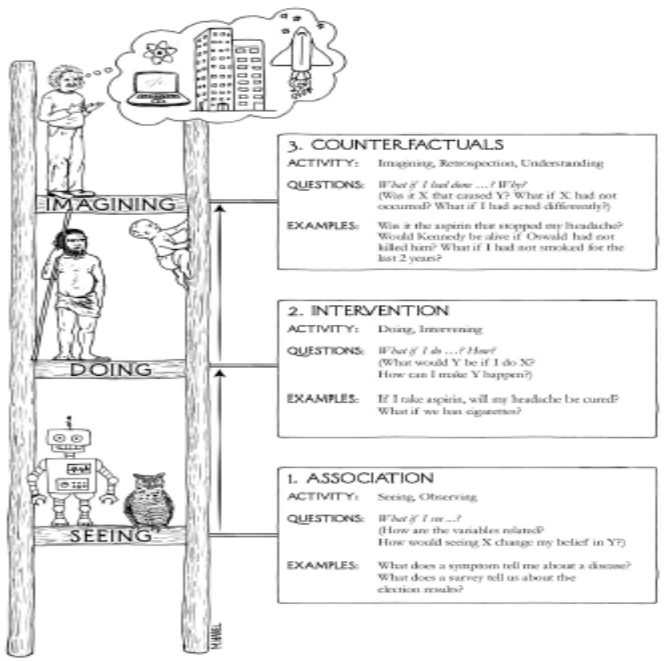
一元线性回归

国际经管学院

大数据管理与应用系

徐娜

人工智能中因果推断的三个层次



图灵奖获得者Pearl & Mackenzie (2019)
“为什么”：提出因果关系的三个层级：

1. 反事实：想象
   张三没打疫苗患了新冠；
   假若当初打疫苗，是否不患新冠？
   介
2. 干预：决策
   如果打疫苗，疫情会减轻吗？
   介
3. 相关：预测
   打疫苗越多的地方或时期，疫情越重

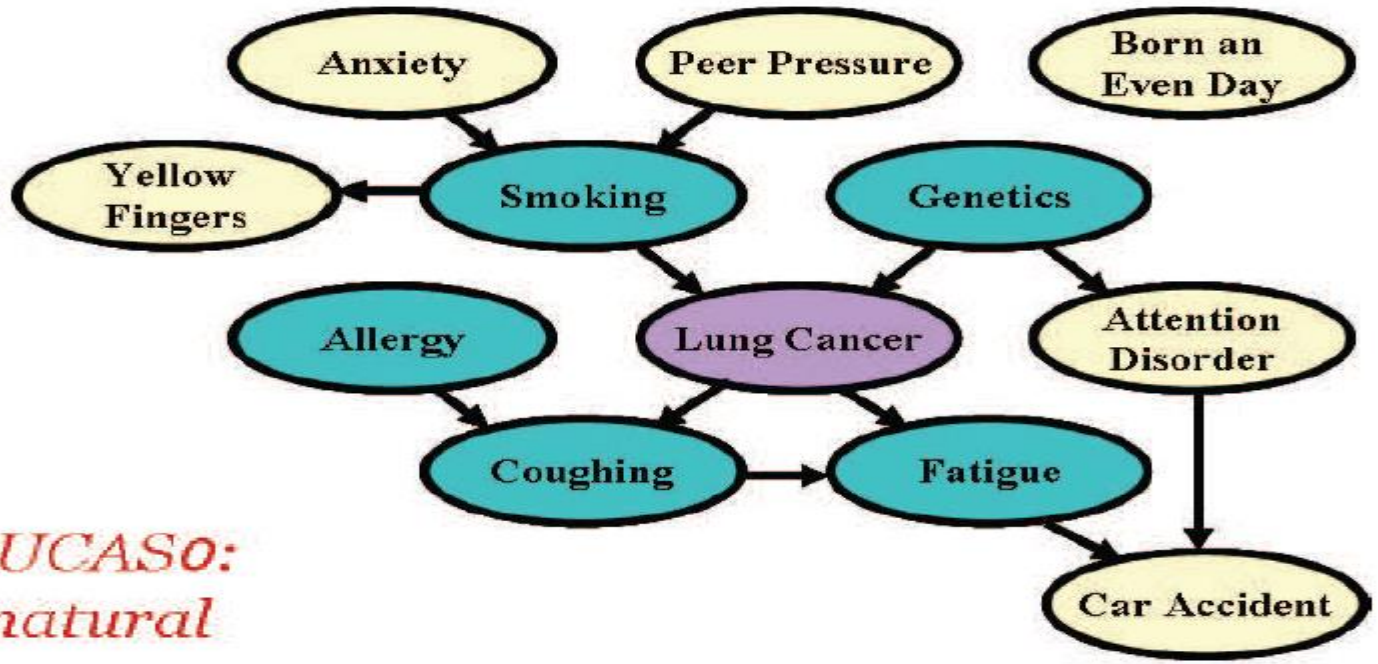
当今人工智能处于最低层次：相关
无论数据多大或神经网络多深，无法回答“干预”问题

“相关关系”不同于“因果关系”

因果与相关是两个不同的重要概念：

- 无因果关系可能会表现出虚假相关关系：
  Freedman （1991）：小学生阅读能力与鞋尺寸有强相关；
  人为地改变鞋的尺寸，不会提高他们的阅读能力。
- 有因果关系也可能表现出无相关关系：
  打太极拳能健身长寿。
  但是，打太极拳人的寿命与不打太极拳人的寿命没有差异。
  因为打太极拳的人都是体弱多病的人，
  他们打了太极拳可以与健康人一样长寿。
  铀矿工人与其它工人的寿命一样长，
  这种现象称为健康工人效应。

各有千秋：相关关系、因果关系



- 相关关系：预测、诊断、模式识别
- 因果关系：决策、病因、吸烟的因果效应、疗效

因果关系

- 确定因果关系很难；
- 如果我们控制了足够多的其他变量，那么估计的其他条件不变效应(Ceteris Paribus)通常可以被认为是因果关系；
- 近年来在识别因果关系方面，计量经济学有很多新的方法，例如PSM, DID, RDD等等。

第 1 部分

概率论与统计学知识回顾

随机变量是指**随机事件**的数量表现。

离散型随机变量：离散值，如0，1，2。随机变量即在一定区间内变量取值为有限个或可数个。例如某地区某年人口的出生数、死亡数，某药治疗某病病人的有效数、无效数等。

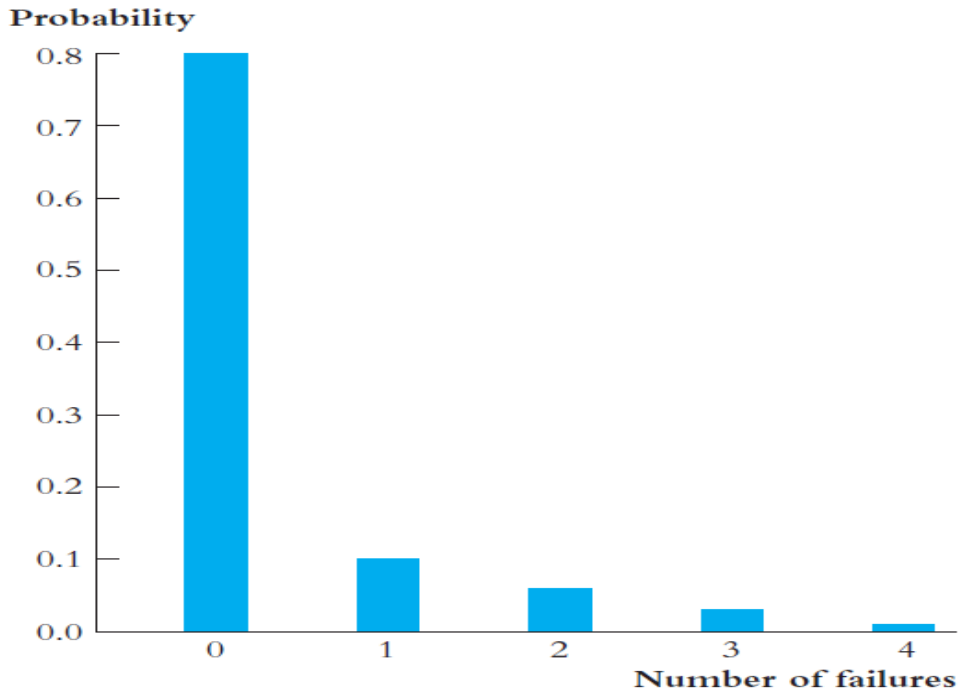
连续型随机变量：是指一系列连续的值。例如等车的时间、成人的体重值身高值、物品的使用寿命等。

概率分布

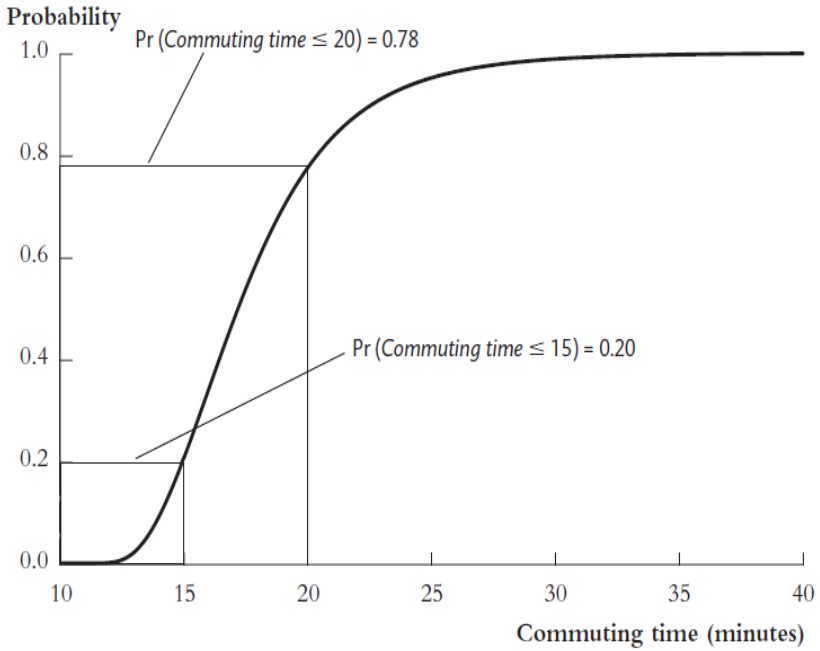
累积概率分布

TABLE 2.1 Probability of Your Wireless Network Connection Failing *M* Times

	Outcome (number of failures)				
	0	1	2	3	4
Probability distribution	0.80	0.10	0.06	0.03	0.01
Cumulative probability distribution	0.80	0.90	0.96	0.99	1.00

