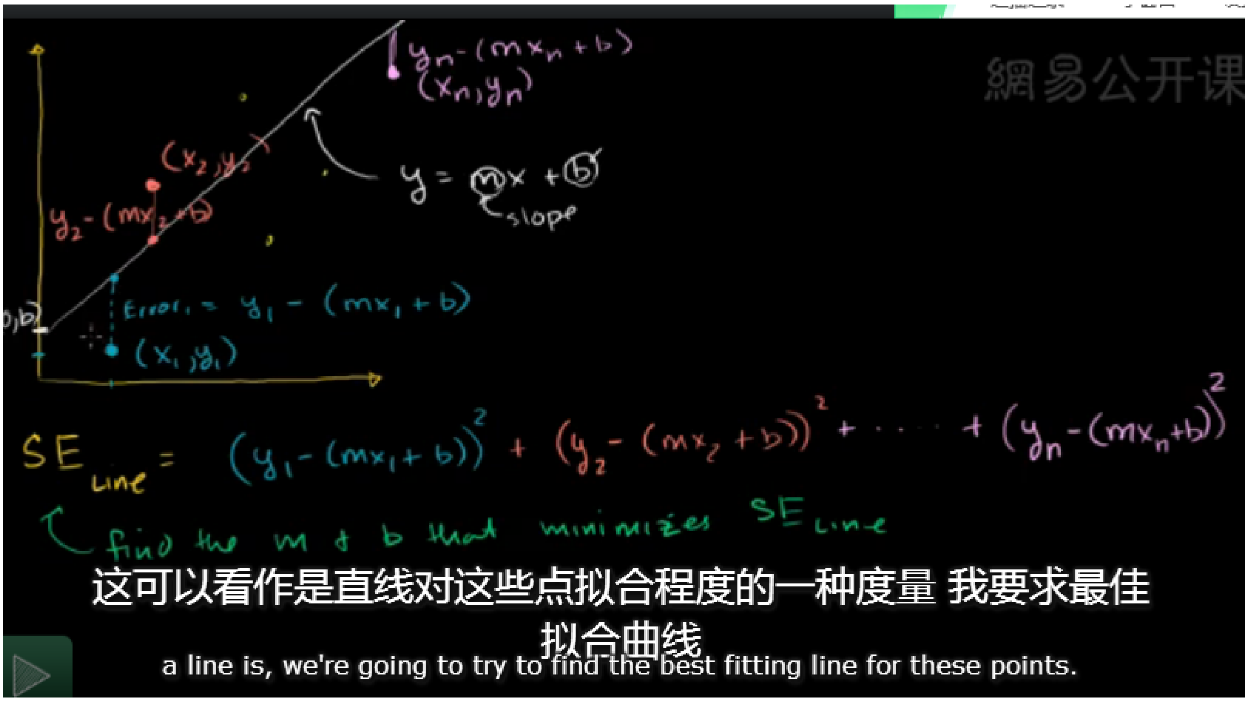

62课

线性回归中的平方误差

线性回归是利用最小平方误差对自变量和因变量之间关系进行建模的一种回归分析。这一节介绍了平方误差的概念，并解释了直线拟合中最基本的原理



63课

线性回归公式的推导1

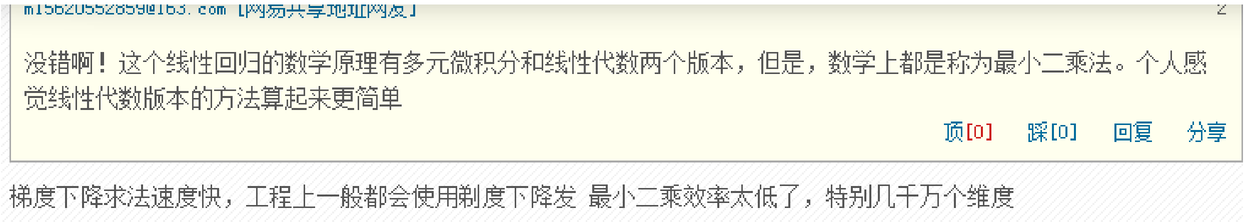
线性回归是利用最小平方误差对自变量和因变量之间关系进行建模的一种回归分析。这一节开始推导线性回归的公式，最佳拟合曲线为y=mx+b，其中m=(x均值·y均值-xy均值)/[(x均值)2-x2均值]，b=ȳ-mx̄。由于推导过程较长，所以分成了四个部分，这是第一部分，进行最初步的代数运算。

64课

线性回归公式的推导2

进行第二步代数运算，并将式子同三维空间的二次曲面联系起来。

three-dimensional parapola


m15620552859@163.com [网易共享地址网友]

没错啊！这个线性回归的数学原理有多元微积分和线性代数两个版本，但是，数学上都是称为最小二乘法。个人感觉线性代数版本的方法算起来更简单

顶[0] 踩[0] 回复 分享

梯度下降求法速度快，工程上一般都会使用剃度下降发 最小二乘效率太低了，特别几千万个维度

65课

线性回归公式的推导3

这是第三部分，利用微积分中的基本偏导知识进行推导，并列出方程。

66课

线性回归公式的推导4

这是第四部分，解出方程，并给出最后结果。

67课

线性回归例题

1,2)、(2,1)、(4,

3)三点如何进行线性回归，这一节利用公式求出了与这三点拟合最好的直线

coliner 共线

通过最小 SE_{line} 求出 m 与截距 b.

