

医学数据分析与可视化技术能力提升培训班

第五讲 回归分析与生存分析

主讲：龚德鑫

2024年12月

广东 深圳

内 容

一

简单相关与回归

二

Logistic回归基础与实践

三

生存分析基础与实践

四

小结



简单相关与回归

相关的基本概念

定义：研究因素间的依存关系，并探讨其**相关方向**以及**相关程度**。

相关系数：定量描述变量之间的关系，系数符号（ \pm ）表明关系方向，数值大小表示关系强弱。

- **Pearson**相关系数：衡量**连续型**、来自**正态**分布总体两变量之间的线性相关程度。
- **Spearman**等级相关系数：衡量**分级定序**变量（满足单调关系）之间的相关程度。
- **Kendall' s Tau**相关系数：一种非参数的等级相关度量。
- 其他：偏相关系数、多分格(polychoric)相关系数和多系列(polyserial)相关系数








相关分析的一般步骤



回归的基本概念

“Regression”最早出现在1886年英国遗传学家Francis Galton的一篇研究身高的论文，他发现子女身高会向整个群体身高的均值回归。

搜索[regression](#)显示的相关词条

 Nonlinear regression Iterative procedure	 Quantile regression Conditional quantiles	Local regression Weighted averages	General linear model Analysis of variance	Discrete choice Binary or multinomial
 Logistic regression Sigmoid curve	 Isotonic regression Monotonic constraint	 Errors-in-variables models Measurement error	Generalized linear model Link function	Binomial regression Binary regression
 Probit Normal CDF	Least-angle regression Variable selection	 Linear regression Simple regression	Vector generalized linear model Multivariate response	Multinomial logistic regression Multiple categories
Mixed logit Random coefficients	Ordered probit Ordinal categories	Fixed effects model Within variation	Nonparametric regression No functional form	Principal component regression Dimension reduction
Multinomial probit Multivariate normal	Poisson regression Count data	Random effects model Between variation	Semiparametric regression Partial functional form	Segmented regression Piecewise linear
Ordered logit Ordinal categories	Multilevel model Hierarchical data	Nonlinear mixed-effects model Nonlinear mixed effects	Robust regression Outlier resistant	Linear mixed-effects model Mixed effects

现今[回归](#)更多是指代一种“[方法](#)”，指研究因变量 $Y_{1,...,i}$ 和自变量 $X_{1,...,k}$ 之间关系的统计分析方法。

相关和回归的关系

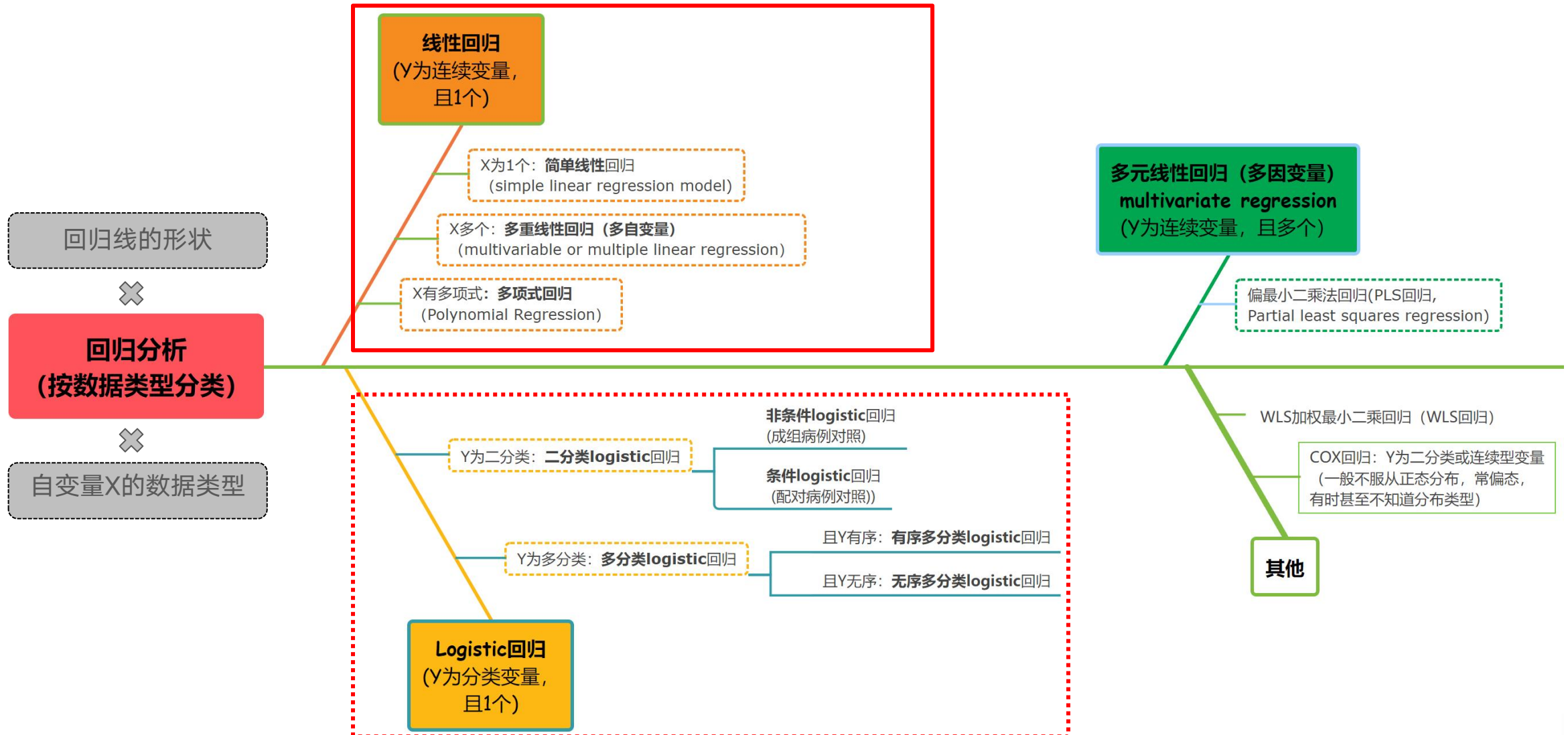
回归分析是一种数学模型，当因变量和自变量为线性关系时，它是一种特殊的线性模型。

一般的，先有相关，才行回归。

回归分析的核心目的在于解释和预测。

Aspect	相关(Correlation)	回归(Regression)
Purpose	Measures strength and direction of the relationship	Predicts and models the relationship
Variables	Two variables (equal roles)	Independent and dependent variables
Calculation	Correlation coefficient (r)	Regression equation ($y = ax + b$)
Direction	+1 to -1 (positive, negative, no Correlation)	Positive, negative (strength and direction)
Causation	It does not imply causation	Can imply causation under controlled conditions
Handling outliers	Relatively robust	Sensitive outliers can distort the Regression line
Data representation	Single coefficient	The equation representing the relationship
Application	Descriptive , identifies patterns	Predictive, forecasts future values
Complexity	Simple method	Variable complexity (simple to multivariate)
Interpretation	Limited to strength and direction	Detailed insights into the relationship
Hypothesis testing	Limited to Correlation significance	Tests coefficients' importance in the model
Time-series analysis	Limited predictive power	Helpful in forecasting future trends
Usage	Preliminary analysis, identifies associations	Prediction, modelling, understanding impact