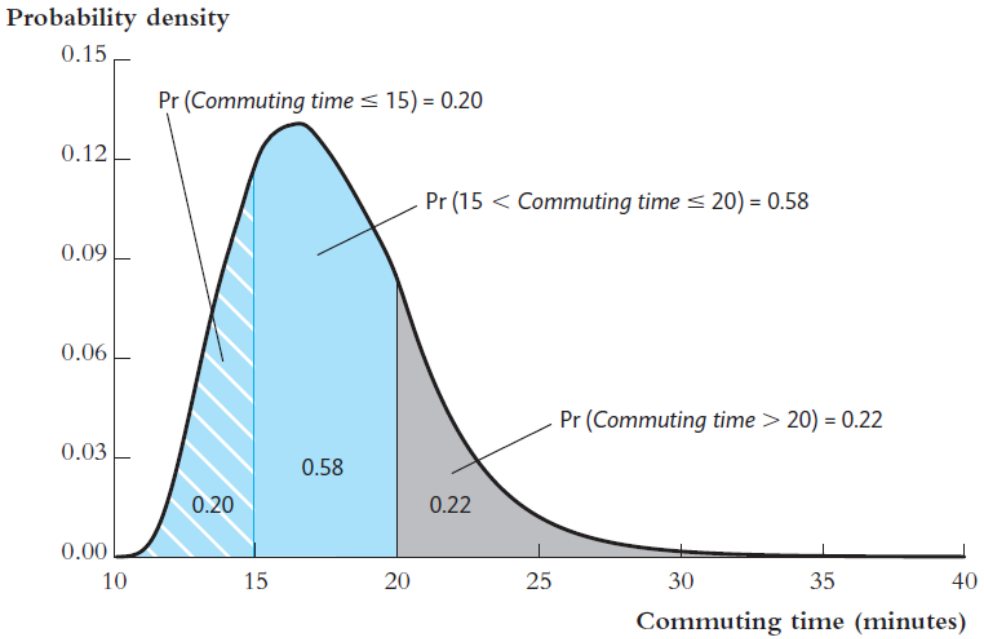


1.12 连续型随机变量



(a) Cumulative probability distribution function of commuting times

累积分布函数（CDF）



(b) Probability density function of commuting times

概率密度函数（PDF）

期望值

随机变量Y的期望值，记作E(Y)，指的是随机变量经过多次重复实验出现的长期平均值，又称为Y的**均值**或者Y的**期望**。

例：写一学期论文电脑崩溃次数的期望值

$$E(M) = 0 \times 0.80 + 1 \times 0.10 + 2 \times 0.06 + 3 \times 0.03 + 4 \times 0.01 = 0.35.$$

假设随机变量Y有k种不同的取值，$y_1, y_2, \ldots, y_k$，其中y_1表示第一个取值，y_2表示第二个取值，以此类推，同时Y取y_1对应的概率是p_1，取y_2的概率为p_2，依次类推。那么Y的期望值$E(Y)$等于

$$E(Y) = y_1p_1 + y_2p_2 + \cdots + y_kp_k = \sum_{i=1}^{k} y_ip_i$$

标准差和方差

标准差和方差度量的是概率分布的分散或者“偏差”程度。

$$Var(Y) = E\left[\left(Y - E(Y)\right)^2\right] \qquad \sigma = \sqrt{Var(Y)}$$

已知：某零件的真实长度为a。

现用甲、乙两台仪器各测量10次
两台仪器的测量结果的均值都是 a 。

无法通过均值评测哪台仪器更好，但是我们发现乙仪器的测量结果集中在均值附近，而甲的测量结果比较分散，很明显，我们会认为乙仪器的性能更好。
因此研究随机变量与其均值的偏离程度是十分必要的。那么，用怎样的量去度量这个偏离程度呢?容易看到E[|Y-E[Y]|]能度量随机变量与其均值E(Y)的偏离程度。但由于上式带有绝对值，运算不方便，通常用量$E[(Y-E(Y)^2]$ 这一数字特征就是方差。

联合分布

天气状况和行使时间的联合分布

	Rain ($X = 0$)	No Rain ($X = 1$)	Total
Long commute ($Y = 0$)	0.15	0.07	0.22
Short commute ($Y = 1$)	0.15	0.63	0.78
Total	0.30	0.70	1.00

边缘概率分布

Y的边缘概率分布可以通过X与Y的联合分布计算得出，即将Y取某一特定值存在的所有可能情况加总。

$$\Pr(Y = y) = \sum_{i=1}^{l} \Pr(X = x_i, Y = y)$$

条件分布

当随机变量***X***给定某一取值的条件下，另一随机变量***Y***的分布称为给定***X***时***Y***的条件分布。

天气状况和行使时间的联合分布

	Rain ($X = 0$)	No Rain ($X = 1$)	Total
Long commute ($Y = 0$)	0.15	0.07	0.22
Short commute ($Y = 1$)	0.15	0.63	0.78
Total	0.30	0.70	1.00

$$\Pr(Y = y | X = x) = \frac{\Pr(X = x,\ Y = y)}{\Pr(X = x)}$$

独立性

如果两个随机变量X和Y中的一个变量无法提供另一个变量的相关信息，那么这两个变量就是**独立分布**，或是**独立的**。特别地，当给定X时Y的条件分布等同于Y的边缘分布时，则X和Y是独立的。

$$\Pr(Y = y|X = x) = \Pr(Y = y) \quad （X和Y独立）$$

$$\Pr(Y = y|X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} \quad （条件分布）$$

$$\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y)$$

如果两个变量是独立的，那么这两个独立随机变量的联合分布等于它们边缘分布的乘积。

协方差和相关系数

协方差：表示两个变量的总体误差，用于度量两个变量同时变动程度的指标。

$$\begin{aligned}\operatorname{cov}(X, Y) &= \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] \\ &= \sum_{i=1}^{k}\sum_{j=1}^{l}(x_j - \mu_X)(y_i - \mu_Y)\Pr(X = x_j, Y = y_i).\end{aligned}$$

相关系数：反映两个变量之间的相关程度。

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\ \operatorname{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$ [-1,1]

相关系数为0时，表明X和Y是不相关的

两个变量独立，一定是不相关的；两个变量不相关，但不一定独立。（可以推导）

随机变量的均值和方差

X和Y之和的均值：

$$E(X+Y)=E(X)+E(Y)=\mu_X+\mu_Y$$

X和Y之和的方差：反映两个变量之间的相关程度。

$$\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)+2\mathrm{cov}(X,Y)=\sigma_X^2+\sigma_Y^2+2\sigma_{XY}.$$

$$\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)=\sigma_X^2+\sigma_Y^2$$ 如果X和Y独立

估计量和估计值

估计量是从总体中随机抽取的样本数据的函数。而估计值是基于特定样本中数据所计算出的估计量的数值。因为样本选择的随机性，因此估计量是随机变量，而估计值是一个非随机的数。

估计量的性质

无偏性

$$E(\hat{Y}) = \mu_Y$$

一致性

随着样本容量的增大，估计量的值越来越接近被估计的总体参数。

有效性

对同一总体参数的两个无偏点估计量，有更小标准差的估计量更有效

$\bar{Y}$ 是最优线性无偏估计量

随机抽样为何如此重要？

估计国家的失业率

估计女性的工资水平

居民的购物意愿调查

国民对热点事件看法的调查

因为非随机的样本会导致$\bar{Y}$有偏。

A牌的汽车百米加速度真的只要6.66秒吗？

A牌和B牌汽车百米加速度是否相同？

原假设：

H_0：$E(Y)=\mu$

备择假设：

H_1：$E(Y)\neq\mu$　（双边备择假设）

P值

P值是用来判定假设检验结果的一个参数。当原假设为真时，比所得到的样本观察结果更极端的结果出现的概率。如果P值很小，说明原假设情况的发生的概率很小，而如果出现了，根据小概率，我们就有理由拒绝原假设，P值越小，我们拒绝原假设的理由越充分。

很多情况下，统计学家们使用5%的显著性水平。

2.3 总体均值的置信区间

误差永远存在，而且不可避免

置信区间是指由样本统计量所构造的总体参数的估计区间。在统计学中，一个概率样本的置信区间（Confidence interval）是对这个样本的某个总体参数的区间估计。置信区间展现的是这个参数的真实值有**一定概率**落在测量结果的周围的程度，其给出的是被测量参数的测量值的可信程度，前面所要求的“一个概率”，就是**置信水平**。

对于一组给定的样本数据，其平均值为μ，标准差为σ，则其整体数据的平均值的100(1-α)%置信区间为$(\mu-Z_{\alpha/2}\sigma\ ,\ \mu+Z_{\alpha/2}\sigma)$

第 2 部分

一元线性回归模型

主要内容

1 简单回归模型及其基本假定

2 参数估计

3 参数最小二乘估计量的统计性质

4 拟合优度

5 显著性检验

1 简单回归模型及其基本假定

举例：教育的回报

- 人力资本投资理论表明：更高的受教育水平会带来更高的回报。
- 我们可以写出一个简单的回归式：

$$Earnings = \beta_0 + \beta_1 education + u$$

举例: 教育回报（续）

- β_1的估计值, 就是教育的回报，但两者之间的因果关系能否成立呢？
- 误差项, u, 包括影响收益的其他因素。
- 许多变量都是不可观测的，而且与教育有关，这就会带来一系列问题。

一些术语(Terminology)

