数 据 挖 掘

Data Mining

主讲: 张仲楠 教授

厦門大學
XIAMEN UNIVERSITY

肆

回归分析

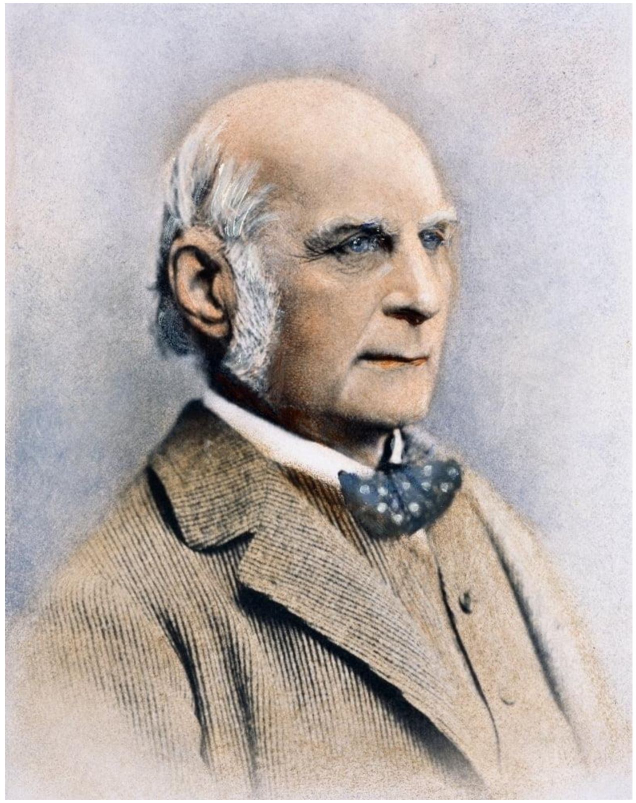
目 录

1. 基本概念

1. 基本定义

■ “回归”一词是由英国生物学家高尔顿(Francis Galton)在遗传学研究中首先提出来的。

■ 高尔顿通过对人体身高的研究发现，子女的身高不仅与父母的身高相关，而且有朝向相同性别的人的平均身高回归的趋势。



1. 基本概念

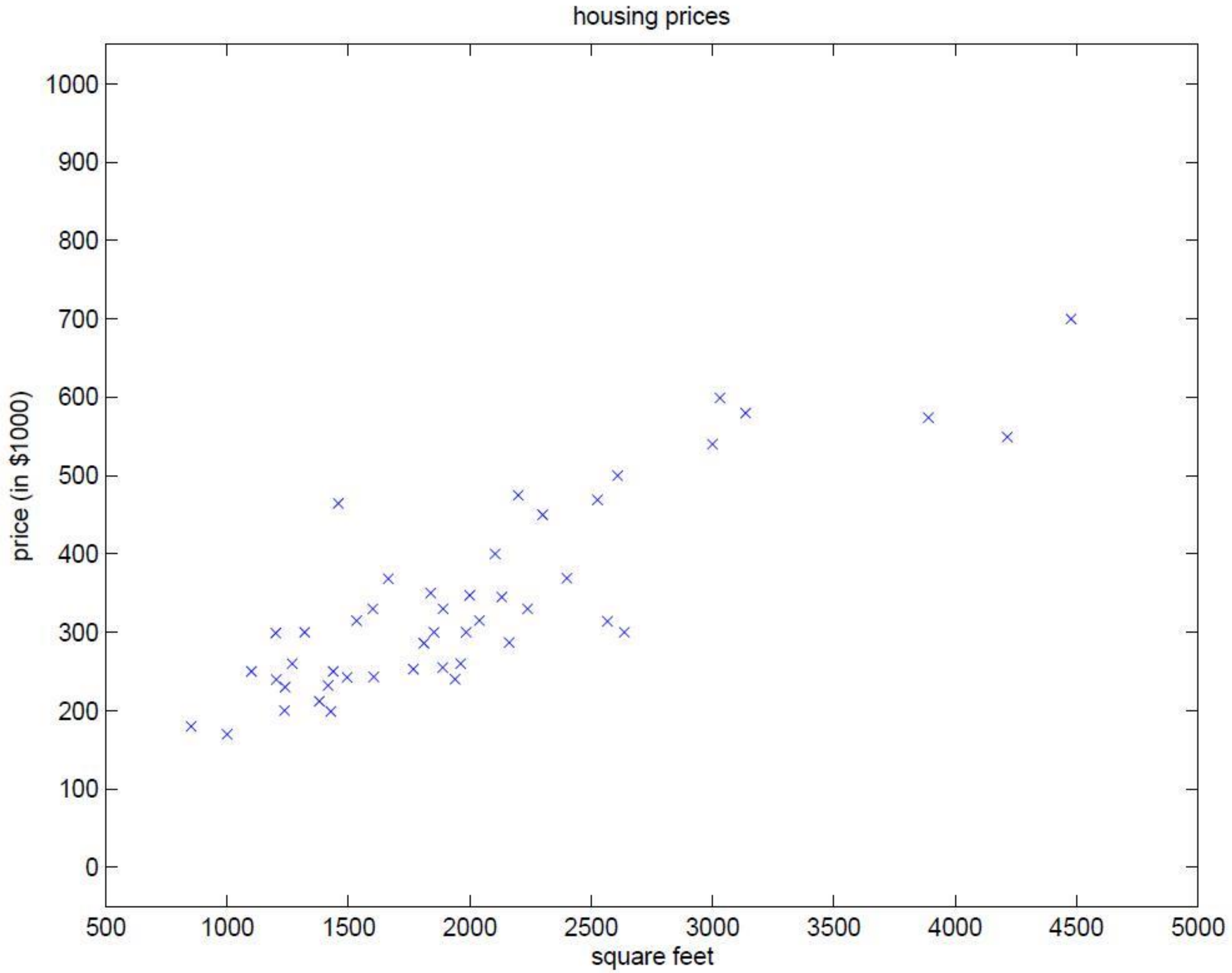
1. 基本定义

- 回归分析(regression analysis)是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。
- 作为一种预测模型，它基于观测数据建立变量间适当的依赖关系，以便分析数据间的内在规律，并应用于预测、控制等问题。