回归分析方法概述



回归分析的常见变体

回归类型	用 途
简单线性	用一个量化的解释变量预测一个量化的响应变量
多项式	用一个量化的解释变量预测一个量化的响应变量，模型的关系是 n 阶多项式
多层	用拥有等级结构的数据预测一个响应变量（例如学校中教室里的学生）。也被称为分层模型、嵌套模型或混合模型
多元线性	用两个或多个量化的解释变量预测一个量化的响应变量
多变量	用一个或多个解释变量预测多个响应变量
Logistic	用一个或多个解释变量预测一个类别型响应变量
泊松	用一个或多个解释变量预测一个代表频数的响应变量
Cox 比例风险	用一个或多个解释变量预测一个事件（死亡、失败或旧病复发）发生的时间
时间序列	对误差项相关的时间序列数据建模
非线性	用一个或多个量化的解释变量预测一个量化的响应变量，不过模型是非线性的
非参数	用一个或多个量化的解释变量预测一个量化的响应变量，模型的形式源自数据形式，不事先设定
稳健	用一个或多个量化的解释变量预测一个量化的响应变量，能抵御强影响点的干扰

简单线性回归

- 经典单因素分析方法
- 简单线性回归方法的拓展
- 定量地描述一个因变量Y和一个自变量X之间的线性依存关系。

$$Y_i = a + \beta X_i + \varepsilon_i$$

多重线性回归

- 经典多因素分析方法
- 简单线性回归方法的拓展
- 定量地描述一个因变量Y和多个自变量X之间的线性依存关系。

$$Y = a + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_m x_m + \varepsilon$$

Y : 因变量, X_1, X_2, \dots, X_m 为 k 个自变量。

α : 常数项

β_m : 偏回归系数, 表示在其它自变量保持不变的情况下,
 X_m 增加或减少一个单位时, Y 的平均改变量。

线性回归对资料的要求

简单线性回归：因变量是连续型变量，且符合正态分布

多重（多自变量）线性回归分析：

- ✓ 因变量：连续型变量
- ✓ 自变量：无要求
 - 连续型变量（如年龄、血压）
 - 分类变量
 - 二分类（如性别）
 - 有序分类（肿瘤分期、疗效分级）
 - 无序分类变量（季节、血型、教育程度等）： $k-1$ 个哑变量赋值。

思考：

如果Y为二分类变量，如何分析？例如：

是否发病：0=否，1=是

死亡与否：0=否，1=是

➡ Logistic回归



PART ONE

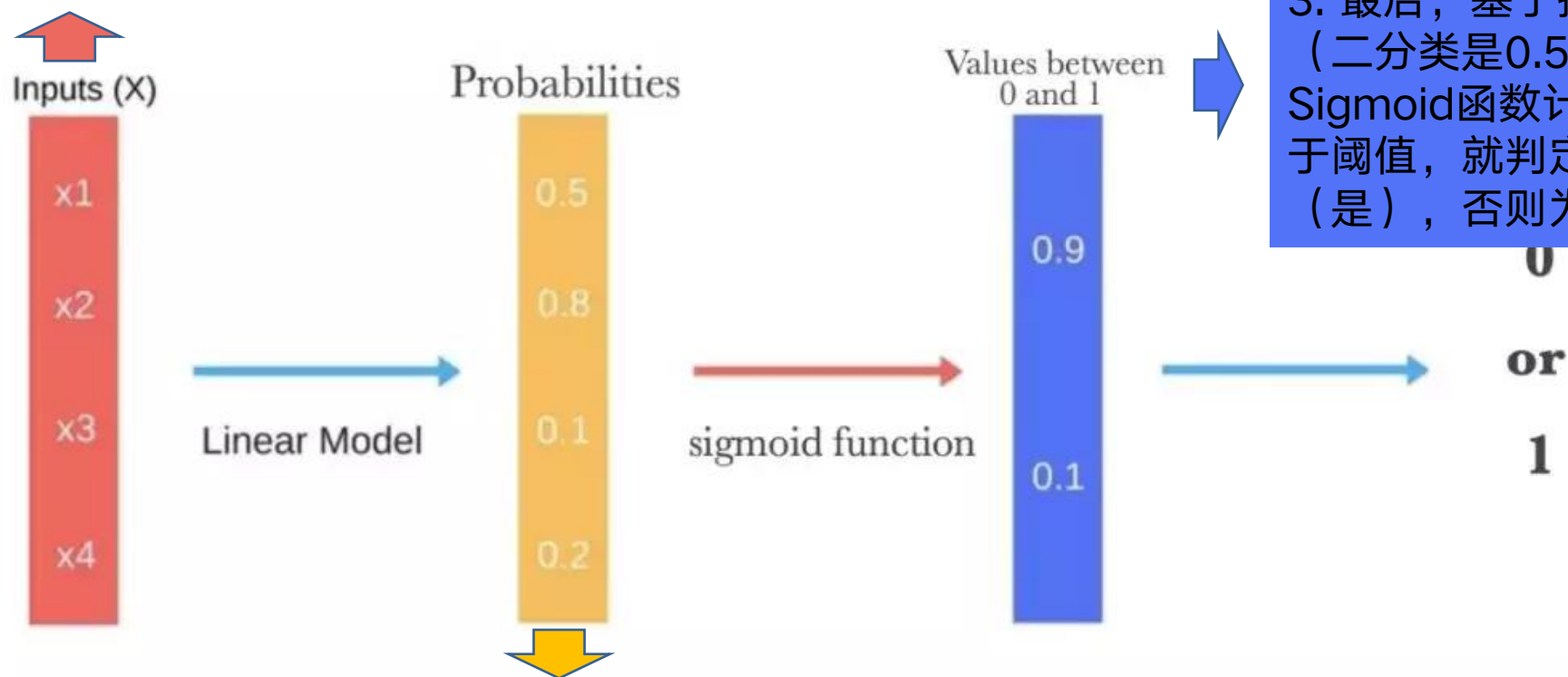
Logistic回归- 基础与实践



Logistic回归基础

Logistic回归原理

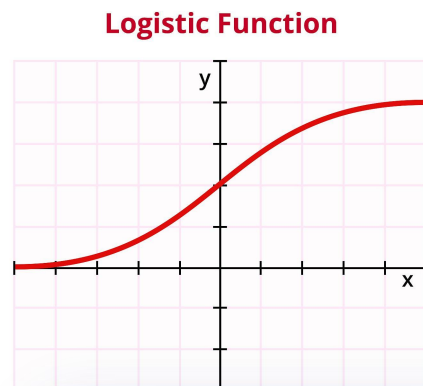
1. 首先，Logistic回归会计算输入特征（如年龄、体重等）的线性组合，这类似于多元线性回归，输出某个值。