- 在如下的简单回归模型中：

$$y = \beta_0 + \beta_1 x + u$$

我们将y 称为：

- 因变量（Dependent variable)
- 被解释变量（Explained Variable）

一些术语（Terminology）续

- 在y 对 x的简单线性回归中，我们把 x 称为：
  - 自变量（Independent Variable）
  - 解释变量（Explanatory Variable）
  - 在这个简单的模型里，不管x的初始值是多少，它变化一单位对y 的影响都是一样的。将来我们还要考虑递增的回报。

简单假定（Assumption）

- 误差项（扰动项）u的均值是0，
- 即：$E(u) = 0$

零条件均值假定：
Zero Conditional Mean

- 我们需要对u 和 x 的关系做出关键性的假定。
- 知道x的信息并不能使我们知道 u的信息，也就是说，它们是相互独立的，即：
  $\mathrm{E}(u|x) = \mathrm{E}(u) = 0$, 同时表明：
- $\mathrm{E}(y|x) = \beta_0 + \beta_1 x$

“线性”的含义

“线性” 可作为两种解释：对变量的线性和对参数的线性。本课“线性”回归一词总是指对参数β为线性的一种回归（即参数只以它的1次方出现）

模型对参数为线性？	模型对变量为线性？	
	是	不是
是	LRM	LRM
不是	NLRM	NLRM
LRM=线性回归模型； NLRM =非线性回归模型		

“线性”的含义

- $Y=\beta_1+\beta_2X+u$ 是线性的！
- $\ln Y=\beta_1+\beta_2\ln X+u$ 是线性的！
- $Y=\beta_1\ln(\beta_2X+u)$ 不是线性的！
- $Y=1/(\beta_1+\beta_2X)+u$ 不是线性的

非线性关系的线性化

$$wage = \exp(\beta_0 + \beta_1 edu + u)$$

对上述两边取对数，可得：

$$\log(wage) = (\beta_0 + \beta_1 edu + u)$$

假如估计的结果为：

$$\log(wage) = 0.584 + 0.083 edu$$

教育每增加一年，工资提高**8.3%**。

模型	因变量	自变量	对β_1的解释
水平值—水平值	y	x	$\Delta y=\beta_1 \Delta x$
水平值—对数	y	log(x)	$\Delta y=(\beta_1/100)\% \Delta x$
对数—水平值	log(y)	x	$\% \Delta y=(100\beta_1) \Delta x$
对数—对数	log(y)	log(x)	$\% \Delta y= \beta_1 \% \Delta x$

LS // Dependent Variable is CONS

Sample(adjusted): 1979 2000

Included observations: 22 after adjusting endpoints

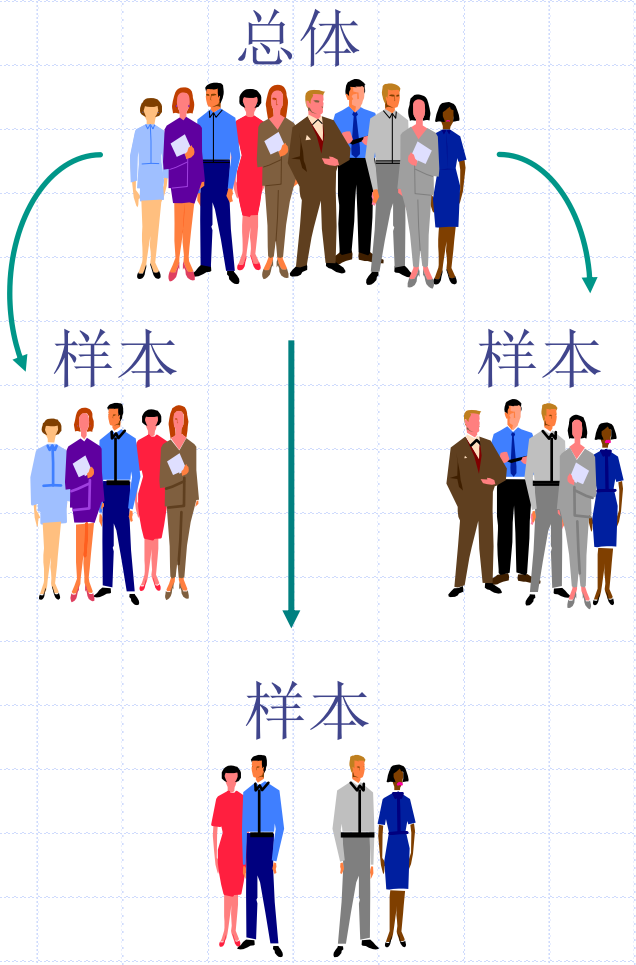
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	120.7000	36.51036	3.305912	0.0037
GDPP	0.221327	0.060969	3.630145	0.0018
CONSP(-1)	0.451507	0.170308	2.651125	0.0158

R-squared	0.995403	Mean dependent var	928.4946
Adjusted R-squared	0.994920	S.D. dependent var	372.6424
S.E. of regression	26.56078	Akaike info criterion	6.684995
Sum squared resid	13404.02	Schwarz criterion	6.833774
Log likelihood	-101.7516	F-statistic	2057.271
Durbin-Watson stat	1.278500	Prob(F-statistic)	0.000000

Q：分析参数估计值的经济学含义

总体与样本

- 总体是我们研究的目的，但是不能知道总体的全部数据
- 用总体中的一部分（样本）来推断总体的性质



样本回归函数（SRF）

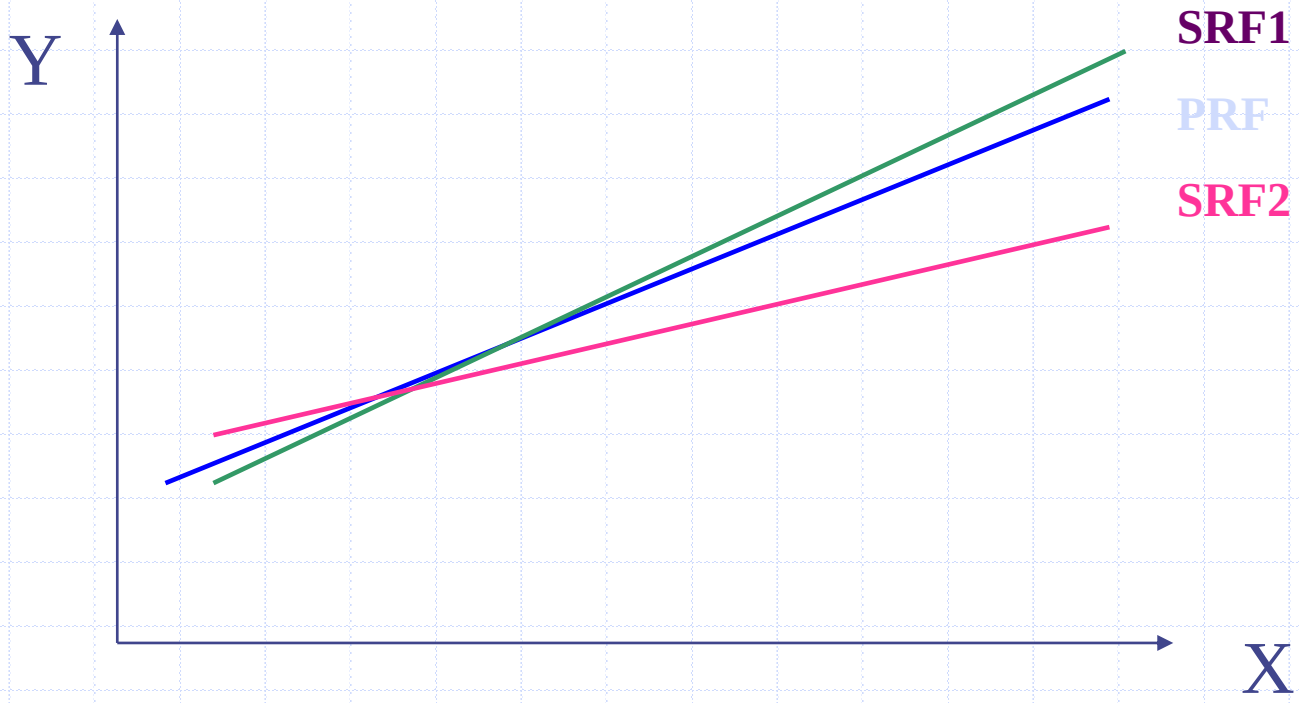
- 两个随机样本，对应给定的每个X_i只有一个Y值，问：能从样本数据中估计出总体回归曲线（ Popular Regression Curve ）吗？

X	Y
80	70
100	65
120	90
⋮	⋮
220	150

X	Y
80	55
100	88
120	90
⋮	⋮
220	175

样本回归线与总体回归线

- 比较两条样本回归线SRF1和SRF2（假定PRF是直线），问哪条样本线代表“真实”的总体回归线？



样本回归函数

$\mathrm{SRF}: \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$　　(相对于$\mathrm{E(Y \mid X_i)} = \beta_1 + \beta_2 X_i$)

其中$\hat{Y}_i$是$\mathrm{E(Y \mid X_i)}$的估计量；

$\hat{\beta}_1$是β_1的估计量；

$\hat{\beta}_2$是β_2的估计量。

估计量（Estimator）：一个估计量又称统计量，是指一个规则、公式或方法，是用已知的样本所提供的信息去估计总体参数。在应用中，由估计量算出的数值称为估计值。

比较PRF和SRF

$$\text{PRF}: \ \mathrm{E}(\mathrm{Y} \mid \mathrm{X_i}) = \beta_1 + \beta_2 X_i$$

$$Y_i = \mathrm{E}(\mathrm{Y} \mid \mathrm{X_i}) + u_i = \beta_1 + \beta_2 X_i + u_i$$

$$\text{SRF}: \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

$$Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

其中$\hat{u}_i$是残差项(*residual*)