1.基本概念

2. 分析案例

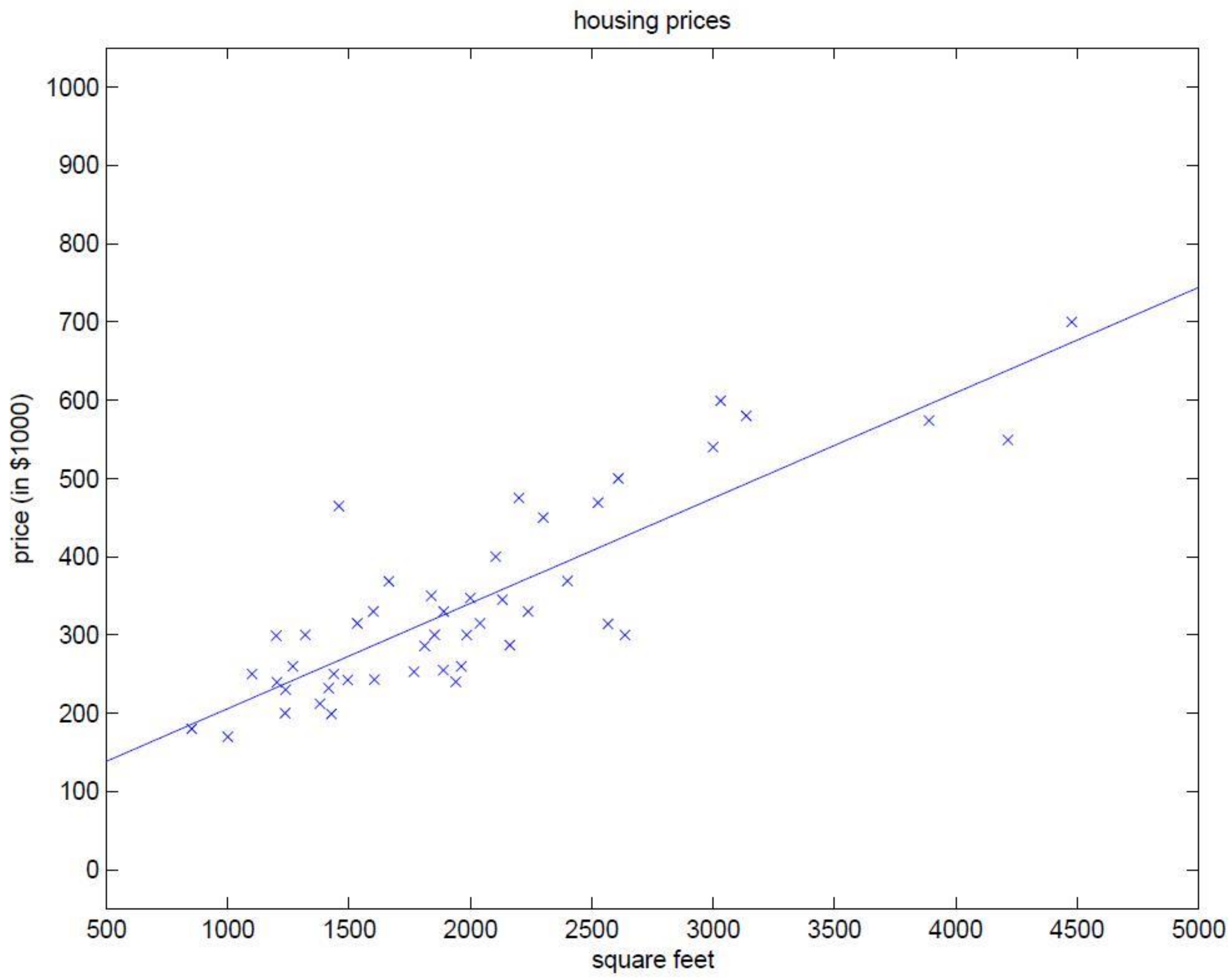
某地区的房屋面积(square feet)与价格($1000)的一个数据集，在该数据集中，只有一个自变量‘面积’和一个因变量‘价格’



1.基本概念

2. 分析案例

利用该数据集，我们可以训练一个线性方程，无限逼近所有数据点，然后利用该方程与给定的某一自变量（本例中为面积），可以预测因变量（本例中为房价）



3. 回归类别

■回归与分类均为有监督学习(supervised learning)问题，其中输入x和输出y的数值是给定的，任务是学习从输入到输出的映射。

■按照问题所涉及变量的多少，可将回归分析分为一元回归分析和多元回归分析。

■按照自变量与因变量之间是否存在线性关系，分为线性回归分析和非线性回归分析。

目 录

2.线性回归

1. 基本概念

■线性回归(Linear Regression)：因变量（y）和一个或多个自变量（x）之间建立一种线性方程关系

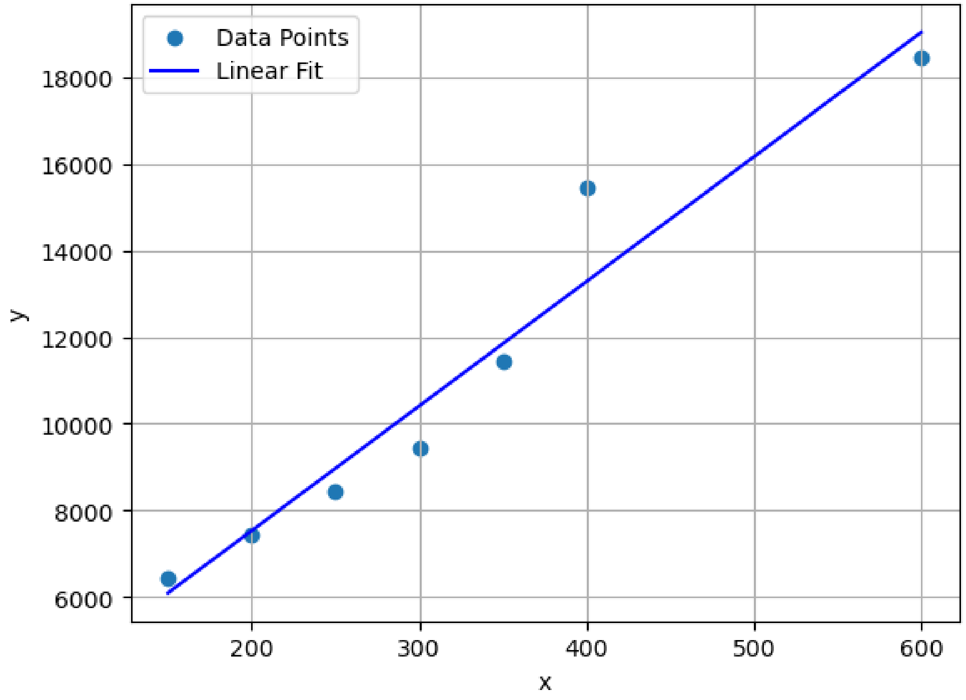
■线性回归是常用的建模技术之一，也通常是人们在学习预测模型时首选的技术之一

■在这种技术中，因变量是连续的，自变量可以是连续的也可以是离散的

2.线性回归

1. 基本概念

■ 如果在某个回归分析问题中，只有两个变量，一个自变量和一个因变量，且自变量与因变量之间的函数关系能够用一条直线来近似表示，那么称其为一元线性回归分析。



2. 一元线性回归

■一元线性回归包含一个自变量(x)和一个因变量(y)