3. 最后，基于提前设定的阈值（二分类是0.5），如果 Sigmoid函数计算出来的值大于阈值，就判定结局为阳性（是），否则为阴性（否）。

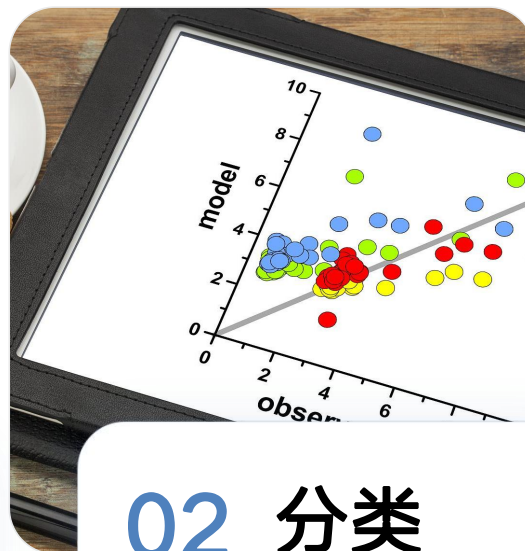
2. 接着，Sigmoid 函数将这个值处理，映射使得其范围介于 0 和 1 之间。

Logistic回归概念



01 定义

Logistic函数，也称为sigmoid函数，将任意实数值映射到 $(0,1)$ 区间，用于概率预测。



02 分类

Logistic回归模型通过设定阈值来确定分类决策边界，将数据分为**两类或多类**。

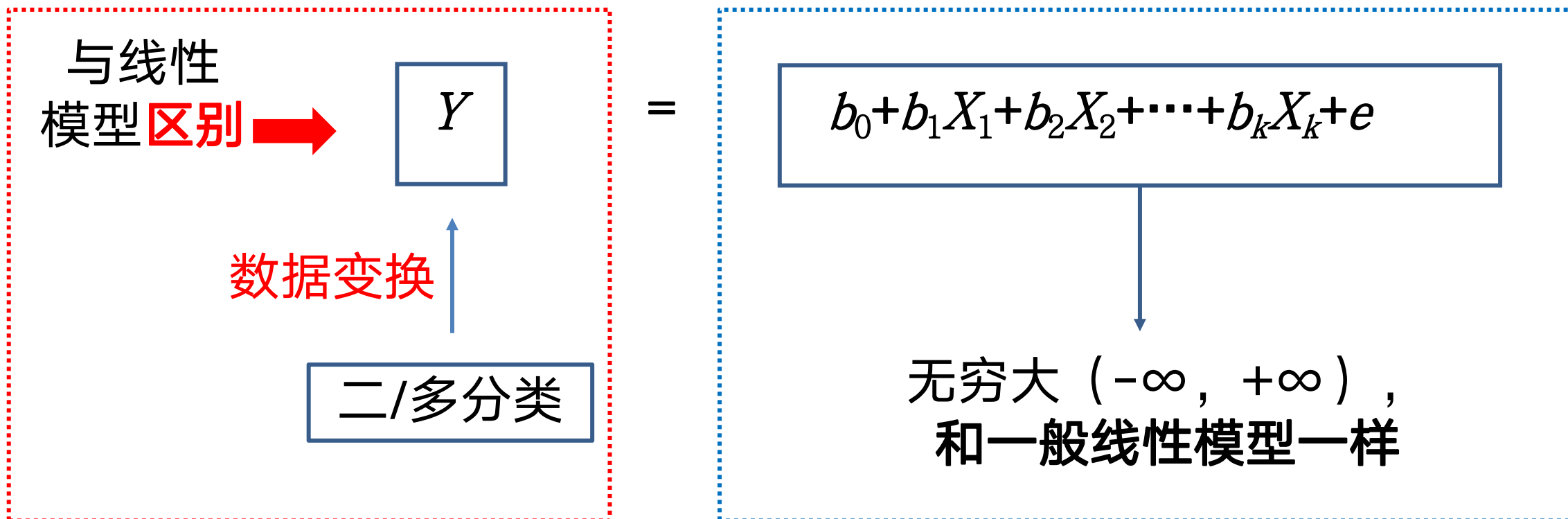


03 解释

Logistic回归提供事件发生的概率估计，并能计算出不同特征值的优势比(OR)来解释影响。

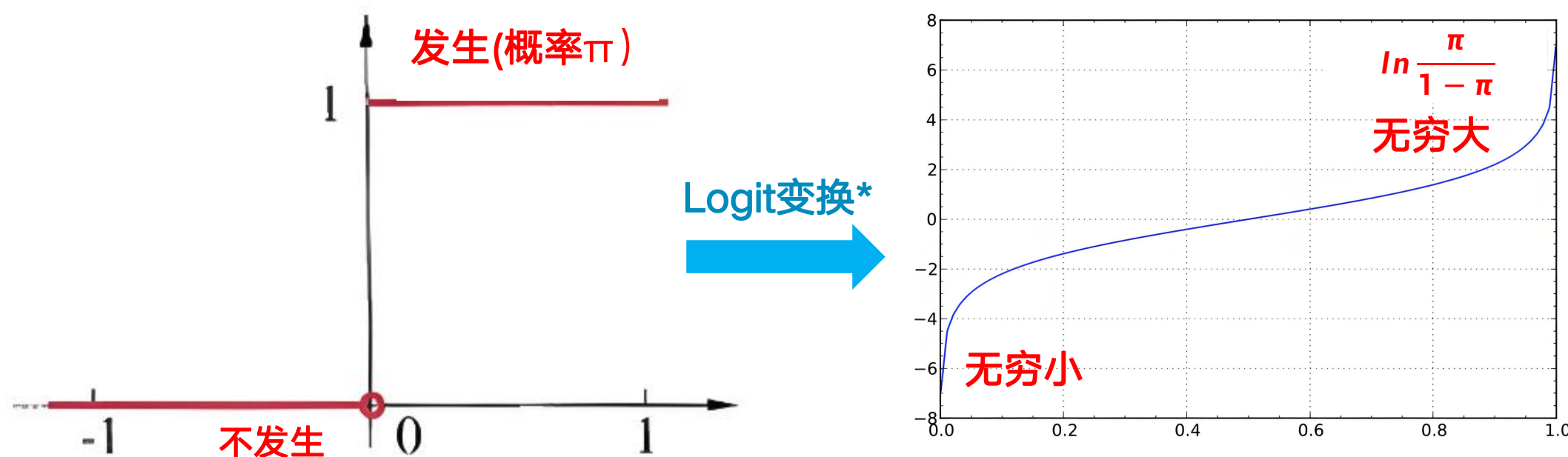
什么是Logistic回归？

定义：属于概率性非线性回归，主要研究**二分类变量**（可扩展到多分类观察结果与影响因素之间关系的一种多变量分析方法。



Logistic回归数学原理

因变量取值范围：以二分类变量为例，某个概率作为方程的因变量的取值范围为**0-1**，但是，一般线性模型右边取值范围是无穷大或者无穷小。所以，直接建立线性模型不合适。但是.....