回归分析的主要目的是根据SRF $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$

来估计PRF: $Y_i = \beta_1 + \beta_2 X_i + u_i$

古典假定

经典线性回归模型（CLRM）的基本假定：

$$Y_i=\beta_1+\beta_2X_i+u_i\ (i=1,2,3\ldots,n)$$

假定1：所有参数为常数，保证模型为线性关系

假定2：随机抽样，有一个来自总体的样本容量为n的随机样本

假定3：解释变量有样本波动，即在样本中取值不都一样

假定4：干扰项的均值为零。即，$E(u_i|X_i)=0$

假定5：同方差性或u_i的方差相等。即$Var(u_i|X_i)=\sigma^2$

假定6：随机干扰项服从0均值、同方差的正态分布。

即：$u_i \sim N(0,\ \sigma^2)$

注：前5个被称为高斯-马尔科夫假定。在实际建模时，除了假定6以外，对模型是否满足假定都要进行检验。对于假定6，由中心极限定理，当样本趋于无穷大时，对于任何实际模型都是满足的。

2. OLS参数估计（β和σ^2）

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)$$

OLS：ordinary least square

求解：

$$\min \ f(\hat{\beta}_1, \hat{\beta}_2) = \min \sum \hat{u}_i^2$$

$$= \min \sum [Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)]^2$$

正规方程（Normal equation）

$$\text{由}\begin{cases}\dfrac{\partial \mathrm{f}}{\partial \hat{\beta}_1}=2\sum[Y_i-(\hat{\beta}_1+\hat{\beta}_2X_i)]\times(-1)=0\\ \dfrac{\partial \mathrm{f}}{\partial \hat{\beta}_2}=2\sum[Y_i-(\hat{\beta}_1+\hat{\beta}_2X_i)]\times(-X_i)=0\end{cases}$$

$\Rightarrow$ 得到的方程组称为正规方程组

整理得：

$$\sum Y_i=n\hat{\beta}_1+\hat{\beta}_2\sum X_i \qquad (1)$$

$$\sum Y_iX_i=\hat{\beta}_1\sum X_i+\hat{\beta}_2\sum X_i^2 \qquad (2)$$

β 的估计

1、公式：

解上述正规方程组得到$\hat{\beta}_1$和$\hat{\beta}_2$的估计值

$$\hat{\beta}_2=\frac{n\sum Y_iX_i-\sum Y_i\sum X_i}{n\sum X_i^2-(\sum X_i)^2}=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}=\frac{\sum y_ix_i}{\sum x_i^2}$$

其中$\overline{X}$和$\overline{Y}$是X和Y的样本均值

定义离差：$x_i=X_i-\bar{X}$；$y_i=Y_i-\bar{Y}$。

用小写字母表示对均值的离差。将$\hat{\beta}_2$代入正规方程（1）得

$$\hat{\beta}_1=\frac{1}{n}\sum Y_i-\hat{\beta}_2\frac{1}{n}\sum X_i=\bar{Y}-\hat{\beta}_2\bar{X}$$

注意“帽子”的含义

$\hat{\beta}_1$是对β_1的估计；

$\hat{Y}_i$是对$E(Y_i \mid X_i)$的估计，
而不是对Y_i的估计；

$\hat{u}_i$是样本残差，而u_i是总体随机扰动项。
所以，$\hat{u}_i$不表示对u_i的估计。

3 参数最小二乘估计量的统计性质

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

一、线性性：$\hat{\beta}$是因变量Y_i的线性函数，也是 扰动项u_i的线性函数

证明：

$$1.\ \hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i (Y_i - \bar{Y})}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum x_i^2} - \frac{\sum x_i \bar{Y}}{\sum x_i^2}$$

$$= \sum \left(\frac{x_i}{\sum x_i^2} \right) Y_i - \frac{\bar{Y} \sum x_i}{\sum x_i^2} = \sum \left(\frac{x_i}{\sum x_i^2} \right) Y_i$$

$$(\because \sum x_i = \sum (X_i - \bar{X}) = \sum x_i - n\bar{X} = n\bar{X} - n\bar{X} = 0)$$

$\therefore \hat{\beta}_2$是Y_i的一个线性函数，是线 性估计量(*Linear estimator*)

3.1线性性（续）

证明：

2. $\because \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$

$\hat{\beta}_2$是Y_i的一个线性函数

$\therefore \hat{\beta}_1$也是Y_i的一个线性函数

注：$\hat{\beta}$也是随机变量

因为$Y_i = \beta_i + \beta_2 X_i + u_i$，$\hat{\beta}$也是 扰动项$u_i$的线性函数

3.2 无偏性

二、无偏性：即 $E(\hat{\beta})=\beta$

证明：1. $\because \hat{\beta}_2=\sum\left(\frac{x_i}{\sum x_i^2}\right)Y_i=\sum\left(\frac{x_i}{\sum x_i^2}\right)(\beta_1+\beta_2X_i+u_i)$

$$=\beta_1\sum\left(\frac{x_i}{\sum x_i^2}\right)+\beta_2\sum\left(\frac{x_iX_i}{\sum x_i^2}\right)+\sum\left(\frac{x_i}{\sum x_i^2}\right)u_i$$

$$=\beta_2\sum\left(\frac{x_i(x_i+\bar{X})}{\sum x_i^2}\right)+\sum\left(\frac{x_i}{\sum x_i^2}\right)u_i=\beta_2+\sum\left(\frac{x_i}{\sum x_i^2}\right)u_i$$

$$\therefore E(\hat{\beta}_2)=\beta_2+\sum\left(\frac{x_i}{\sum x_i^2}\right)E(u_i)=\beta_2$$

3.2 无偏性（续）

证明：2. $\because \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = \frac{1}{n}\sum Y_i - \sum\left(\frac{x_i}{\sum x_i^2}\right) Y_i \bar{X}$

$$= \sum\left(\frac{1}{n} - \frac{x_i \bar{X}}{\sum x_i^2}\right) Y_i = \sum\left(\frac{1}{n} - \frac{x_i \bar{X}}{\sum x_i^2}\right)(\beta_1 + \beta_2 X_i + u_i)$$

$$= \beta_1 \sum\left(\frac{1}{n} - \frac{x_i \bar{X}}{\sum x_i^2}\right) + \beta_2 \sum\left(\frac{X_i}{n} - \frac{x_i X_i \bar{X}}{\sum x_i^2}\right) + \sum\left(\frac{1}{n} - \frac{x_i \bar{X}}{\sum x_i^2}\right) u_i$$

$$\because \sum\left(\frac{1}{n} - \frac{x_i \bar{X}}{\sum x_i^2}\right) = 1 - \frac{\bar{X}}{\sum x_i^2} \sum x_i = 1$$

$$\sum\left(\frac{X_i}{n} - \frac{x_i X_i \bar{X}}{\sum x_i^2}\right) = \frac{1}{n}\sum X_i - \bar{X}\frac{\sum x_i X_i}{\sum x_i^2} = \bar{X} - \bar{X} = 0$$

$$\therefore E(\hat{\beta}_1) = \beta_1 + \sum\left(\frac{1}{n} - \frac{x_i \bar{X}}{\sum x_i^2}\right) E(u_i) = \beta_1$$

3.3 最小方差性

高斯-马尔可夫定理：

在给定古典线性回归模型的假定之下，OLS估计量在无偏线性估计量一类中，有最小方差。称之为最优线性无偏估计量（BLUE）。

BLUE-Best Linear Unbiasedness Estimator

有最小方差的无偏估计量叫做有效估计量（efficient estimator）

最小二乘估计量的方差

$$\mathrm{var}(\hat{\beta}_2) = \sigma^2 \Big/ \sum x_i^2$$

证明：

$$1.\ \hat{\beta}_2 = \sum \left(\frac{x_i}{\sum x_i^2} \right) Y_i = \sum \left(\frac{x_i}{\sum x_i^2} \right) (\beta_1 + \beta_2 X_i + u_i)$$

$$= \beta_2 + \sum \left(\frac{x_i}{\sum x_i^2} \right) u_i$$

$$\therefore \mathrm{var}(\hat{\beta}_2) = \sum \left(\frac{x_i}{\sum x_i^2} \right)^2 \mathrm{var}(u_i) = \sigma^2 \Big/ \sum x_i^2$$

最小二乘估计量的方差（续）

$$\mathrm{var}(\hat{\beta}_1)=\sigma^2\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum x_i^2}\right)=\sigma^2\frac{\sum X_i^2}{n\sum x_i^2}$$

证明：

$$1.\ \hat{\beta}_1=\beta_1+\sum\left(\frac{1}{n}-\frac{x_i\bar{X}}{\sum x_i^2}\right)u_i$$

$$\therefore \mathrm{var}(\hat{\beta}_1)=\sum\left(\frac{1}{n}-\frac{x_i\bar{X}}{\sum x_i^2}\right)^2\mathrm{var}(u_i)$$

$$=\sigma^2\sum\left(\frac{1}{n^2}-\frac{2x_i\bar{X}}{\sum x_i^2}+\frac{(x_i\bar{X})^2}{(\sum x_i^2)^2}\right)$$

$$=\sigma^2\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum x_i^2}\right)\ \underline{\underline{\text{通分后}}}\ \sigma^2\frac{\sum X_i^2}{n\sum x_i^2}$$