■一元线性回归方程：

$$E(y) = \beta_0 + \beta_1 x$$

■这个方程对应的图像是一条直线，称作回归线，其中，β_0是回归线的截距，β_1是回归线的斜率，$E(y)$是在一个给定x值下y的期望值（均值）

2. 一元线性回归

■如何找出适合一元线性回归模型的最佳回归线?

■一种拟合方法：最小二乘法（least square method）

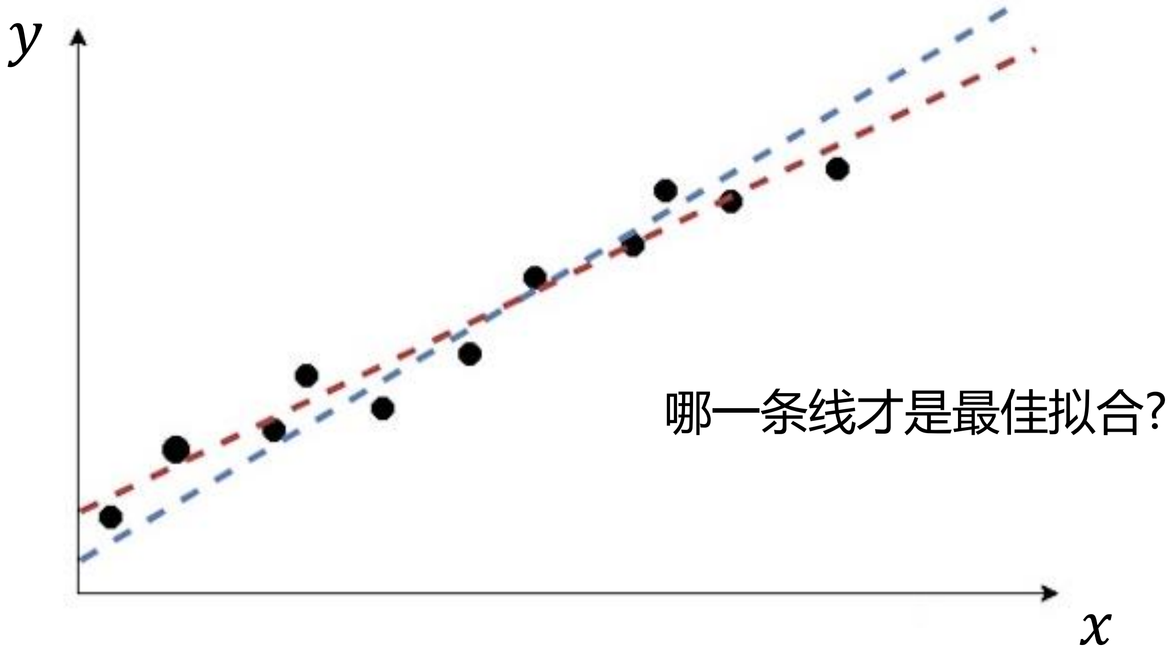
■最小二乘法的出发点是使实际测量数据y_i与拟合直线上对应的估计值$\hat{y}_i$的差(残差)的平方和为最小，即：

$$min\sum(y_i-\hat{y}_i)^2$$

2.线性回归

2. 一元线性回归

■ 法国数学家阿德里安-马里·勒让德（1752 - 1833）提出让总的误差的平方最小的y就是真值，这是基于“如果误差是随机的，应该围绕真值上下波动”



2. 一元线性回归

- 假设我们的线性方程为：$f(x) = \beta_0 + \beta_1 x$
- 样本的误差为：

$$e_i = y_i - f(x_i) = y_i - \beta_0 - \beta_1 x_i$$

- 根据最小二乘法思想，总误差为：

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- 通过最小化Q来确定直线方程，此时该问题变成了求函数极值的问题。

2. 一元线性回归

■ 求关于未知参数β_0和β_1的偏导数:

$$\begin{cases}\frac{\partial Q}{\partial \beta_0}=2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)(-1)\\ \frac{\partial Q}{\partial \beta_1}=2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)(-x_i)\end{cases}$$

■ 通过令偏导数为0，可求解极值点，即:

$$\beta_0=\frac{\sum x_i^2\sum y_i-\sum x_i\sum x_i y_i}{n\sum x_i^2-(\sum x_i)^2} \qquad \beta_1=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{n\sum x_i^2-(\sum x_i)^2}$$