备注：若是右图变换为左图，则为Sigmoid函数，其在形式上与Logit变换相反，可看作其逆过程。

*Logit变换：发生概率除以没有发生概率再取对数，即logit变换，将发生和未发生的概率成为了比值。

Logistic回归公式

- Logistic回归模型: $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$
- β_i 表示自变量 x_i 改变一个单位时, $\text{logit}(p)$ 的改变量。
- 其它形式:

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m}}$$
$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m)}}$$

Logistic回归系数与结果解释

基于公式，回归系数 β_k 的解释为：保持其他自变量不变，自变量 X_k 每改变一个单位， $\text{logit}(P)$ 的改变量，与优势比（odds ratio, OR)有一个对应关系。

自变量 X_k 的回归系数：

- $\beta_k > 0$ 时， $OR_k > 1$ ，提示 X_k 是危险因素；
- 当 $\beta_k < 0$ 时， $OR_k < 1$ ，提示 X_k 为保护因素；
- 回归系数 $\beta_k = 0$ 时， $OR_k = 1$ ，提示 X_k 为无关因素

Logistic回归应用场景



预测/判别/分类

Logistic回归在医疗领域用于疾病预测，如心脏病风险评估；预测患者疾病复发的概率



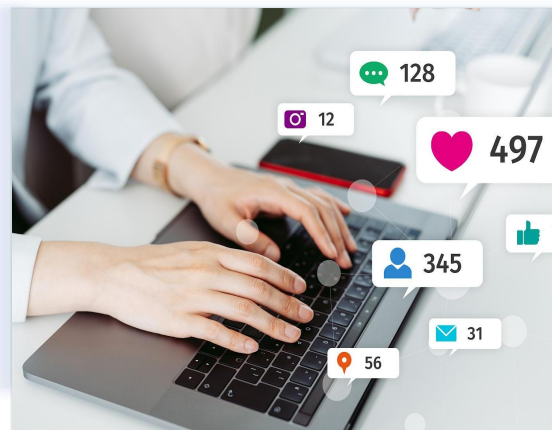
分析影响因素

医学研究中，可以用来分析吸烟、饮酒和特定职业是否是食管癌的危险因素。



校正混杂因素

控制混杂变量，从而更准确地评估特定因素与结果之间的关系。



临床/公卫干预

比较不同因素对疾病发生风险的贡献程度，从而确定哪些因素具有更高的相对重要性。



Logistic回归实践

logistic回归实践 - 数据

(一) 实践数据

数据文件 :.PyData2403/GdAdultPhy1000_05.xlsx。

Number	Sex	Height	Weight	Age	BMI	PhysiLv1	PhysiLv2	AgeGroup	AgeGroup 2
1	1	163.3	69.3	52	25.99	overweight	0	45-	45-
2	2	152	54.6	42	23.63	normal	0	35-	30-
3	1	166.2	83.7	65	30.3	obese	1	65-	60-
4	2	152.6	59.5	45	25.55	overweight	0	45-	45-
5	2	159.8	56.2	75	22.01	normal	0	75-	75-

logistic回归实践 - 任务

(二) 实践任务

以体质分类（二分类）为因变量，以性别、身高和体重为自变量拟合Logistic回归。

logistic回归实践程序

```
import pandas as pd; import statsmodels.api as sm
dataFram=pd.read_excel('./PyData2403/GdAdultPhy1000_04.xlsx',index_col='Number')
Gender_dummy=pd.get_dummies(dataFram['Gender'],prefix='Gender',dtype=int)
# 将Gender列转换为哑变量，设置哑变量名的前缀为Gender
# 哑变量名会自动设置为Gender_1和Gender_2
# 设置哑变量数据类型为整型（0/1），默认为布尔型（True/False）
dataFram_encoded=pd.concat([dataFram,Gender_dummy],axis=1)
# 合并原始特征与哑变量
Y=dataFram['PhysiLv2'] # 获取体质分类（二分类）值作因变量
X=dataFram_encoded[['Gender_2','Height','Weight']] # 获取自变量值
# 以性别哑变量Gender_1为对照，因此性别自变量取Gender_2
X1=sm.add_constant(X) # 添加模型截距
logit_model=sm.Logit(Y,X1); logit_results=logit_model.fit( ) # 构建Logit模型拟合模型
print('1-1 Logit模型拟合结果：\n',logit_results.summary( ))
predicts=logit_results.predict(X1) # 计算预测值
predictsDF=pd.DataFrame(predicts) # 将预测结果转化为数据帧便于控制输出数据
print('1-2 Logit模型拟合预测值(前5个)：\n',predictsDF.head( ))
```

logistic回归实践结果

Warning: Maximum number of iterations has been exceeded.

Current function value: 0.000039

Iterations: 35

1-1 Logit 模型拟合结果:

Logit Regression Results

=====			
Dep. Variable:	PhysiLv2	No. Observations:	944
Model:	<u>Logit</u>	<u>Df</u> Residuals:	940
Method:	MLE	<u>Df</u> Model:	3
Date:	Thu, 21 Mar 2024	Pseudo <u>R-squ.</u> :	0.9999
Time:	21:14:25	Log-Likelihood:	-0.036955
converged:	False	LL-Null:	-287.99
Covariance Type:	<u>nonrobust</u>	LLR p-value:	1.676e-124
=====			

拟合优度的指标

似然比检验p值，检验模型与空模型的差异是否显著

logistic回归实践结果

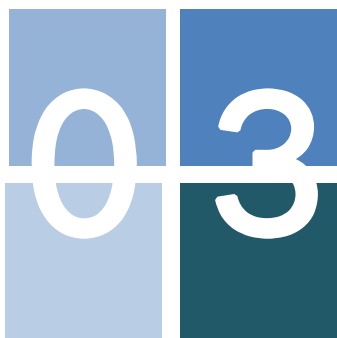
	<u>coef</u>	std err	z	P> z	[0.025	0.975]
<u>const</u>	7483.4814	9056.036	0.826	0.409	-1.03e+04	2.52e+04
Gender_2	10.2665	200.004	0.051	0.959	-381.735	402.268
Height	-97.6567	118.269	-0.826	0.409	-329.461	134.147
Weight	113.6054	137.573	0.826	0.409	-156.032	383.243

Possibly complete quasi-separation: A fraction 1.00 of observations can be perfectly predicted.