最小二乘估计量的方差（续）

如果古典假定6成立

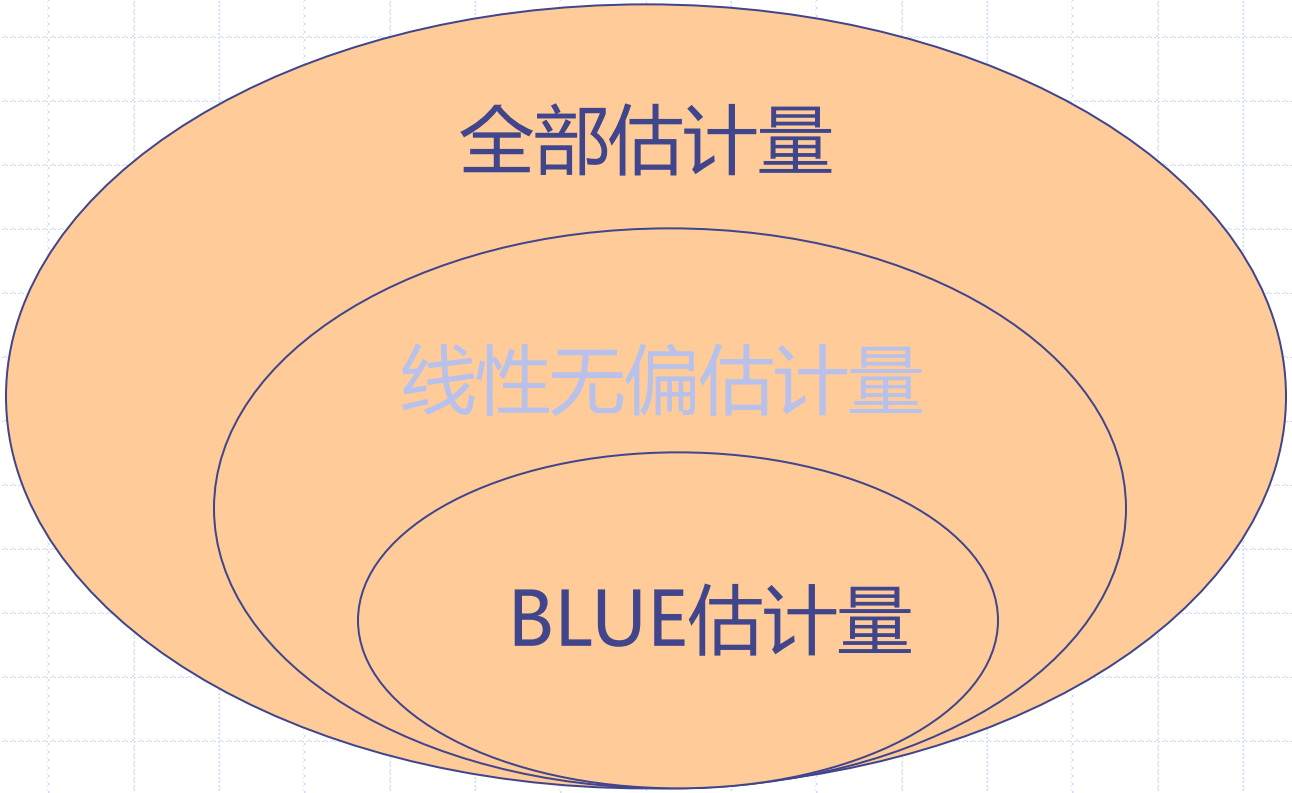
即，u_i服从$N(0,\sigma^2)$

则$\hat{\beta}_2$服从$N(\beta_2,\frac{\sigma^2}{\sum x_i^2})$

$\hat{\beta}_1$服从$N(\beta_1,\sigma^2\frac{\sum X_i^2}{n\sum x_i^2})$

（最小方差性的证明略）

BLUE估计量的图形表示



(最小方差性的证明略)

4 拟合优度检验（统计检验之一）

问：样本回归线对数据拟合得有多好？

如果全部观测点都落在样本回归线上，则得到的是一个“完美”的拟合。

一般情形：总有一些正的残差或负的残差。我们希望这些围绕着回归线的残差尽可能小。

判定系数R^2就是用来做拟合优度检验的。

平方和公式中各项的解释

- 总平方和（SST）

  是实测的Y值围绕其均值的总变异。

- 解释平方和（SSE）

  是估计的Y值围绕其均值的变异。

- 残差平方和（SSR）

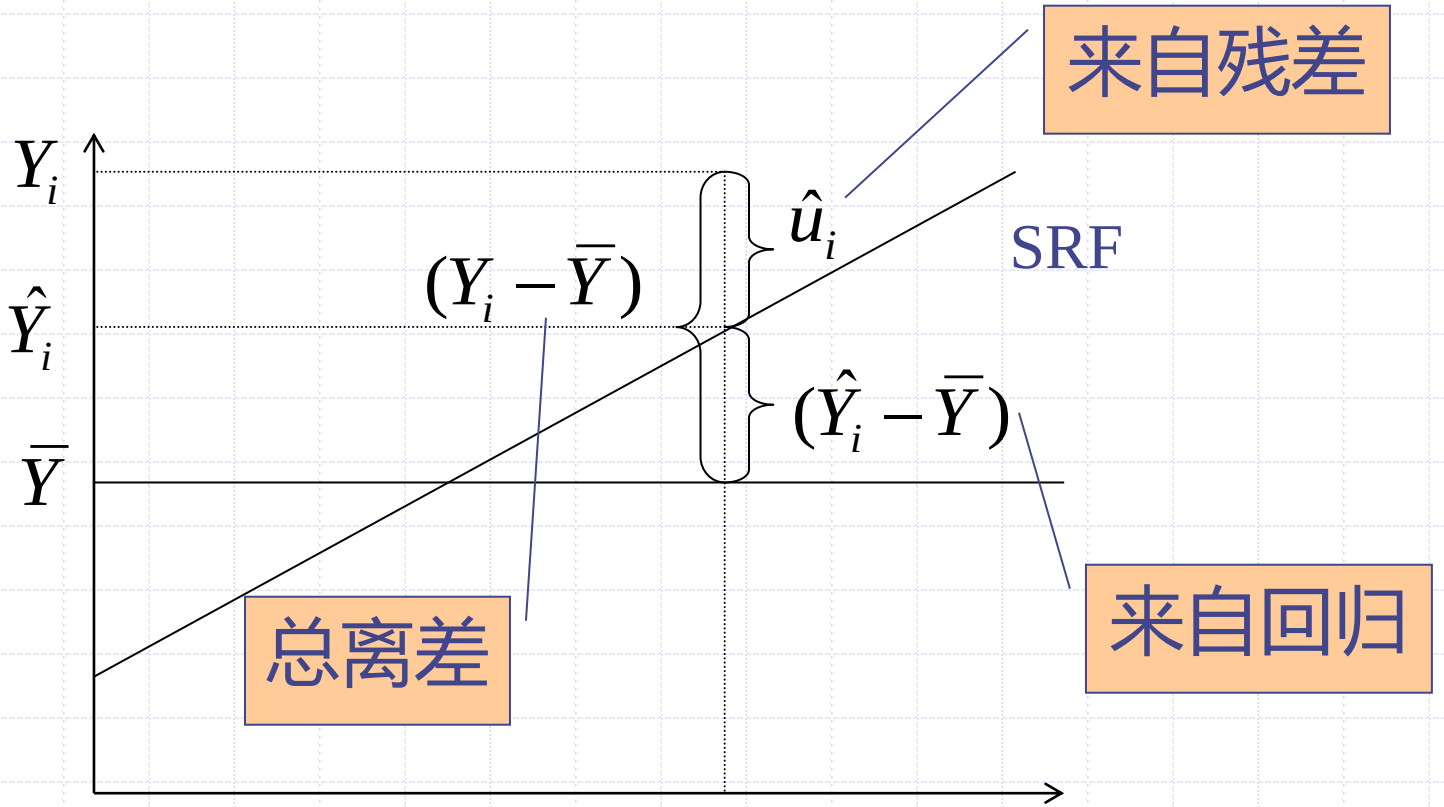
  是未被解释的围绕回归线的Y的变异。

$$\sum y_i^2 = \sum (Y_i - \bar{Y})^2$$

$$\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{\hat{Y}})^2$$

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

平方和公式的几何表示



平方和公式

由$Y_i=\hat{Y}_i+\hat{u}_i,\ \bar{Y}=\bar{\hat{Y}}$前后两个等式相减得：

$y_i=\hat{y}_i+\hat{u}_i$　两边求平方得

$y_i^2=\hat{y}_i^2+\hat{u}_i^2+2\hat{y}_i\hat{u}_i$　对i求总和得

$$\sum y_i^2=\sum \hat{y}_i^2+\sum \hat{u}_i^2+2\sum \hat{y}_i\hat{u}_i$$

$$\because \sum \hat{y}_i\hat{u}_i=\sum \hat{\beta}_2 x_i\hat{u}_i=\sum \hat{\beta}_2 x_i(y_i-\hat{y}_i)$$