■ 将样本(x_1,y_1)，$(x_2,y_2),\ldots,(x_n,y_n)$代入，即可得到参数$\beta_0$和$\beta_1$的值

2.线性回归

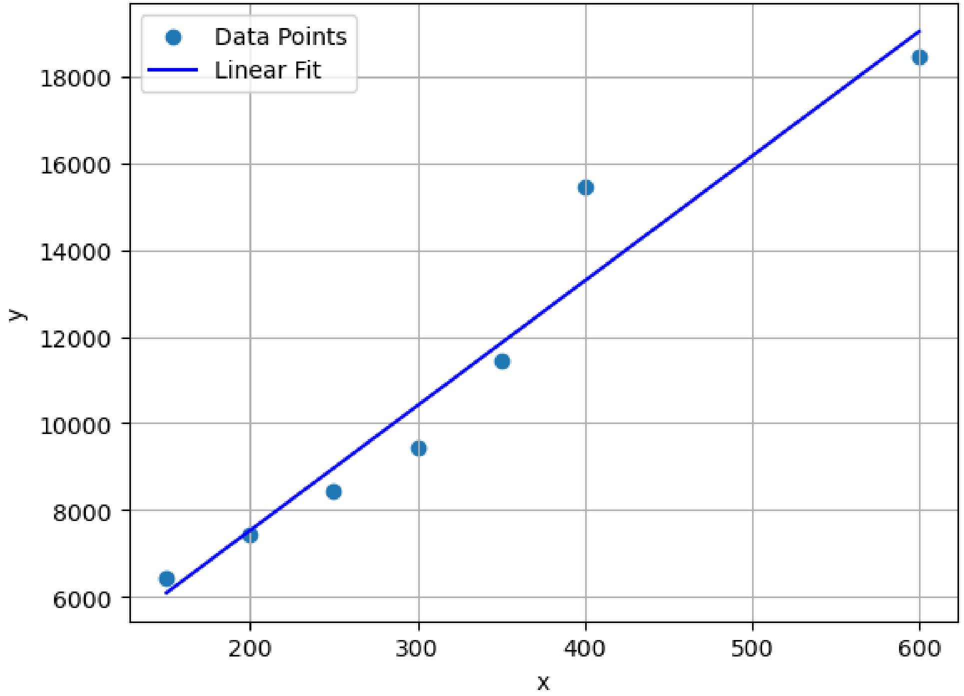
$\sum x_i$=2250 $\sum y_i$=77150 $\sum x_i^2$=857500 $\sum x_i y_i$=28662500

$$\beta_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{857500*77150-2250*28662500}{7*857500-2250*2250} = 1771.80851$$

2. 一元线性回归

$$\beta_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{7*28662500-2250*77150}{7*857500-2250*2250} = 28.77659574$$

x	y
150	6450
200	7450
250	8450
300	9450
350	11450
400	15450
600	18450



$$f(x) = 1771.81 + 28.78x$$

2.线性回归

2. 多元线性回归

- 给定数据集$D=\{(\boldsymbol{x_1},y_1),(\boldsymbol{x_2},y_2),\cdots,(\boldsymbol{x_m},y_m)\}$, 其中$\boldsymbol{x_i}$由 d 个属性描述，表示为$\boldsymbol{x_i}=(x_{i1},x_{i2},\cdots,x_{id})$， x_{ij} 是$\boldsymbol{x_i}$在第 j 个属性X_j上的取值，$y_i \epsilon \mathbb{R}$
- 线性模型(linear model)试图学得一个通过属性的线性组合来进行预测的函数，即

$$f(X)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_dX_d$$

- $\beta_1,\beta_2,\cdots,\beta_d$称为偏回归系数， β_i的意义是当其他自变量$X_j(j\neq i)$都固定时，自变量X_i每变化一个单位而使因变量平均改变的数值

2. 多元线性回归

- 样本的误差为：

$$e_i = y_i - f(x_i) = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_d x_{id}$$

- 用矩阵表示：

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\substack{\boldsymbol{Y} \\ n \times 1}} = \underbrace{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & & x_{2d} \\ & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix}}_{\substack{\boldsymbol{X} \\ n \times (d+1)}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}}_{\substack{\boldsymbol{\beta} \\ (d+1) \times 1}} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}}_{\substack{\boldsymbol{e} \\ n \times 1}}$$

2. 多元线性回归

- 样本的误差为：

$$e_i = y_i - f(x_i) = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_d x_{id}$$

- 根据最小二乘法思想，总误差为：

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_d x_{id})^2$$

- 通过最小化Q来确定直线方程，求关于未知参数$\beta_0,\beta_1,\cdots,\beta_d$的偏导数，并令它们为0

$$\frac{\partial Q}{\partial \beta_j} = 0 (0 \le j \le d)$$

2. 多元线性回归

$$\begin{cases} \frac{\partial Q}{\partial \beta_0}=2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_{i1}-\beta_2 x_{i2}-\cdots-\beta_d x_{id})(-1)=0 \\ \frac{\partial Q}{\partial \beta_j}=2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_{i1}-\beta_2 x_{i2}-\cdots-\beta_d x_{id})(-x_{ij})=0 \quad (1\le j\le d) \end{cases}$$

$\Rightarrow \Sigma e_i = 0$

$\Rightarrow \Sigma x_{ij} e_i = 0 \quad (1 \le j \le d)$

- 对上述方程整理后，可以得到：

$$\begin{cases} n\beta_0 + \beta_1 \sum x_{i1} + \beta_2 \sum x_{i2} + \cdots + \beta_d \sum x_{id} = \sum y_i \\ \beta_0 \sum x_{ij} + \beta_1 \sum x_{i1} x_{ij} + \beta_2 \sum x_{i2}\, x_{ij} + \cdots + \beta_d \sum x_{id}\, x_{ij} = \sum y_i x_{ij} \; (1 \le j \le d) \end{cases}$$

- 参数的最小二乘估计量为: $\beta = (X^T X)^{-1} X^T Y$

2.线性回归

2. 多元线性回归

■ 一家快递公司送货：X1:运输里程，X2:运输次数，Y:总运输时间

任务	X1	X2	Y
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

问题：如果一个运输任务是跑102公里，运输6次，预计时间是多长?

2.线性回归

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d$$

$$\beta = (X^T X)^{-1} X^T Y$$

=[-0.86870147 0.0611346 0.92342537]

2. 多元线性回归

■ 一家快递公司送货：X1:运输里程，X2:运输次数，Y:总运输时间

任务	X1	X2	Y
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

Y = -0.869 + 0.0611*X1 + 0.923*X2

当X1=102, X2=6, Y=10.9

$$\underset{X}{\begin{bmatrix} 1 & 100 & 4 \\ 1 & 50 & 3 \\ 1 & 100 & 4 \\ 1 & 100 & 2 \\ 1 & 50 & 2 \\ 1 & 80 & 2 \\ 1 & 75 & 3 \\ 1 & 65 & 4 \\ 1 & 90 & 3 \\ 1 & 90 & 2 \end{bmatrix}} \qquad \underset{Y}{\begin{bmatrix} 9.3 \\ 4.8 \\ 8.9 \\ 6.5 \\ 4.2 \\ 6.2 \\ 7.4 \\ 6.0 \\ 7.6 \\ 6.1 \end{bmatrix}}$$