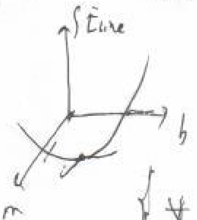
$$SE_{line} = \sum_{i}^{n} [y_i^2 - 2y_i(mx_i+b) + (mx_i+b)^2]$$

$$= \sum_{i=0}^{n} [y_i^2 - 2mx_iy_i - 2by_i + m^2x_i^2 + \cancel{2mx_ib}\ 2mbx_i + b^2]$$ ⇒ step2

$= \sum_{i=0}^{n}$ { 全部转化为均值

$$= n\overline{y^2} - 2mn\overline{xy} - 2nb\bar{y} + m^2n\overline{x^2} + 2mnb\bar{x} + nb^2$$

next: Optimize this! : minimize this expression ⇒ 分别对m和b求偏导



(1) 目标 $\frac{\partial SE}{\partial m} = 0$ and $\frac{\partial SE}{\partial b} = 0$

$(-2n\overline{xy} + 2mn\overline{x^2} + 2nb\bar{x}) = 0$ $\cancel{(\overline{y^2} - 2n\overline{xy} - 2b\bar{y} + m^2\overline{x^2} + 2mb\bar{x})}$

同理

$(-2n\bar{y} + 2mn\bar{x} + 2nb) = 0$

除以2n

$$\begin{cases} -\overline{xy} + m\overline{x^2} + b\bar{x} = 0 \\ -\bar{y} + (m\bar{x} + b) = 0 \end{cases} \Rightarrow \begin{cases} m\overline{x^2} + b\bar{x} = \overline{xy} & ① \\ m\bar{x} + b = \bar{y} & ② \end{cases}$$

均值 $(\bar{x}, \bar{y})$ 在拟合最优曲线

方程①: $m\frac{\overline{x^2}}{\bar{x}} + b = \frac{\overline{xy}}{\bar{x}}$ ③ 即 $\left(\frac{\overline{x^2}}{\bar{x}}, \frac{\overline{xy}}{\bar{x}}\right)$ 也在最优曲线上. 两点一线

结果: 方程③-②: $m\frac{\overline{x^2}}{\bar{x}} - m\bar{x} = \frac{\overline{xy}}{\bar{x}} - \bar{y}$ ⇒ $m = \dfrac{\frac{\overline{xy}}{\bar{x}} - \bar{y}}{\frac{\overline{x^2}}{\bar{x}} - \bar{x}} = \dfrac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \dfrac{\bar{x}\bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}$

$b = \bar{y} - \bar{x} \times m = \bar{y} - \dfrac{\overline{xy} - \bar{x}\bar{y}}{\frac{\overline{x^2}}{\bar{x}} - \bar{x}}$

(1,2) (2,1) (4,3) 的线性回归

$\overline{xy} = \frac{2+2+12}{3} = \frac{16}{3}$　$\overline{x^2} = \frac{1+4+16}{3} = 7$

$\bar{x} = \frac{1+2+4}{3} = \frac{7}{3}$

$\bar{y} = \frac{2+1+3}{3} = 2$

则 $m = \dfrac{\bar{x}\bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}} = \dfrac{\frac{7}{3}\times 2 - \frac{16}{3}}{\frac{49}{9} - 7} = \dfrac{\frac{2}{3}}{\frac{14}{9}} = \frac{3}{7}$

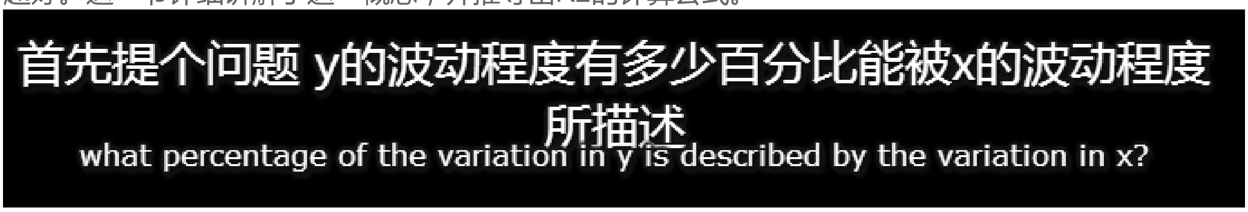
$b = \bar{y} - \bar{x}m = 2 - \frac{7}{3} \times \frac{3}{7} = 1$