$$=\hat{\beta}_2\left(\sum x_i y_i-\hat{\beta}_2\sum x_i^2\right)=0$$

$$\therefore y_i^2=\hat{y}_i^2+\hat{u}_i^2$$

R^2公式

$$R^2 = \frac{SSE}{SST} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

$$\text{或} R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}$$

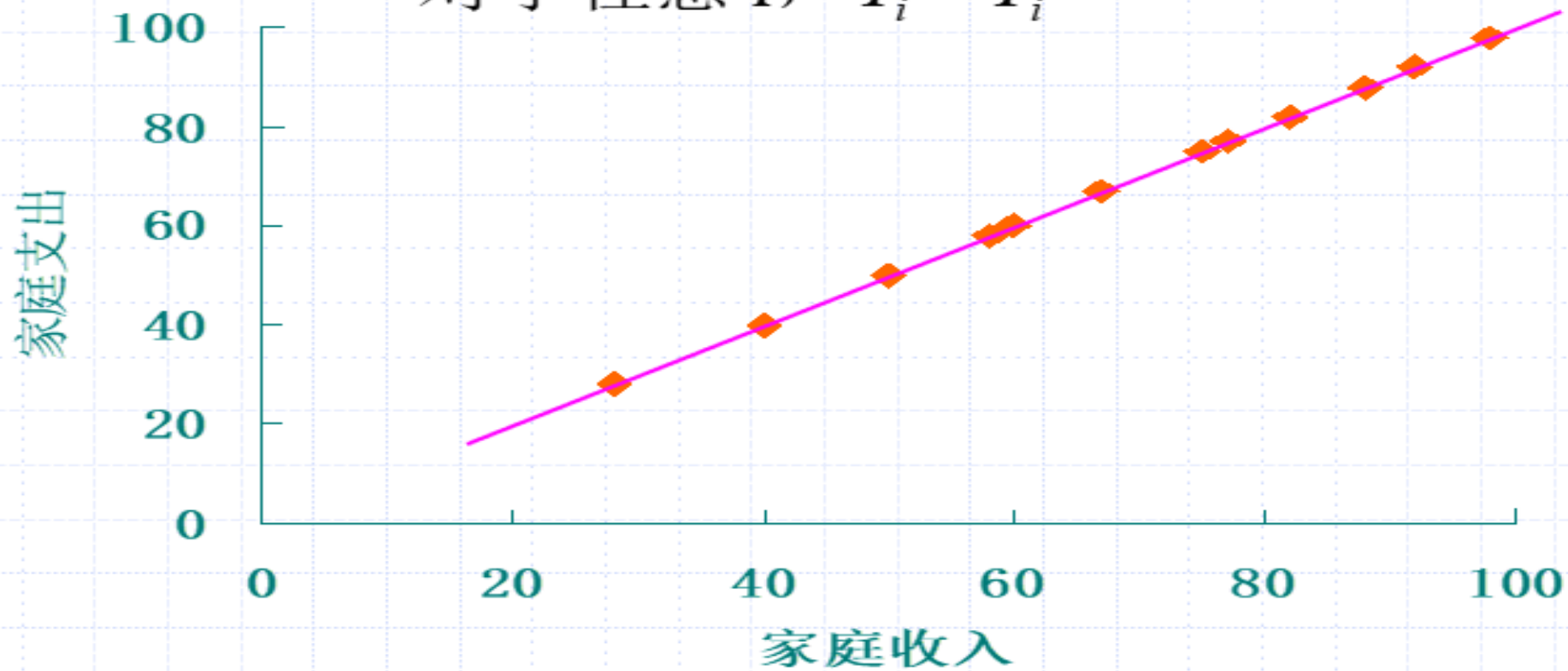
性质：$0 \leqslant R^2 \leqslant 1$

问：$R^2 = 0$ 意味着什么？

$R^2 = 1$ 意味着什么？

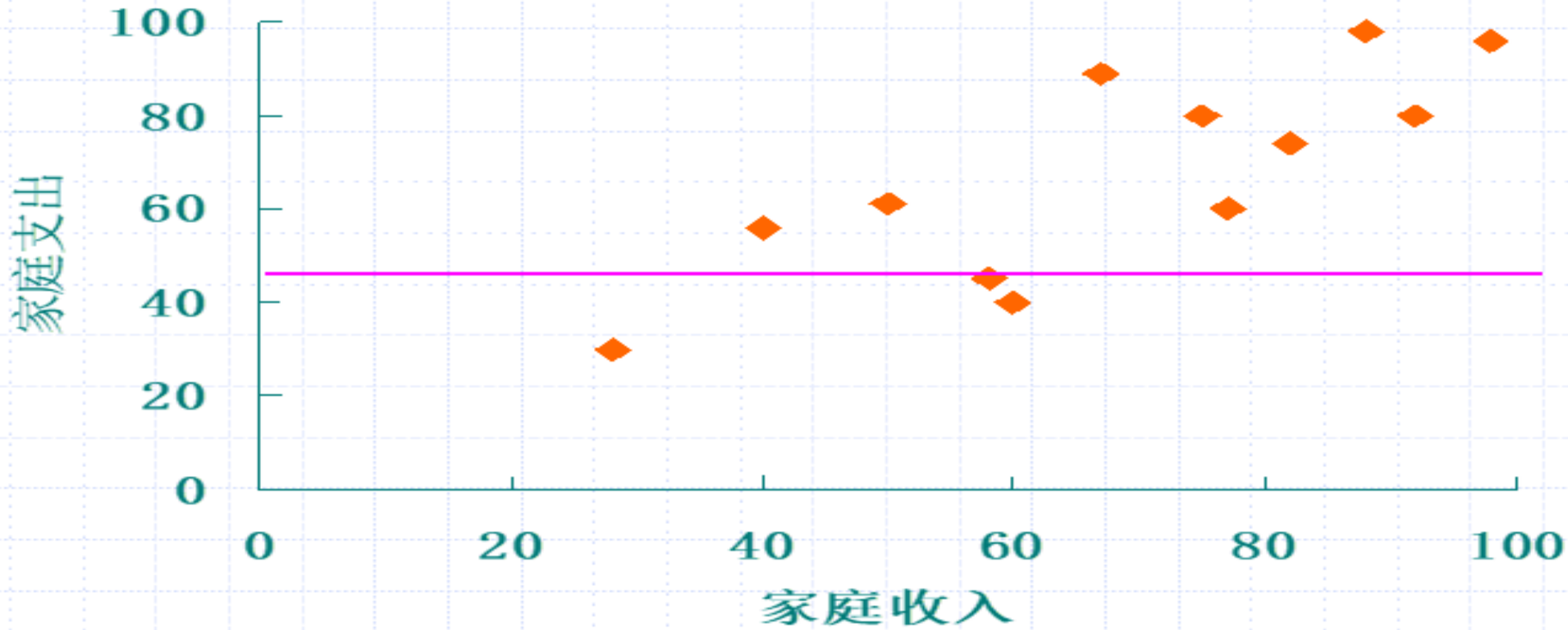
$R^2 = 1$

对于任意 i，$\hat{Y}_i = Y_i$



$R^2 = 0$

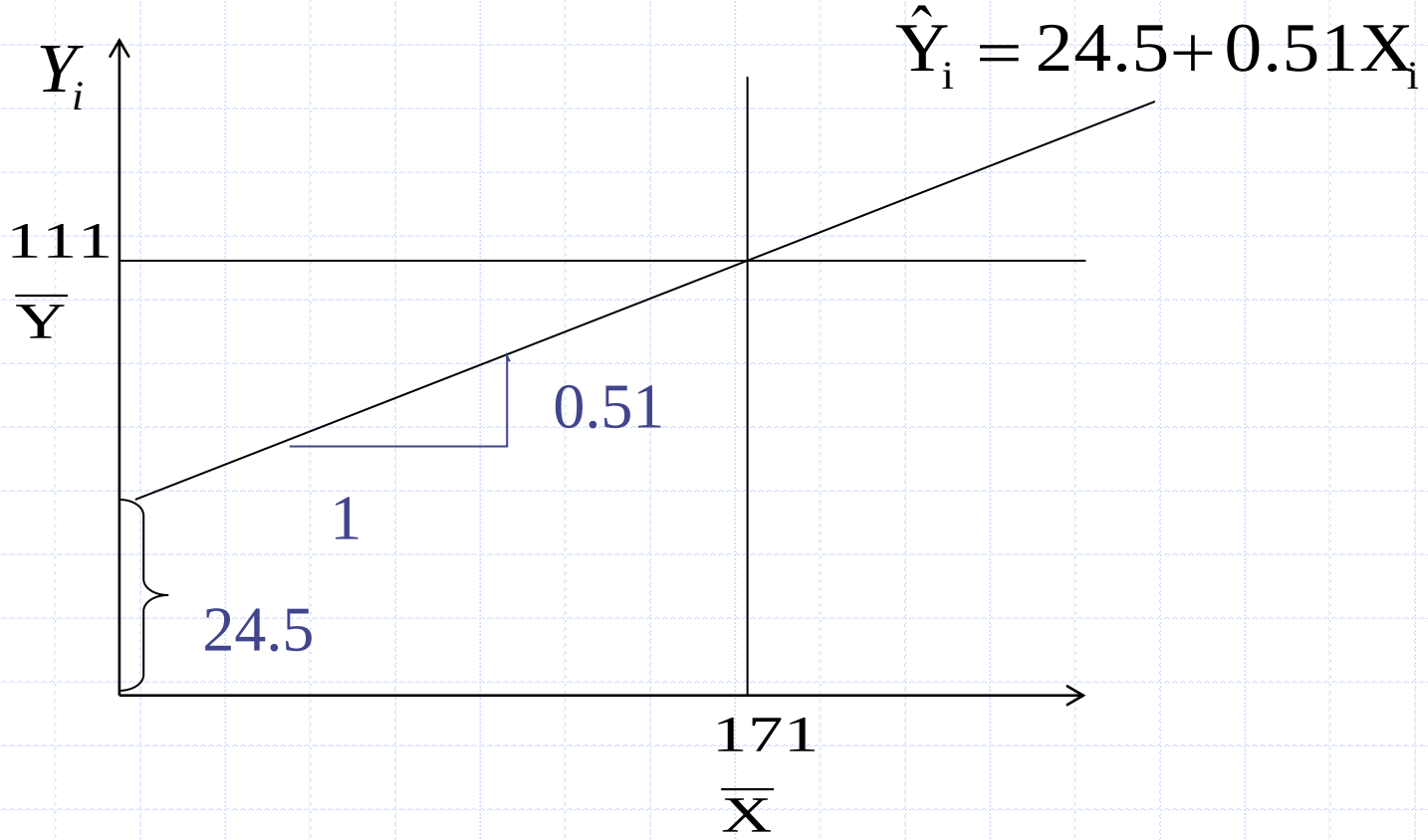
对于任意 i，$\hat{Y}_i = \hat{\beta}_1 = \bar{Y}$



R^2 =0.86

- R^2 =0.86
- 表示约有86%的因变量Y的变异能由解释变量X来说明。
- 或者说，解释变量解释了因变量Y变异中的86%。
- 注意：不表示有86%的样本观测点落在了样本回归线上！！（此处容易错。）