目 录

3.非线性回归

1. 基本概念

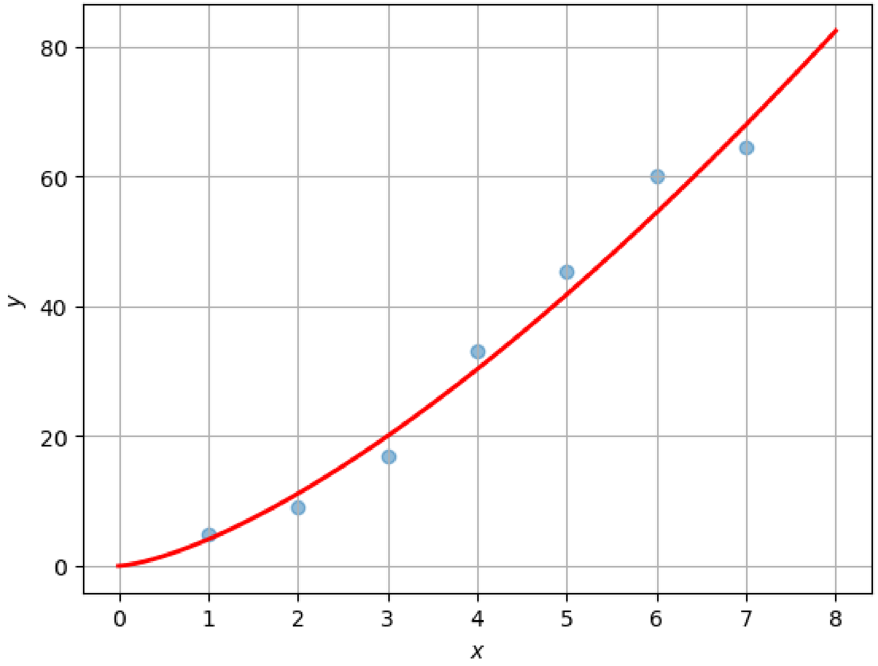
- 两个现象变量之间的相关关系并非线性关系，而呈现某种非线性的曲线关系，如：双曲线、二次曲线、三次曲线、幂函数曲线、指数函数曲线、S型曲线、对数曲线等。
- 对于非线性回归问题，常采用适当的变量代换，把问题转化为线性回归问题，求出线性回归模型后代回，得到非线性回归方程。

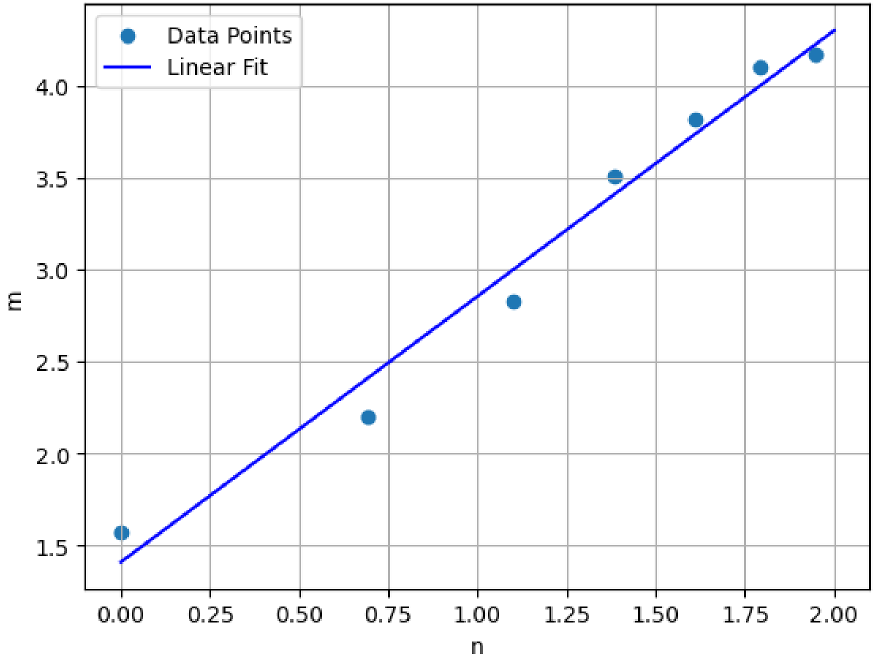
3.非线性回归

x	1	2	3	4	5	6	7
y	4.82	9.0	16.894	33.204	45.338	60.114	64.549
$\ln x\,(n)$	0	0.693	1.099	1.386	1.609	1.792	1.946
$\ln y(m)$	1.573	2.197	2.827	3.503	3.814	4.096	4.167

2. 幂函数型

- 样本点分布在某幂函数曲线$y = c_1 x^{c_2}$的周围，其中c_1, c_2是待定参数。
- 变量代换:令$m = \ln y$，$n = \ln x$，变换后样本点应该分布在直线 $m = an + b$的周围,其中$a = c_2$，$b = \ln c_1$





$$m = \underbrace{\boxed{1.444}}_{c_2} n + \underbrace{\boxed{1.410}}_{\ln c_1}$$