This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

1-2 Logit 模型拟合预测值(前 5 个):

Number	0
1	2.149846e-257
2	0.000000e+00
3	1.000000e+00
4	1.208973e-282
5	0.000000e+00



PART THREE

生存分析基础与实践



生存分析基础

生存资料案例

案例：某肿瘤医院调查了1991-1995年间经手术治疗的大肠癌患者150例，对可能影响大肠癌术后生存时间的因素进行了调查，如性别、年龄、组织学分类、肿瘤大小、Dure'S分期等。随访截止日期为2000年12月30日，随访记录见下表。

关心结局（生/死），也关心经历的时间（生存时间）

编号	性别	年龄	...	手术日期	随访终止日期	随访结局	生存时间(天)	
1	男	45	...	1991.05.20	1995.06.04	死亡	1476	
2	男	50	...	1992.01.12	1998.08.25	死亡	2417	
3	女	36	...	1991.10.24	1994.03.18	失访	876 ⁺	删失
4	男	52	...	1994.11.02	2000.12.30	存活	2250 ⁺	(censoring)
5	女	56	...	1994.06.25	1995.03.17	死亡	265	截尾
6	女	60	...	1993.12.05	1996.08.16	死于其它	985 ⁺	(truncation)
...								

生存分析相关概念

在医学研究中，除了考虑某事件发生与否（ $Y=0/1$ ），还考虑产生结局时**经历的时长**，此种分析即为生存分析，此数据为生存数据。

Survival analysis (time-to-event analysis / event history analysis)

生存时间(survival time, failure time)

- 终点事件与起始事件之间的时间间隔。
- 终点事件指研究者所关心的特定结局。
- 起始事件是反映研究对象生存过程的起始特征的事件。

生存时间举例

起始事件		终点事件
服药	→	痊愈
手术切除		死亡
染毒		死亡
化疗		缓解
缓解		复发

生存分析的目的

1. 构建生存函数及危险函数曲线

==》描述/估计

- 生存函数（Survival Function）：描述了在一段时间内**个体存活的概率**。通过生存函数，我们可以了解个体在特定时间点的生存概率，这对于评估治疗效果、预测患者预后等具有重要意义。
- 危险函数（Hazard Function）：表示在**给定时间点，个体发生事件的瞬时率**。危险函数有助于识别高风险时期，比如某些治疗可能在特定时间段内更有效或风险更高。

2. 比较不同生存函数和危险函数曲线的差异

==》比较

- 评估治疗效果、风险因素以及患者预后的差异

3. 使用数学建模方法，评估解释变量与生存时间之间的关系(例如Cox 比例风险模型)

==》解释

- 是一种半参数模型，用于评估**多个解释变量（如年龄、性别、生活方式等）对生存时间的影响**。Cox模型允许研究者评估这些变量对生存风险的相对影响，而不需要假设生存时间的特定分布。

生存分析常用方法

1. Kaplan-Meier估计:

==》描述/估计

- 用于非参数估计生存函数，适用于右删失数据。
- 通过将每个时间点的生存概率相乘来估计生存函数。

2. Nelson-Aalen估计:

==》描述/估计

- 用于非参数估计累积危险函数。
- 通过将每个时间点的危险函数相加来估计累积危险函数。

3. Log-Rank检验:

==》比较

- 用于比较两个或多个生存曲线的差异。
- 基于事件发生时间的秩次进行检验。

4. Cox比例风险模型:

==》解释

- 一种半参数回归模型，用于评估一个或多个自变量对生存时间的影响。
- 假设危险函数是基线危险函数与自变量的指数函数的乘积。

生存分析相关概念

累积生存概率 / 生存率：实际上描述的是同一个概念。

累积死亡概率 / 累积死亡概率曲线：描述相同的概念，一个是数值表达，另一个是图形表示。

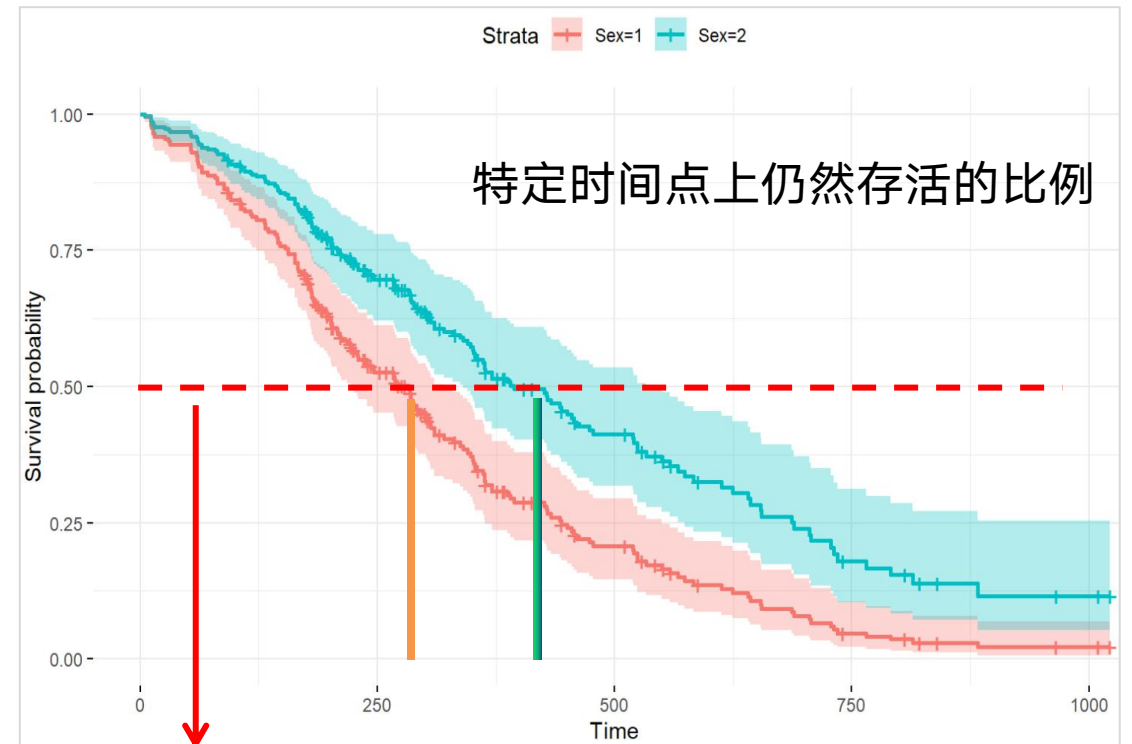
累积生存概率和累积死亡概率是互补的，即一个增加时另一个减少。当一个时间点上的累积死亡概率增加，意味着该时间点的累积生存概率相应地减少。