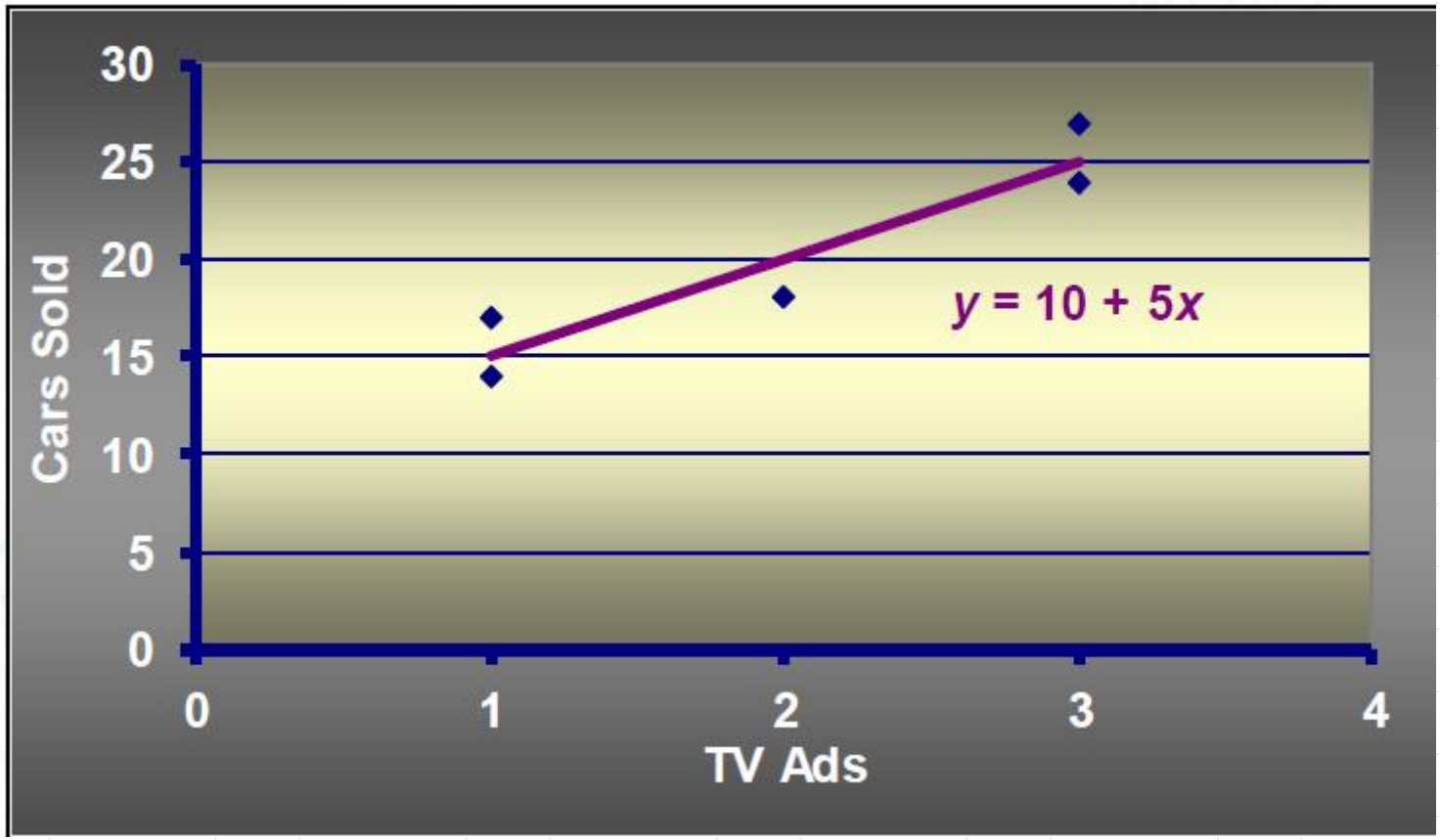
一个数值例子：支出与收入



5 显著性检验

- 在估计了回归方程以后，为了更好地评价自变量X与因变量Y之间的因果关系，我们还需要对系数进行统计推断。

例 1:广告投入和汽车销售

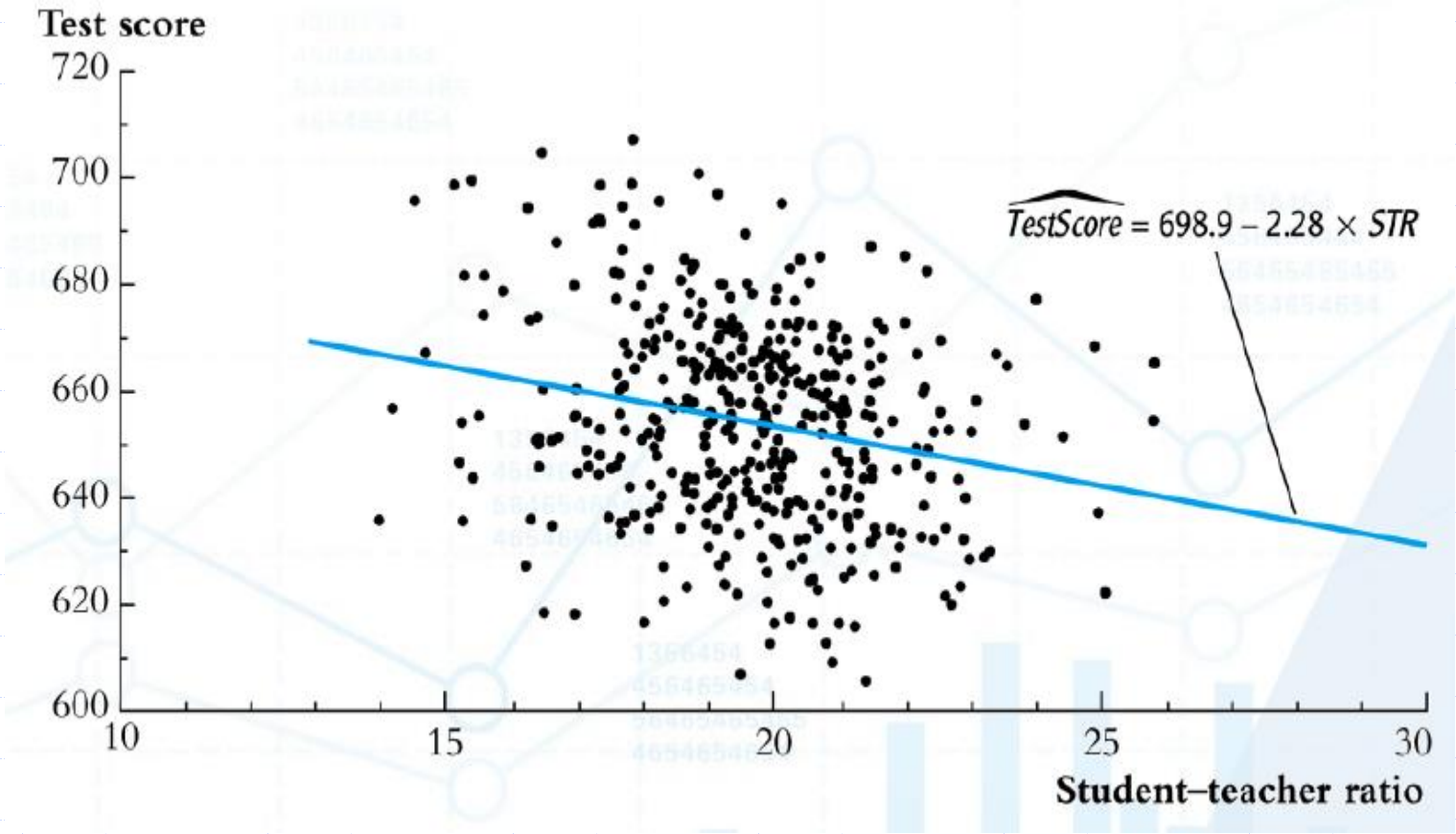


在估计出$\hat{\beta}_1 = 5$之后，我们是否能够拒绝"广告数量对汽车销量没有显著影响"这一假设？

$$H_0: \beta_1 = 0 \text{ vs } H_1: \beta_1 \neq 0$$

如果$\beta_1 \neq 0$，则说广告数量的影响是显著(significant)的。

例2：班级规模如何影响学生考试成绩？



t 检验:单双侧备择假设和显著水平

01 除了原假设H_0外， 我们还需要备择假设H_1 和置信水平

02 H_1可以是单侧的也可以是双侧的

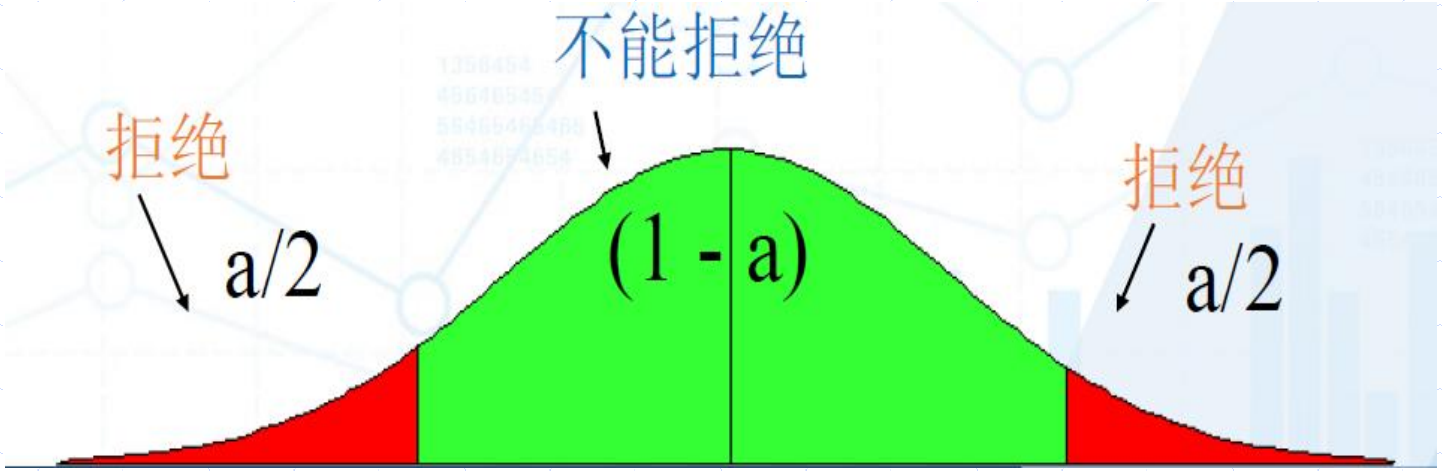
---$H_1: \mu > 0$ 和$H_1: \mu < 0$是单侧备择假设

---$H_1: \mu \neq 0$是双侧备择假设

03 如果H_0为真，而我们的结论是拒绝了H_0, 则称为犯了第一类（拒真）错误；显著性水平α就是犯第一类错误的概率，通常可选1%，5%或10%；在没有特别说明的情况下α指定为5%。