$$c_1 = e^{1.410} = 4.096$$

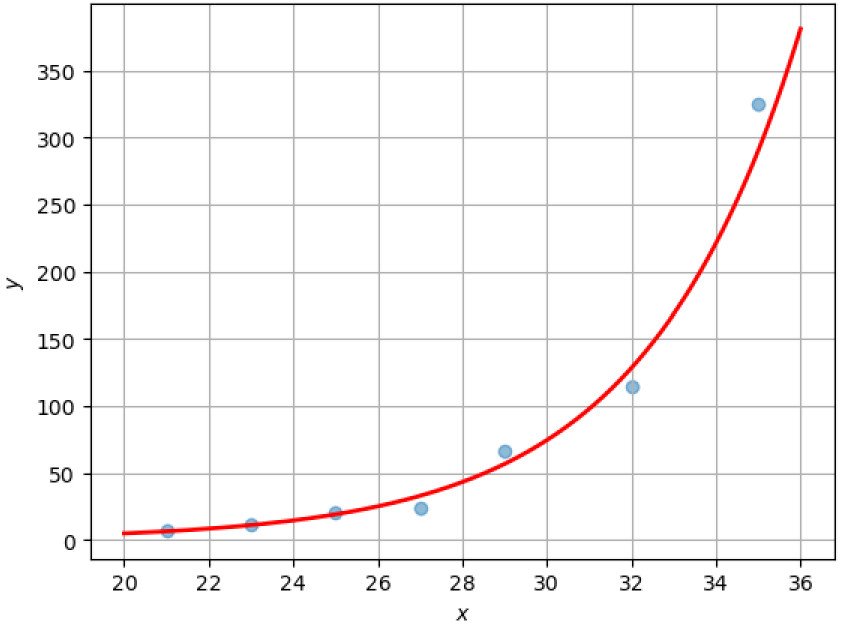
$$y = 4.096x^{1.444}$$

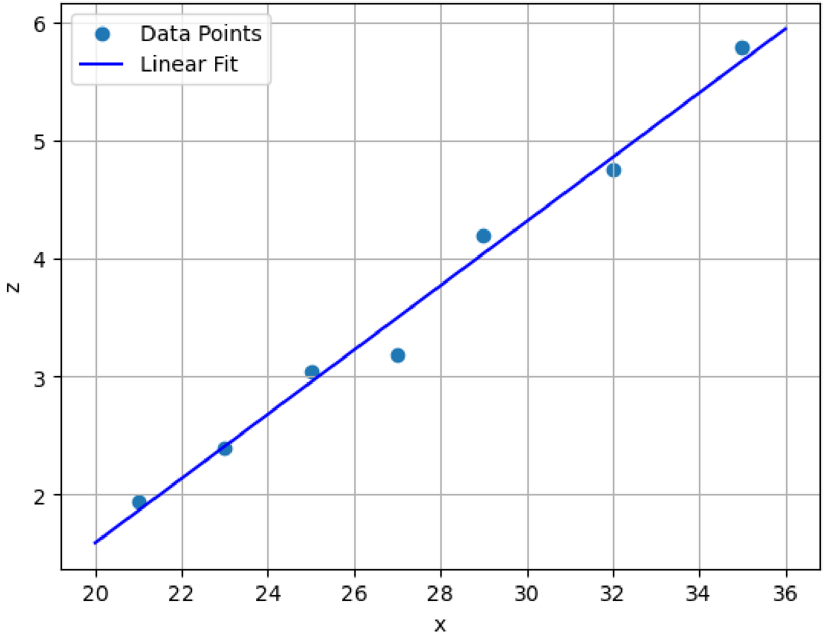
3.非线性回归

x	21	23	25	27	29	32	35
y	7	11	21	24	66	115	325
$\ln y\ (z)$	1.946	2.398	3.045	3.178	4.190	4.745	5.784

3. 指数函数型

- 样本点分布在某一条指数函数曲线$y = c_3 e^{c_4 x}$的周围，其中c_3, c_4是待定参数。
- 变量代换:令$z = \ln y$，变换后样本点应该分布在直线 $z = ax + b$的周围，其中$a = c_4$，$b = \ln c_3$

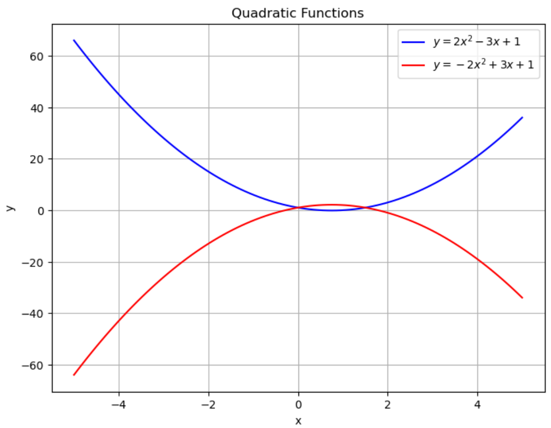


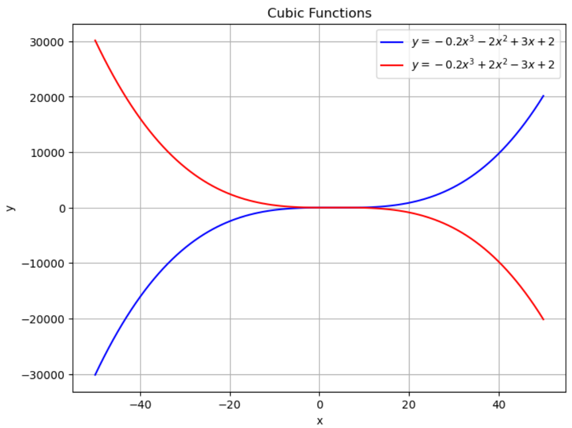


$$z = \underbrace{\boxed{0.272}}_{c_4}x \underbrace{\boxed{-\ 3.849}}_{\ln c_3}$$

$$y = e^{0.272x - 3.849}$$

3.非线性回归





4. 多项式函数型

- 进行多项式回归分析，首先要确定多项式的次数。
- 次数的确定一般是根据经验和实验。
- 假设确定了用一个一元k次多项式来拟合训练样本集。
- 多项式函数表示为$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$，令$x_1 = x$，$x_2 = x^2, \ldots$，$x_k = x^k$，则该式可化为：$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$。
- 后续采用类似于多元线性回归的方法求解各β值。
- 多项式回归的优势在于它可以拟合出各种形状的曲线。然而，随着多项式阶数的增加，模型的复杂度也会增加，可能会导致过拟合的问题。

4. 多项式函数型

■ 假设有以下数据集，表示某产品的广告支出（x）与销售额（y）的关系：

广告支出 (x)	销售额 (y)
1	1
2	2.5
3	3.5
4	5
5	7
6	8.5
7	10
8	12.5

用二次多项式（k=2）来拟合这组数据

3.非线性回归

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

4. 多项式函数型

■ 假设有以下数据集，表示某产品的广告支出（x）与销售额（y）的关系：

广告支出 (x)	销售额 (y)
1	1
2	2.5
3	3.5
4	5
5	7
6	8.5
7	10
8	12.5

$$X = \begin{bmatrix} 1 & 1 & 1^2 \\ 1 & 2 & 2^2 \\ 1 & 3 & 3^2 \\ 1 & 4 & 4^2 \\ 1 & 5 & 5^2 \\ 1 & 6 & 6^2 \\ 1 & 7 & 7^2 \\ 1 & 8 & 8^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 7 & 49 \\ 1 & 8 & 64 \end{bmatrix} \qquad Y = \begin{bmatrix} 1 \\ 2.5 \\ 3.5 \\ 5 \\ 7 \\ 8.5 \\ 10 \\ 12.5 \end{bmatrix}$$

$$\beta = (X^T X)^{-1} X^T Y$$

=[0 1.01785714 0.06547619]