概念	描述	数值范围	图形表示	相关公式及关系
累积生存概率 / 生存率 ($S(t)$)	到某个时间点为止个体仍然存活的概率；特定时间段内的存活比例	0 到 1	下降的曲线或生存曲线	$S(t) = P(T > t)$
累积死亡概率 / 累积死亡概率曲线 ($F(t)$)	到某时间点为止的死亡概率总和；展示累积死亡概率随时间的变化	0 到 1	增加的曲线	$F(t) = 1 - S(t)$
瞬时死亡率/危险率 ($h(t)$)	在给定时间点发生事件的风险	非负实数	曲线或函数	$h(t) = \frac{f(t)}{S(t)}$
累积危害风险 ($H(t)$)	到某时间点为止的事件发生风险累积	非负实数	增加的曲线	$H(t) = -\ln(S(t))$
死亡概率密度 ($f(t)$)	死亡发生的瞬时概率	非负实数	曲线下面积代表概率	$f(t) = h(t)S(t)$
死亡风险	特定时间段内死亡的可能性	0 到 1	可能是柱状图或线图	通过将每个时间段的死亡风险相加以获得 $F(t)$

生存分析相关概念

生存曲线：是生存分析中的一种图形表示方法，用于描述和展示一个群体随时间变化的**生存概率**。生存曲线以时间为横轴（X轴），生存概率为纵轴（Y轴），通过绘制连续型的阶梯形曲线，直观地反映在不同时间点上，**群体中存活下来的比例**。

生存曲线



中位生存时间：50%的患者存活或未发生特定事件的时间点。

Kaplan-Meier估计里面，找到生存概率为0.5的那个时间点，就是中位生存时间



生存分析实践

--以肿瘤患者的生存时间及其影响因素分析为例

生存分析包

在Python中，可以利用lifelines包进行生存分析。

其方法包括：

- ①通过Kaplan-Meier plots绘图使生存曲线可视化（描述）；
- ②通过Nelson-Aalen plots绘图可视化累积危害风险（描述）；
- ③通过Log-Rank test检验比较两组或更多组的生存曲线；
- ④通过Cox比例风险回归，揭示（解释）不同变量对生存的影响（见第三节）。

使用lifelines包前，应先通过Anaconda Prompt采用pip install lifelines或conda install lifelines命令完成该包的安装。

生存分析实践数据

一、实践数据

A、B两种治疗方案（简称A组、B组）分别治疗某恶性肿瘤患者25人和22人，随访记录患者的生存时间（月）如下，“+”表示删失数据（无法得知随访对象的确切生存时间者）。

	A	B	C	D
1		Group	Time	Status
2	0	A组	10	1
3	1	A组	2	0
4	2	A组	12	1
5	3	A组	13	1
6	4	A组	18	1
7	5	A组	6	0
8	6	A组	19	1
9	7	A组	26	1
10	8	A组	9	0
11	9	A组	8	1

生存分析实践任务

- (1) 分析生存率和绘制生存曲线。
- (2) 分析死亡风险和绘制累积死亡概率曲线。
- (3) 对两组生存率进行比较。
- (4) 对两组的累积风险进行比较。

核心实践程序

- 1. 分析生存率和绘制生存曲线

```
kmf=KaplanMeierFitter( )    # 创建KaplanMeier拟合对象
kmf.fit(durations=data['Time'], event_observed=data['Status'])    # 数据拟合

median=kmf.median_survival_time_    # 计算中位生存时间
median_CI=median_survival_times(kmf.confidence_interval_)    # 计算中位生存时间的95%CI

kmf.fit(data['Time'][Egroup],data['Status'][Egroup],label='A组')    # A组数据拟合
ax=kmf.plot(show_censors=True,ci_show=False)    # 绘制A组生存率曲线图
TreatG_median=kmf.median_survival_time_    # 计算A组中位生存时间
TreatG_median_CI=median_survival_times(kmf.confidence_interval_)
    # A组中位生存时间的95%CI
print("1-3 A组中位生存时间及其95%CI: \n", TreatG_median,'\n',TreatG_median_CI)
```

核心实践程序

- 2. 分析死亡风险和绘制累积死亡概率曲线

```
from lifelines import KaplanMeierFitter    # 导入生存分析KaplanMeierFitter方法
```

```
kmf=KaplanMeierFitter( )    # 创建KaplanMeier拟合对象
```

```
kmf.fit(durations=data['Time'], event_observed=data['Status'])    # 数据拟合
```

```
groups=data['Group']    # 获取分组数据
```

```
kmf.fit(data['Time'][(groups=='A组')], data['Status'][(groups=='A组')], label='A组')
```

```
# A组数据拟合, (groups=='A组')的( )号可有可无
```

```
ax=kmf.plot_cumulative_density( )    # 绘制A组累积死亡概率曲线图
```

```
kmf.fit(data['Time'][groups=='B组'],data['Status'][groups=='B组'],label='B组')
```

```
# B组数据拟合
```

```
ax=kmf.plot_cumulative_density(ax=ax)    # 绘制B组累积死亡概率曲线图
```

核心实践程序

- 3. 对两组生存率进行比较

```
from lifelines import KaplanMeierFitter # 导入生存分析KaplanMeierFitter方法
```

```
from lifelines.statistics import logrank_test
```

```
kmf=KaplanMeierFitter( ) # 创建KaplanMeier拟合对象
```

```
kmf.fit(durations=data['Time'][data['Group']=='A组'],event_observed=data['Status']
```

```
      [data['Group']=='A组'], label='A组') # A组数据拟合
```

```
kmf.fit(data['Time'][~(data['Group']=='A组')],
```

```
      data['Status'][~(data['Group']=='A组')],label='B组')
```

```
# B组数据拟合，注意~(data['Group']=='A组')中的( )号不能省
```

```
lr_test=logrank_test(data['Time'][data['Group']=='A组'],
```

```
                    data['Time'][data['Group']=='B组'],
```

```
                    data['Status'][data['Group']=='A组'],
```

```
                    data['Status'][data['Group']=='B组'], alpha=0.95)
```

```
# 两组生存率比较检验
```

核心实践程序

- 4. 对两组的累积风险进行比较

```
from lifelines import KaplanMeierFitter    # 导入生存分析KaplanMeierFitter方法
from lifelines import NelsonAalenFitter    # 导入生存分析NelsonAalenFitter方法
data_1=DataF[DataF['Group']=='A组']      # 获取A组数据
kmf_1=KaplanMeierFitter( )               # 创建KaplanMeier拟合对象
kmf_1.fit(durations=data_1['Time'], event_observed=data_1['Status'])    # 数据拟合
kmf_1.plot_cumulative_density(label='cumulative_density' )    # 绘制累积风险密度图
naf_1.plot_cumulative_hazard(label='cumulative_hazard')    # 绘制累积风险图

# 绘制两组的累积风险概率密度比较图
kmf_1.plot_cumulative_density(label='A组' )
kmf_2.plot_cumulative_density(label='B组' )
```