则 回归曲线 $y = \frac{3}{7}x + 1$

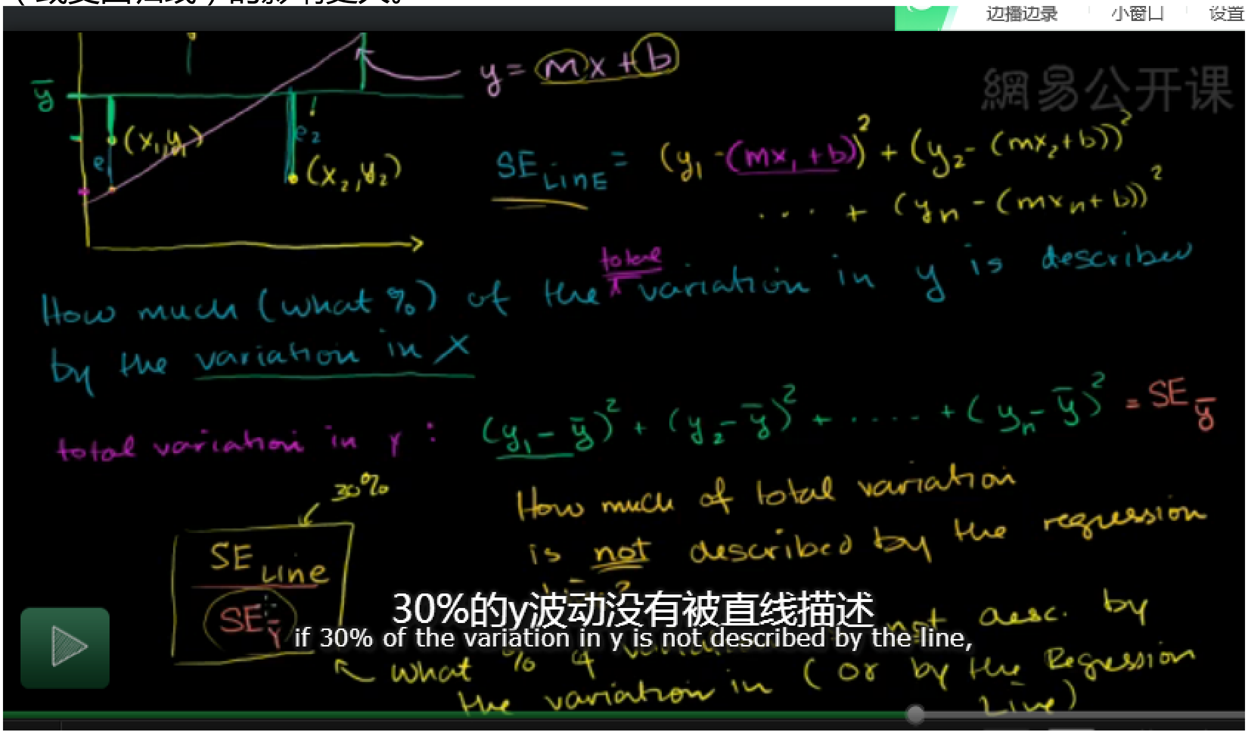
70课

决定系数R2

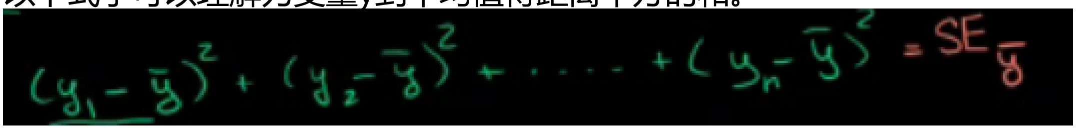
决定系数R2，是指y的总波动情况中，可以以直线关系说明的部分所占的比率。R2越大，表示直线拟合得越好。这一节详细讲解了这一概念，并推导出R2的计算公式。



首先搞明白变量y波动的量化方式。
决定系数r^2 = 1 - 未被回归线拟合的损失/y的波动值。
y的波动值= y- y的平均值的方差。未拟合的损失越小、决定系数越大，则说明y受x的影响（或受回归线）的影响更大。