目 录

4.逻辑回归

1. 基本概念

■ logistic 回归(logistic regression)：一种概率判别模型，它直接利用其属性值来估计数据实例$\boldsymbol{x}$的概率

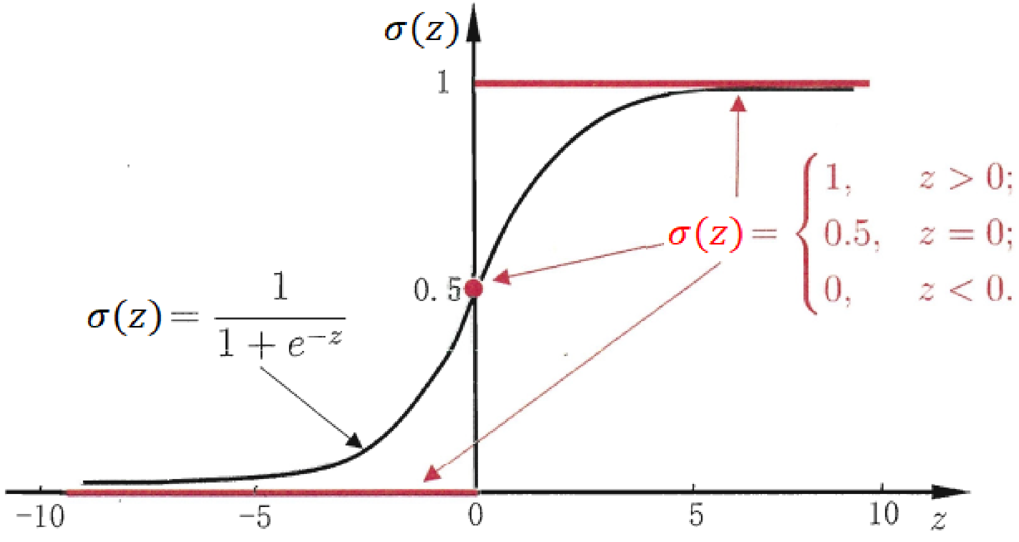
■ logistic 回归的基本思想是使用线性预测器$z = \boldsymbol{w}^T\boldsymbol{x} + b$表示 $\boldsymbol{x}$的概率：

比率值(odds) ⟹ $$\frac{P(y=1|\boldsymbol{x})}{P(y=0|\boldsymbol{x})}=e^{z}=e^{\boldsymbol{w}^T\boldsymbol{x}+b}$$

■ 如果 $\boldsymbol{w}^T\boldsymbol{x} + b$ >0, 那么 $\boldsymbol{x}$ 属于第 1 类，因为它的比率值大于 1；否则，$\boldsymbol{x}$ 属于第 0 类

4.逻辑回归

1. 基本概念



- 由于 $P(y=0|\boldsymbol{x}) + P(y=1|\boldsymbol{x}) = 1$,

  可以重写: $\frac{P(y=1|\boldsymbol{x})}{1-P(y=1|\boldsymbol{x})} = e^z$

- 进一步简化，将 $P(y=1|\boldsymbol{x})$表示为 z 的函数：$P(y=1|\boldsymbol{x}) = \frac{1}{1+e^{-z}} = \sigma(z)$
- 函数$\sigma(\cdot)$称为 logistic 或者 S 形函数(Sigmoid function)
- $P(y=0|\boldsymbol{x})$可以表示为：$P(y=0|\boldsymbol{x}) = 1 - \sigma(z) = \frac{1}{1+e^{z}}$
- 如果知道了参数$\boldsymbol{w}$和b的合适值，可以用上式来估计任何数据实例 $\boldsymbol{x}$的后验概率，并确定其类别标签

4.逻辑回归

2. 学习模型参数

- logistic 回归的参数 $(\boldsymbol{w}, b)$ 是在训练过程中使用最大似然估计法来估计的。这种方法需要计算观察给定$(\boldsymbol{w}, b)$的训练数据的可能性，然后确定最大似然下的模型参数$(\boldsymbol{w}^*, b^*)$。
- 让$D.train = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)\}$,表示一组$n$个训练实例，其中$y_i$是一个二分类变量(0或1)。

2. 学习模型参数

- 给定$\boldsymbol{x}_i$，$\boldsymbol{w}$和b，可以表示观察到y_i的可能性：

$$P(y_i|\boldsymbol{x}_i,\ \boldsymbol{w},\ b) = P(y=1|\boldsymbol{x}_i)^{y_i} \times P(y=0|\boldsymbol{x}_i)^{1-y_i}$$

$$= \left(\sigma\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right)\right)^{y_i} \times \left(1 - \sigma\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right)\right)^{1-y_i}$$

- 所有训练实例$\mathcal{L}(\boldsymbol{w},\ b)$的可能性可以通过取单个似然积(假设训练实例中的独立性)来计算:
- $\mathcal{L}(\boldsymbol{w},\ b) = \prod_{i=1}^{n} P(y_i|\boldsymbol{x}_i,\ \boldsymbol{w},\ b) = \prod_{i=1}^{n} P(y=1|\boldsymbol{x}_i)^{y_i} \times P(y=0|\boldsymbol{x}_i)^{1-y_i}$

4.逻辑回归

2. 学习模型参数

■ 考虑似然函数的负对数(以e为底)，也称为交叉熵：

$$-log\mathcal{L}(\boldsymbol{w},\ b) = -\sum_{i=1}^{n} y_i logP(y = 1|\boldsymbol{x_i}) + (1 - y_i)\, logP(y = 0|\boldsymbol{x_i})$$

$$= -\sum_{i=1}^{n} y_i log(\sigma(\boldsymbol{w}^T\boldsymbol{x_i} + b)) + (1 - y_i)\, log(1 - \sigma(\boldsymbol{w}^T\boldsymbol{x_i} + b))$$

■ 我们想找到模型参数$(\boldsymbol{w}^*, b^*)$使得交叉熵$-log\mathcal{L}(\boldsymbol{w}^*, b^*)$最小:

$$(\boldsymbol{w}^*, b^*) = \underset{(\boldsymbol{w},\ b)}{\operatorname{argmin}} -log\mathcal{L}(\boldsymbol{w},\ b)$$