实践结果（详见代码）

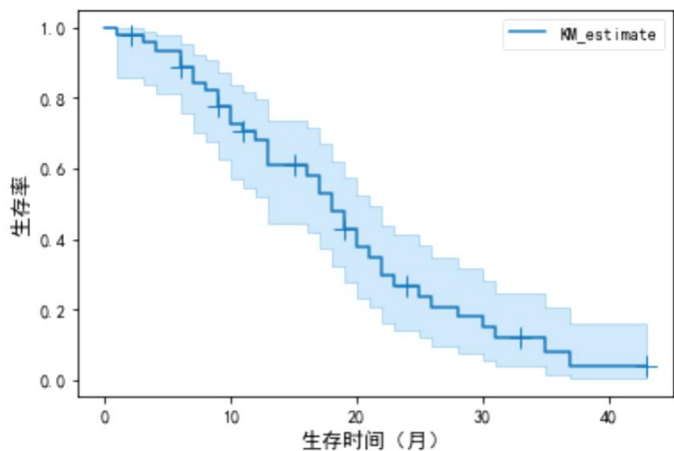


图5-3 两组合计生存率及其95%CI曲线图

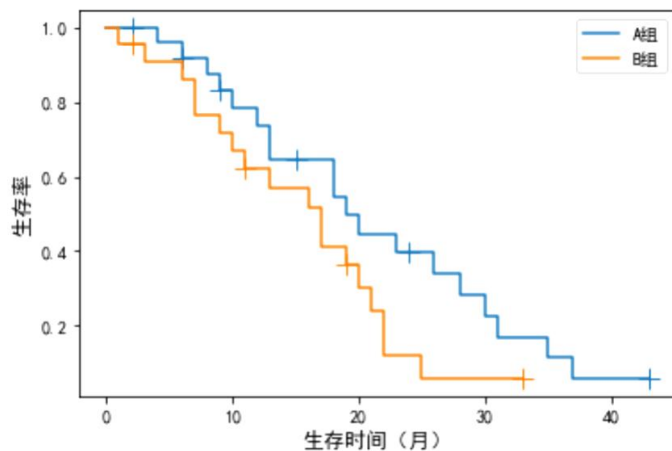


图5-4 A、B两组的生存率曲线对比图

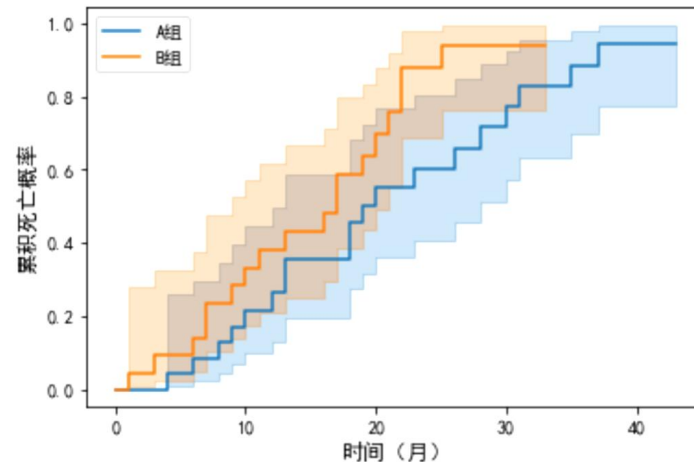


图5-5 A、B两组的累积死亡概率对比图

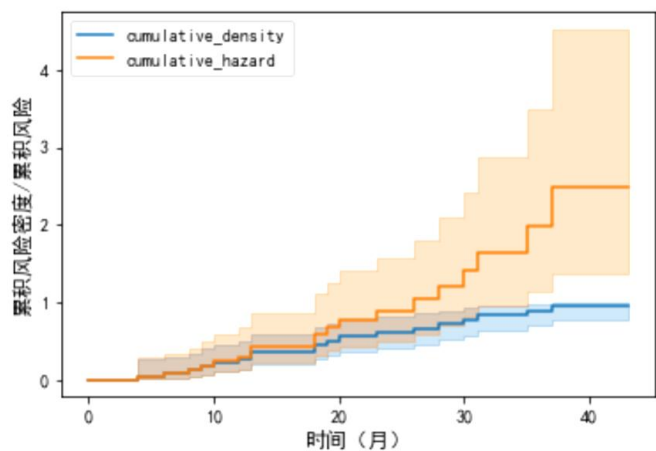


图5-6 A组累积死亡风险概率密度与累积风险

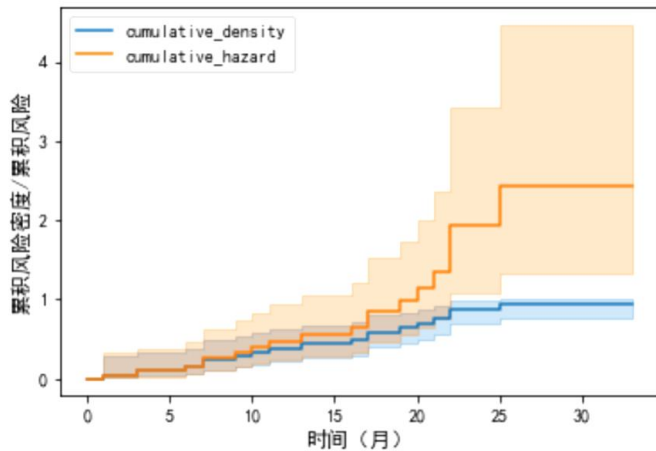


图5-7 B组累积死亡风险概率密度与累积风险图

2-1 A 组累积死亡概率:

timeline	A 组
0.0	0.000000

2-2 B 组累积死亡概率:

timeline	B 组
0.0	0.000000

2.0	0.000000
4.0	0.041667
...	...
35.0	0.886583
37.0	0.943291
43.0	0.943291

1.0	0.045455
2.0	0.045455
...	...
22.0	0.879356
25.0	0.939678
33.0	0.939678



Cox比例风险回归模型

——以肿瘤患者的生存时间及其影响因素分析
为例

Cox模型介绍

Cox模型是由英国统计学家D.R. Cox于1972年首次提出的一种半参数回归模型，是一种广泛用于生存分析领域的统计方法。

原理：

Cox模型**基于比例风险假设**，即不同个体的危险率（hazard rate）之间的比值是恒定的，不随时间变化。这意味着模型中的协变量对生存时间的影响是乘性的，即它们影响的是危险率的比例，而不是危险率的绝对值。

目的：

- 评估**解释变量**（如年龄、性别、治疗方式等）对生存时间的影响。通过模型，研究者可以量化每个解释变量对生存风险的影响，并且可以比较不同组别之间的生存风险
- 构建生存**预测模型**，帮助进行临床决策和资源分配

