0.983Loc1 - 0.216Loc2 - 1.759Dist + 2.703Road$$

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
商务 办公 (NF = 2)	常数项	1.155	1.331	0.753	1	0.385	
	LA	-5.220	2.250	5.382	1	0.020	0.005
	FA	2.977	1.322	5.067	1	0.024	19.620
	Loc1	0.983	1.030	0.910	1	0.340	2.671
	Loc2	0.216	1.082	0.040	1	0.842	1.241
	Dist	-1.759	0.622	8.004	1	0.005	0.172
	Road	2.703	1.034	6.838	1	0.009	14.930

- 用地面积、到区中心的距离，对于更新为商务办公的可能性（相对于生产物流）具有显著的负效应；

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
商务 办公 (NF = 2)	常数项	1.155	1.331	0.753	1	0.385	
	LA	-5.220	2.250	5.382	1	0.020	0.005
	FA	2.977	1.322	5.067	1	0.024	19.620
	Loc1	0.983	1.030	0.910	1	0.340	2.671
	Loc2	0.216	1.082	0.040	1	0.842	1.241
	Dist	-1.759	0.622	8.004	1	0.005	0.172
	Road	2.703	1.034	6.838	1	0.009	14.930

- 用地面积、到区中心的距离，对于更新为商务办公的可能性（相对于生产物流）具有显著的负效应；
- 建筑面积、与干路相邻，则具有显著的正效应。

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
商业 服务 (NF = 3)	常数项	-0.861	1.224	0.495	1	0.482	
	LA	0.835	1.464	0.326	1	0.568	2.306
	FA	-0.208	1.141	0.033	1	0.855	0.812
	Loc1	1.985	0.937	4.486	1	0.034	7.282
	Loc2	-1.446	1.276	1.284	1	0.257	0.236
	Dist	-0.651	0.503	1.679	1	0.195	0.521
	Road	2.621	0.957	7.496	1	0.006	13.743

更新为商业服务($NF = 3$)的模型:

$$\begin{aligned}\ln(\text{odds}^{3,Ref}) &= \ln\left(\frac{\Pr(NF = 3)}{\Pr(NF^{Ref} = 4)}\right) \\ &= -0.861 + 0.835LA - 0.208FA + 1.985Loc1 - 1.446Loc2 - 0.651Dist + 2.621Road\end{aligned}$$

多元Logistic回归

NF	变量	系数(B)	标准误 (S.E.)	Wald 值	自由度 (df)	显著性 (Sig.)	优势比 Exp(B)
商业 服务 (NF = 3)	常数项	-0.861	1.224	0.495	1	0.482	
	LA	0.835	1.464	0.326	1	0.568	2.306
	FA	-0.208	1.141	0.033	1	0.855	0.812
	Loc1	1.985	0.937	4.486	1	0.034	7.282
	Loc2	-1.446	1.276	1.284	1	0.257	0.236
	Dist	-0.651	0.503	1.679	1	0.195	0.521
	Road	2.621	0.957	7.496	1	0.006	13.743

- 位于内环内、与干路相邻，对于更新为商业服务的可能性(相对于生产物流)具有显著的正效应。

为什么要反复强调“相对于生产物流”呢？

因为当参照水平发生变化时，结论也会发生变化！

- 当距离增加时，工业用地更新为商务办公的可能性(相对于生产物流)降低，反映出商务办公具有更高的中心可达性要求。

$$\ln(\text{odds}^{2,Ref}) = \ln\left(\frac{\Pr(NF = 2)}{\Pr(NF^{Ref} = 4)}\right) = 1.155 - 5.220LA + 2.977FA + 0.983Loc1 - 0.216Loc2 - \mathbf{1.759Dist} + 2.703Road$$

- 然而，如果是跟创意产业园比呢？

$$\ln(\text{odds}^{1,Ref}) = \ln\left(\frac{\Pr(NF = 1)}{\Pr(NF^{Ref} = 4)}\right) = -1.216 - 0.964LA + 2.931FA + 0.757Loc1 - 0.012Loc2 - \mathbf{2.954Dist} + 3.787Road$$

- 当转换比较基准时，会发现工业用地更新为商务办公的可能性(相对于创意产业园)提高，反映出商务办公对于中心可达性的要求不如创意产业园。

多元Logistic回归

- 虽然没有直接比较商务办公($NF = 2$) 和创意产业园($NF = 1$), 但可以通过生产物流为媒介进行间接比较:

$$\text{odds}^{2,1} = \frac{\Pr(NF = 2)}{\Pr(NF = 1)} = \frac{\Pr(NF = 2)/\Pr(NF^{Ref} = 4)}{\Pr(NF = 1)/\Pr(NF^{Ref} = 4)} = \frac{\text{odds}^{2,Ref}}{\text{odds}^{1,Ref}}$$

- 距离每增加1km, $\text{odds}^{2,Ref}$ 变为原来的17.2%, $\text{odds}^{1,Ref}$ 变为原来的5.2%。
- $\text{odds}^{2,1}$ ——更新为商务办公的概率与更新为创意产业园的概率之比——将变为原来的 $17.2\%/5.2\% = 3.303$ 倍。
- 如果要判断显著性, 需要将被比较的水平设置为参照水平, 然后重新估计模型。

多元Logistic回归预测类别概率

- 本案例中 113 块工业用地，预测结果(取概率最大的类别)和实际结果的列联表如下：（对角线上的数字代表预测正确）

实际结果	预测结果				
	创意产业园	商务办公	商业服务	生产物流	正确率
创意产业园	7	3	2	1	53.8%
商务办公	2	20	7	4	60.6%
商业服务	0	9	20	1	66.7%
生产物流	0	5	6	26	70.3%
总计占比	8.0%	32.7%	31.0%	28.3%	64.6%

多元Logistic回归

多元Logistic回归预测类别概率

- 本案例中 113 块工业用地，预测结果(取概率最大的类别)和实际结果的列联表如下：（对角线上的数字代表预测正确）

实际结果	预测结果				
	创意产业园	商务办公	商业服务	生产物流	正确率
创意产业园	7	3	2	1	53.8%
商务办公	2	20	7	4	60.6%
商业服务	0	9	20	1	66.7%
生产物流	0	5	6	26	70.3%
总计占比	8.0%	32.7%	31.0%	28.3%	64.6%

- 模型总体的预测正确率是 64.6%，较为理想，对生产物流的预测效果最好(70.3%)，对创意产业园的预测效果相对最差(53.8%)。

总结

系数估计结果		二元Logistic回归		多元Logistic回归	
b_i	e^{b_i} (OR)	odds的变化	解释	odds的变化	解释
>0	>1	事件发生相对于不发生的优势提高	x_i 的增加使事件更有可能发生	类别 n 相对于参照水平的优势提高	x_i 的增加使样本属于类别 n 的可能性（相对于参照水平）增加
=0	=1	事件发生相对于不发生的优势不变	x_i 的变化对事件的发生没有影响	类别 n 相对于参照水平的优势不变	x_i 的变化对样本属于类别 n 的可能性（相对于参照水平）无影响
<0	<1	事件发生相对于不发生的优势降低	x_i 的增加使事件更有可能不发生	类别 n 相对于参照水平的优势降低	x_i 的增加使样本属于类别 n 的可能性（相对于参照水平）减少