3. logistic回归模型的特点

- 1) logistic回归是一种用来直接计算概率的判别模型，它不做任何关于条件概率的假设。因此，它是相当通用的，可以应用于不同的应用程序。它也可以轻松地扩展到多分类，那时，它被称为多项式logistic回归(multinomial logistic regression)。然而，它的表达能力仅限于学习线性决策边界。
- 2)因为每个属性都有不同的权重(参数)，因此可以分析logistic回归的学习参数来理解属性和类别标签之间的关系。

4. logistic回归例题

■ 有如下的关于客户年龄、收入和是否购买产品的数据：

客户	年龄	收入	是否购买（0=不购买，1=购买）
1	25	50000	0
2	30	60000	1
3	22	45000	0
4	35	70000	1
5	28	52000	1
6	40	80000	1
7	23	48000	0

■ 逻辑回归模型的形式为：

$$P(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1\cdot\text{年龄}+\beta_2\cdot\text{收入})}}$$

初始参数：

$$\beta_0=-10$$

$$\beta_1=0.1$$

$$\beta_2=0.0001$$

4. logistic回归例题

■ 有如下的关于客户年龄、收入和是否购买产品的数据：

客户	年龄	收入	是否购买（0=不购买，1=购买）
1	25	50000	0
2	30	60000	1
3	22	45000	0
4	35	70000	1
5	28	52000	1
6	40	80000	1
7	23	48000	0

■ 逐个计算每个客户的预测概率：

1. **客户 1**: 年龄=25，收入=50000，是否购买=0
   - $z = -10 + 0.1 \cdot 25 + 0.0001 \cdot 50000$
   - $z = -10 + 2.5 + 5 = -2.5$
   - $P(Y = 1|X) = \frac{1}{1+e^{2.5}} \approx \frac{1}{1+12.182} \approx 0.076$
2. **客户 2**: 年龄=30，收入=60000，是否购买=1
   - $z = -10 + 0.1 \cdot 30 + 0.0001 \cdot 60000$
   - $z = -10 + 3 + 6 = -1$
   - $P(Y = 1|X) = \frac{1}{1+e^{1}} \approx \frac{1}{1+2.718} \approx 0.268$

4. logistic回归例题

■ 有如下的关于客户年龄、收入和是否购买产品的数据：

客户	年龄	收入	是否购买（0=不购买，1=购买）
1	25	50000	0
2	30	60000	1
3	22	45000	0
4	35	70000	1
5	28	52000	1
6	40	80000	1
7	23	48000	0

■ 逐个计算每个客户的预测概率：

3. **客户 3**: 年龄=22，收入=45000，是否购买=0
   - $z = -10 + 0.1 \cdot 22 + 0.0001 \cdot 45000$
   - $z = -10 + 2.2 + 4.5 = -3.3$
   - $P(Y = 1|X) = \frac{1}{1+e^{3.3}} \approx \frac{1}{1+27.14} \approx 0.036$
4. **客户 4**: 年龄=35，收入=70000，是否购买=1
   - $z = -10 + 0.1 \cdot 35 + 0.0001 \cdot 70000$
   - $z = -10 + 3.5 + 7 = 0.5$
   - $P(Y = 1|X) = \frac{1}{1+e^{-0.5}} \approx \frac{1}{1+0.6065} \approx 0.623$

4. logistic回归例题

■ 有如下的关于客户年龄、收入和是否购买产品的数据：

客户	年龄	收入	是否购买（0=不购买，1=购买）
1	25	50000	0
2	30	60000	1
3	22	45000	0
4	35	70000	1
5	28	52000	1
6	40	80000	1
7	23	48000	0

■ 逐个计算每个客户的预测概率：

5. **客户 5**: 年龄=28，收入=52000，是否购买=1
   - $z = -10 + 0.1 \cdot 28 + 0.0001 \cdot 52000$
   - $z = -10 + 2.8 + 5.2 = -2$
   - $P(Y = 1|X) = \frac{1}{1+e^2} \approx \frac{1}{1+7.389} \approx 0.119$
6. **客户 6**: 年龄=40，收入=80000，是否购买=1
   - $z = -10 + 0.1 \cdot 40 + 0.0001 \cdot 80000$
   - $z = -10 + 4 + 8 = 2$
   - $P(Y = 1|X) = \frac{1}{1+e^{-2}} \approx \frac{1}{1+0.1353} \approx 0.881$

4. logistic回归例题

■ 有如下的关于客户年龄、收入和是否购买产品的数据：

客户	年龄	收入	是否购买（0=不购买，1=购买）
1	25	50000	0
2	30	60000	1
3	22	45000	0
4	35	70000	1
5	28	52000	1
6	40	80000	1
7	23	48000	0

■ 逐个计算每个客户的预测概率：

7. **客户 7**: 年龄=23，收入=48000，是否购买=0
    - $z = -10 + 0.1 \cdot 23 + 0.0001 \cdot 48000$
    - $z = -10 + 2.3 + 4.8 = -3.9$
    - $P(Y = 1|X) = \frac{1}{1+e^{3.9}} \approx \frac{1}{1+49.57} \approx 0.020$

4. logistic回归例题

■ 有如下的关于客户年龄、收入和是否购买产品的数据：

客户	年龄	收入	是否购买（0=不购买，1=购买）	预测概率
1	25	50000	0	0.076
2	30	60000	1	0.268
3	22	45000	0	0.036
4	35	70000	1	0.623
5	28	52000	1	0.119
6	40	80000	1	0.881
7	23	48000	0	0.020

结果分析：

- 客户6的预测概率最高(0.881)，符合其购买情况
- 客户 4 的预测概率为 0.623，也表明可能会购买
- 客户1和客户 3 的预测概率较低，表明不太可能购买，符合实际情况
- 客户2和客户 5 的预测概率较低，表明不太可能购买，不符合实际情况

4. logistic回归例题

■ 有如下的关于客户年龄、收入和是否购买产品的数据：

客户	年龄	收入	是否购买（0=不购买，1=购买）	预测概率
1	25	50000	0	0.377
2	30	60000	1	0.802
3	22	45000	0	0.176
4	35	70000	1	0.964
5	28	52000	1	0.560
6	40	80000	1	0.994
7	23	48000	0	0.256

将β_1调整到0.18，结果分析：

- 客户 2、4、5 和 6 的预测概率较高，表明他们可能会购买，预测合理。
- 客户 1、3 和 7 的预测概率较低，符合他们没有购买的实际情况。