Cox模型的公式

$$h(t, X) = h_0(t) \exp(\beta_1 X_1 + \dots + \beta_p X_p)$$

- $h(t|X)$ 是给定协变量向量 X 下的时间 t 的风险函数。
- $h_0(t)$ 是基线风险函数，表示没有任何协变量影响下的风险。
- β_i 是与第 i 个协变量相关的系数，反映了该协变量对风险的影响程度。
- $\exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)$ 是由协变量组成的指数项，称为风险比 (Hazard Ratio, HR)。如果 $\beta_i > 0$ ，则表示相应的协变量会增加风险；如果 $\beta_i < 0$ ，则表示它会降低风险。

简化表达式：

$$\ln(h(t, X)) = \ln(h_0(t)) + (\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m)$$

进一步简化：

$$\ln \left[\frac{h(t, X)}{h_0(t)} \right] = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m$$

每个 β_i 表示当其他条件不变时，协变量 X_i 每单位变化对风险的对数变化量。

Cox模型的注意事项

比例风险假设

Cox模型假设所有个体的风险比率在整个观察期间保持恒定。如果这一假设不成立，模型的估计可能不准确。

线性关系

对于连续型自变量，Cox模型假设它们与对数风险比之间存在线性关系。

无信息删失

Cox模型要求删失数据应是随机的，即删失的发生与个体的生存时间无关。

独立性假设

观察样本之间数据应相互独立，即每个个体不受其他个体的影响。

样本量

模型需要足够的样本量以得到可靠的参数估计。

Cox模型分析实践数据

A、B两种治疗方案（简称A组、B组）分别治疗某恶性肿瘤患者25人和22人，患者基本情况包括性别（男性=1，女性=2）、年龄（岁）和体重（kg），病情包括严重程度分级（分为1、2、3级）、是否转移（转移=1，未转移=0）。随访记录患者的生存时间（月）和结局（死亡=1，删失=0）。

编号	疗法	性别	年龄	体重	分级	是否转移	生存时间	结局
1	A组	2	42	54.6	3	1	10	1
2	A组	1	65	83.7	3	1	2	0
3	A组	2	45	59.5	3	1	12	1
4	A组	2	49	44.9	3	1	13	1
5	A组	1	46	62.8	2	0	18	1

Cox模型分析实践任务

以性别、年龄、疗法等因素为协变量，采用Cox回归模型分析患者的生存时间及状态的影响因素（**解释**）。

Cox模型分析实践程序

```
import pandas as pd
from lifelines import CoxPHFitter # 从lifelines库导入CoxPHFitter方法
data=pd.read_excel('./PyData2403/AB组COX生存分析数据.xlsx')
data.drop('编号',axis=1,inplace=True) # 删除不纳入分析的列数据
data.loc[data['疗法']=='A组','疗法']=1 # 将A组疗法赋值为1
data.loc[data['疗法']=='B组','疗法']=2 # 将B组疗法赋值为2
cph=CoxPHFitter() # 构建模型
cph.fit(data,'生存时间', event_col='结局') # 创建拟合对象，利用数据拟合
print('（1）Cox回归预分析结果：'); cph.print_summary() # 输出预分析结果
data=data[['疗法','是否转移','体重','分级','生存时间','结局']] # 根据预分析结果选取纳入分析数据
cph=CoxPHFitter()
cph.fit(data,'生存时间', event_col='结局') # 创建拟合对象，利用数据拟合
print('（2）某种恶性肿瘤患者生存时间影响因素的Cox回归分析结果：'); cph.print_summary() #
输出分析结果
```

(1) Cox模型回归预分析结果

```
(1) Cox回归预分析结果:
<lifelines.CoxPHFitter: fitted with 47 total observations, 10 right-censored observations>
    duration col = '生存时间'
    event col = '结局'
    baseline estimation = breslow
    number of observations = 47
    number of events observed = 37
    partial log-likelihood = -68.23
    time fit was run = 2024-12-06 01:14:38 UTC
```

covariate									
疗法	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%		
性别	1.26	3.54	0.39	0.50	2.03	1.64	7.60		
年龄	-0.22	0.80	0.41	-1.01	0.58	0.36	1.78		
体重	0.00	1.00	0.01	-0.02	0.03	0.98	1.03		
分级	0.03	1.04	0.02	-0.00	0.07	1.00	1.07		
是否转移	2.66	14.25	0.62	1.45	3.87	4.26	47.71		
	2.99	19.81	1.23	0.58	5.39	1.79	218.59		

covariate									
疗法	cmp to	z	p	-log2(p)					
性别	0.00	3.23	<0.005	9.68					
年龄	0.00	-0.54	0.59	0.76					
体重	0.00	0.17	0.87	0.21					
分级	0.00	1.90	0.06	4.13					
是否转移	0.00	4.31	<0.005	15.90					
	0.00	2.44	0.01	6.08					

(2) Cox模型回归分析结果

(2) 某种恶性肿瘤患者生存时间影响因素的Cox回归分析结果:

```
<lifelines.CoxPHFitter: fitted with 47 total observations, 10 right-censored observations>
    duration col = '生存时间'
    event col = '结局'
    baseline estimation = breslow
    number of observations = 47
    number of events observed = 37
    partial log-likelihood = -68.41
    time fit was run = 2024-12-06 01:14:38 UTC
```

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef)	lower 95%	exp(coef)	upper 95%		
covariate											
疗法	1.28	3.60	0.39	0.51	2.05	1.67	7.74				
是否转移	2.96	19.33	1.23	0.56	5.37	1.75	213.91				
体重	0.04	1.04	0.02	0.01	0.07	1.01	1.07				
分级	2.62	13.80	0.61	1.42	3.82	4.15	45.82				
	cmp to	z	p	-log2(p)							
covariate											
疗法	0.00	3.27	<0.005	9.87							
是否转移	0.00	2.41	0.02	5.99							
体重	0.00	2.31	0.02	5.57							
分级	0.00	4.28	<0.005	15.74							

参考文献

- 曾四清.Python卫生健康统计分析与可视化——方法与实践[M].1版.广州:中山大学出版社,2024.
- Maalouf,Maher.Logistic regression in data analysis: an overview[J].International Journal of Data Analysis Techniques & Strategies, 2011, 3(3):281-299.DOI:10.1504/IJDATS.2011.041335.
- Cox and Oakes, Analysis of Survival Data, Chapman & Hall, 1984
- Fleming and Harrington, Counting Processes and Survival Analysis, Wiley, 1991
- Carpenter M .Survival Analysis: A Self-Learning Text[J].Technometrics, 1997.DOI:10.1080/00401706.1997.10485091.
- Andersen P K .Fifty Years with the Cox Proportional Hazards Regression Model[J].Journal of the Indian Institute of Science, 2022, 102(4):1135-1144.DOI:10.1007/s41745-021-00283-9.
- Harrell F E .Cox Proportional Hazards Regression Model[J]. 2015.DOI:10.1007/978-3-319-19425-7_20.

谢谢！