双侧备择假设

$$H_0: \mu = \mu_0 \text{ v.s. } H_1: \mu \neq \mu_0$$



使用t统计量检验方法

使用t统计量进行假设检验的步骤分为以下四步：

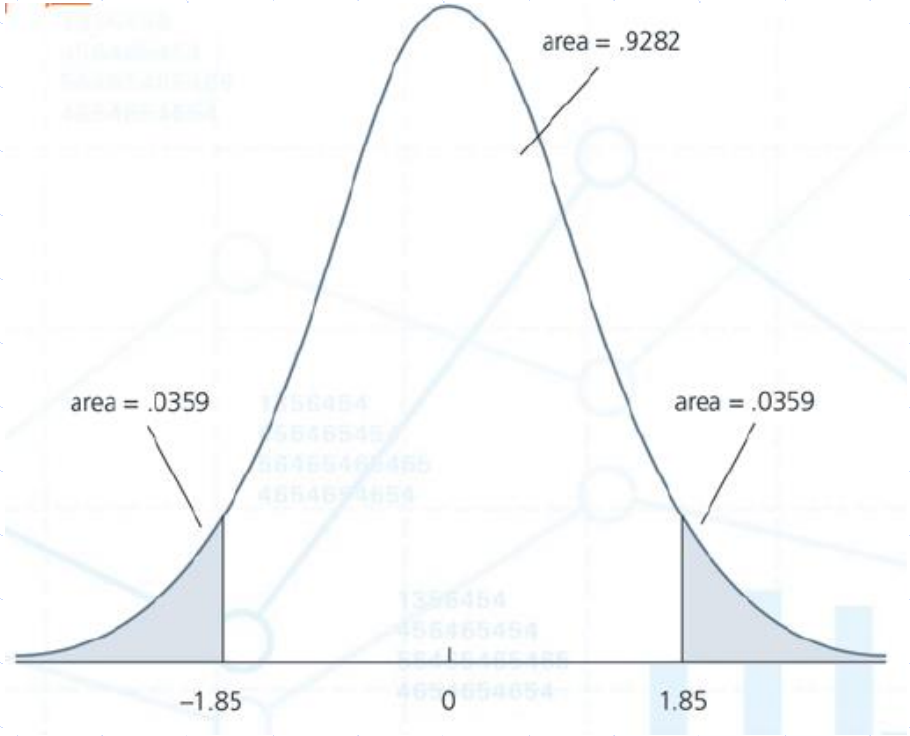
- 1. 列出原假设和备择假设
- 2. 确定置信水平（或显著性水平）
- 3. 计算t统计量的值 $t = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$
- 4. 若t值落在拒绝域中就拒绝原假设。

（使用正态分布近似t分布的临界值如下表）

H_1	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.01$
$\mu \neq \mu_0$	$\|t\| > 1.64$	$\|t\| > 1.96$	$\|t\| > 2.58$
$\mu > \mu_0$	$t > 1.28$	$t > 1.64$	$t > 2.33$
$\mu < \mu_0$	$t < -1.28$	$t < -1.64$	$t < -2.33$

使用P值的检验方法

P值定义为当原假设为真时，根据数据计算的t统计量出现的概率。下图是一个自由度为40，t 统计量为1.85所对应的双侧检验的p值。



例：班级规模如何影响学生考试成绩？

估计结果为：

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

$$\hat{\beta}_1 = -2.28\ ,\ SE(\hat{\beta}_1) = 0.522$$

因此

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{-2.28 - 0}{0.52} = -4.38$$

因为$|t| = 4.38 \gg 2.58$，所以即使在0.01的显著水平上，我们也拒绝H_0: $\beta_1 = 0$。

置信区间

OLS估计值$\hat{\beta}_2$是一个点估计值，

问：它离真实的β_2有多近？有多可靠？

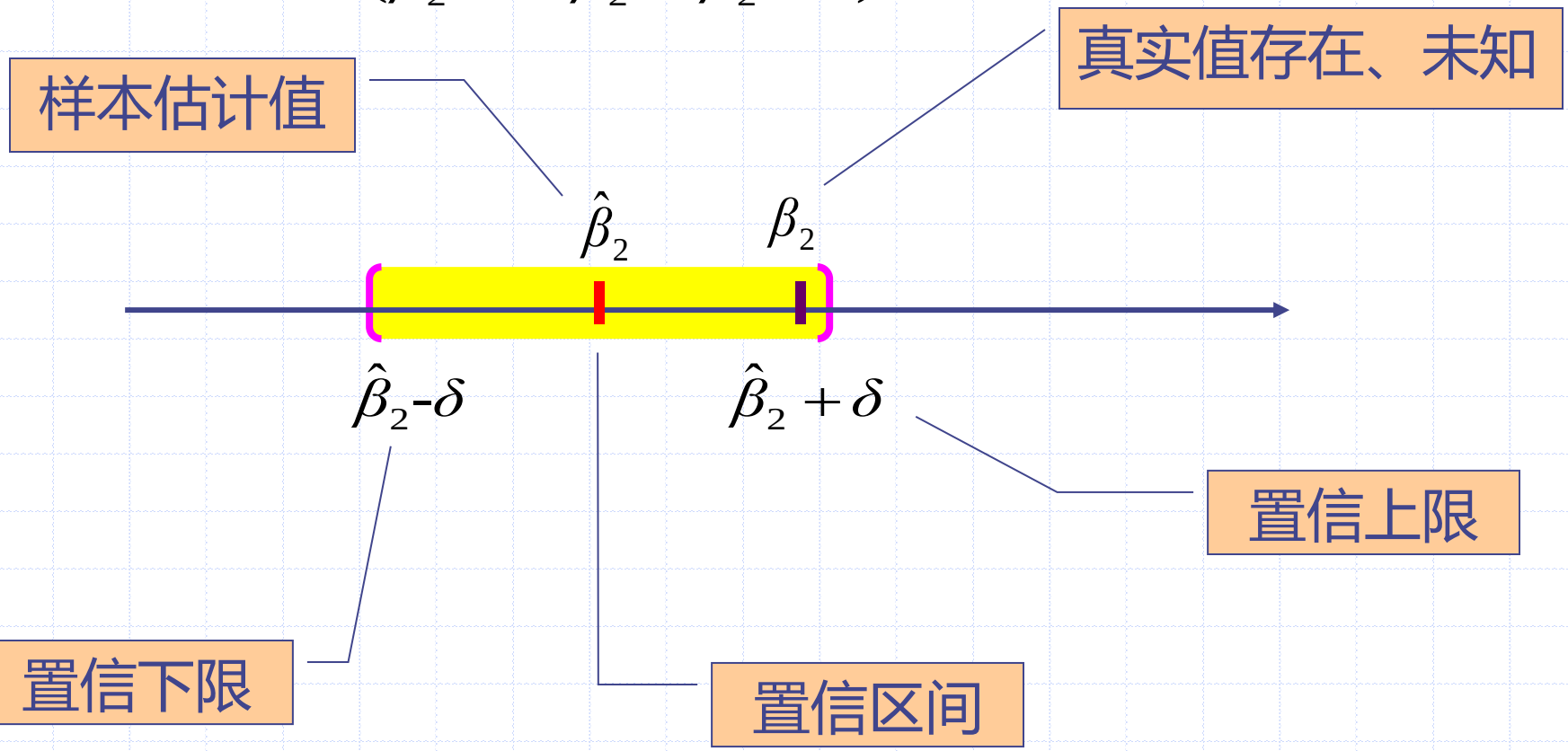
由于抽样的波动，单个估计值很可能不同于真值。

在统计学中，一个点估计量的可靠性由它的标准误来衡量。因此，我们不能完全信赖一个点估计值，而要围绕点估计量构造一个区间。

如：在点估计量的两边各宽2个或3个标准误的一个区间，使得它有95%的概率包含着真实的参数值。

置信区间的图形表示

$$\Pr(\hat{\beta}_2-\delta \le \beta_2 \le \hat{\beta}_2+\delta) = 1-\alpha$$



$$\Pr(\hat{\beta}_2-\delta \le \beta_2 \le \hat{\beta}_2+\delta)=1-\alpha$$

置信区间(Confidence interval): 这样的一个区间如果存在的话，就称为置信区间。

置信系数(Confidence coefficient): 1-α称为置信系数。

显著性水平(Level of significance)： α（0<α<1）。

置信限(Confidence limit):置信区间的端点。

置信下限(Lower Confidence limit): $\hat{\beta}_2-\delta$

置信上限(upper Confidence limit): $\hat{\beta}_2+\delta$

回归系数β_1和β_2的置信区间(续)

$$t=\frac{\hat{\beta}_2-\beta_2}{se(\hat{\beta}_2)}=\frac{\hat{\beta}_2-\beta_2}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2}}}\sim t(n-2)$$

注：$t=\dfrac{\text{估计量-参数}}{\text{估计量的标准误的估计值}}$

Estimated standard error

即：$se(\hat{\beta}_2)=\sqrt{\dfrac{\hat{\sigma}^2}{\sum x_i^2}}$ 是估计量 $\hat{\beta}_2$ 的标准差 $\sqrt{\dfrac{\sigma^2}{\sum x_i^2}}$

回归系数β_1和β_2的置信区间(续)

由$\Pr(-t_{\frac{\alpha}{2}} \le t \le t_{\frac{\alpha}{2}}) = 1-\alpha$得

$$\Pr\left(-t_{\frac{\alpha}{2}} \le \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2}}} \le t_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Rightarrow \Pr[\hat{\beta}_2 - t_{\frac{\alpha}{2}} se(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{\frac{\alpha}{2}} se(\hat{\beta}_2)] = 1-\alpha$$

$\therefore \beta_2$的显著水平为α的置信区间为

$$[\hat{\beta}_2 - t_{\frac{\alpha}{2}} se(\hat{\beta}_2),\ \hat{\beta}_2 + t_{\frac{\alpha}{2}} se(\hat{\beta}_2)]$$

同理，β_1的显著水平为α的置信区间为

$$[\hat{\beta}_1 - t_{\frac{\alpha}{2}} se(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\frac{\alpha}{2}} se(\hat{\beta}_1)]$$

回归系数β_1和β_2的置信区间(续)

- 置信区间的宽度与估计量的标准误成正比，因此，估计量的标准误常被喻为估计量的精度(precision)