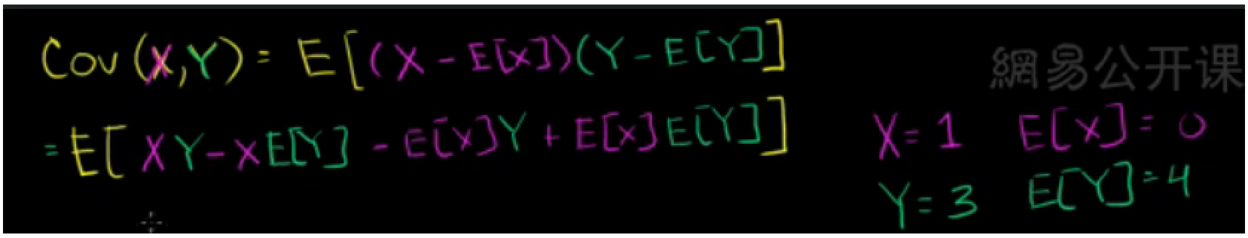
以下式子可以理解为变量y到平均值得距离平方的和。

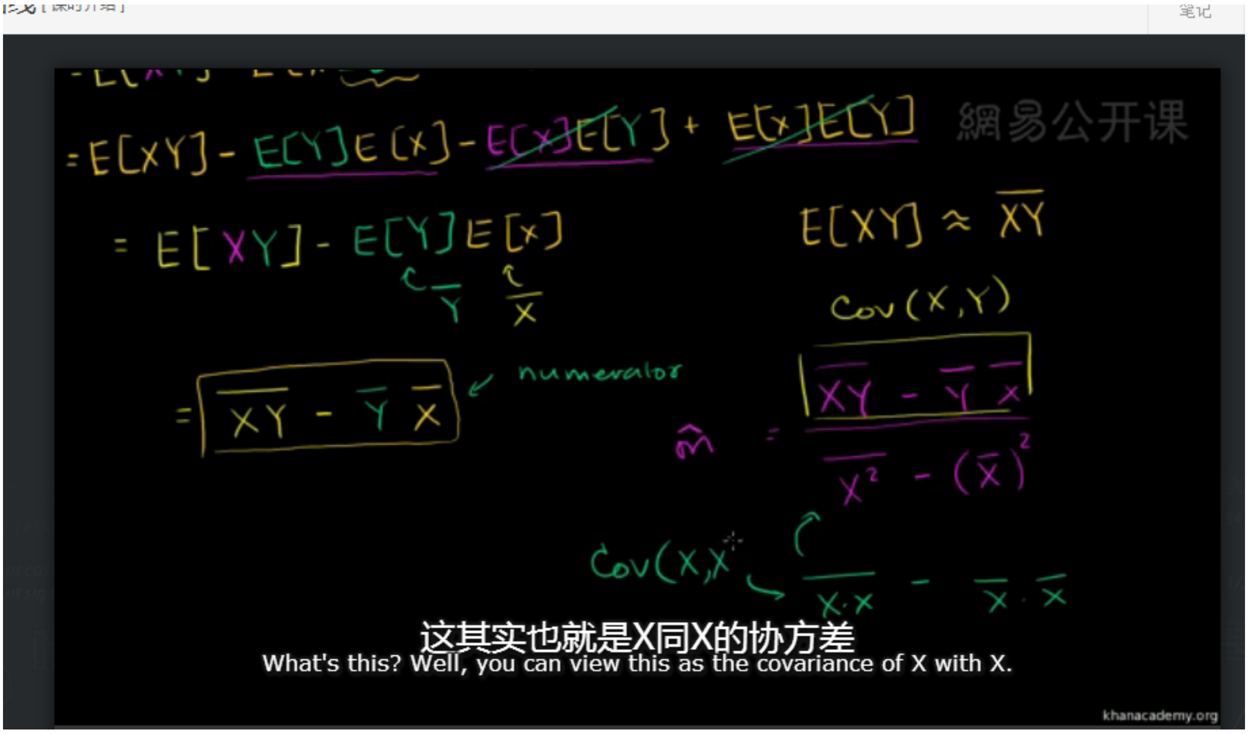


71课

协方差和回归线

协方差的定义是Cov(X,Y)=E[(X-E[X])·(Y-E[Y])]，这一节通过对该定义公式的推导，最后将协方差同线性回归良好地结合了起来。推导出，回归线的斜率m=Cov(X,Y)/Var(X)。


Cov(X,Y) = E[(X - E[X])(Y - E[Y])]
= E[XY - XE[Y] - E[X]Y + E[X]E[Y]]
X = 1 E[X] = 0
Y = 3 E[Y] = 4
網易公开课


= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y]
= E[XY] - E[Y]E[X]
E[XY] ≈ XY
Cov(X,Y)
XY - Y X
numerator
m̂ =
X² - (X)²
Cov(X,X)
X·X - X·X
網易公开课
这其实也就是X同X的协方差
What's this? Well, you can view this as the covariance of X with X.
khanacademy.org


随机变量同自身的协方差等于该随机变量的方差
is really just the variance of that random variable.

slope斜率