中山大學
SUN YAT-SEN UNIVERSITY

实验设计与数据分析

第十讲：线性回归

肖叶，副教授
环境科学与工程学院
中山大学

广州，2023

回归分析

基本概念

- 回归分析的起源
- 回归分析的研究内容和研究目的
- 变量之间的统计关系
- 回归模型与回归参数
- 总体回归函数和样本回归函数的关系

回归分析的起源

- 回归分析（regression analysis）

  A return to a former or less developed state

- 起源
  - 1886年英国生物学家、统计学家F. Galton（高尔顿， 1822-1911）指出：在同一种族里，父亲高的，儿子的平均高度比父亲矮，但高于种族平均高度；父亲矮的，儿子的平均高度比父亲高，但矮于种族平均高度。也就是说，儿子的高度有”回归“于种族平均高度的趋势。
- 发展
  - 现代统计应用学的重要分支
  - 回归的现代含义要比其原始含义广得多

回归分析的含义

回归分析的研究内容

- 一个变量或者一组变量（即自变量）之变化对另一个变量（即因变量）之变化的影响程度

回归分析的研究目的

- 根据已知的自变量的变异来估计或者预测因变量的变异情况
- 例如：探索在已知父亲们身高的条件下，儿子们的平均身高是怎样变动的

变量之间的关系

- 可以分为两大类
- 函数关系（functional relation，deterministic relation）
- 统计关系（statistic relation）

变量之间的函数关系

- 定义
  - 变量之间数量上的确定性关系
- 数学表达
  - 设X为自变量，Y为因变量，X和Y的函数关系可以表示为：$Y = f(X)$
- 特点
  - 通过函数式，一个或一组变量在数量上的变化就可以**完全确定**另一个变量在数量上的变化。
- 举例
  - 圆面积与半径之间的关系：$A = \pi r^2$

变量之间的统计关系

定义

- 在客观世界里，许多变量之间也存在相随变动，并具有某种规律性；
- 但是，数量关系往往**并非完全确定**

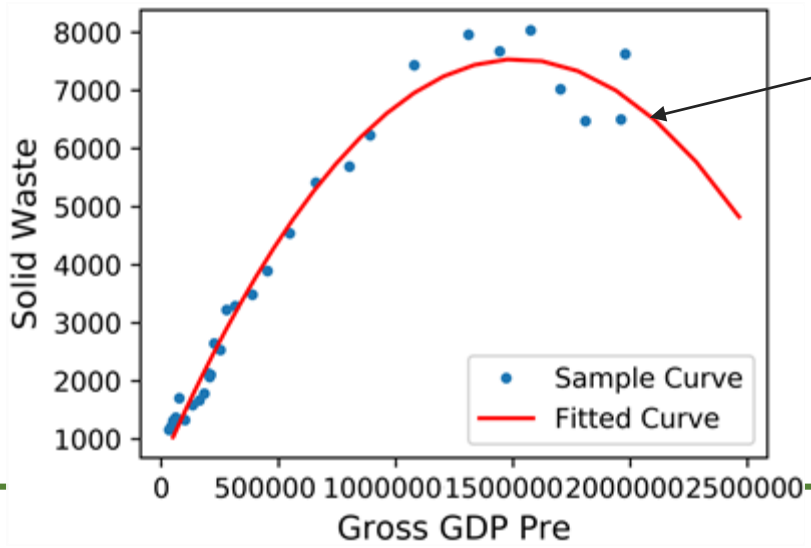
举例

- 固体废物产生量与经济发展水平（GDP）之间的关系

饮食文化习惯?

地区气候?

产业结构?



库兹涅茨曲线

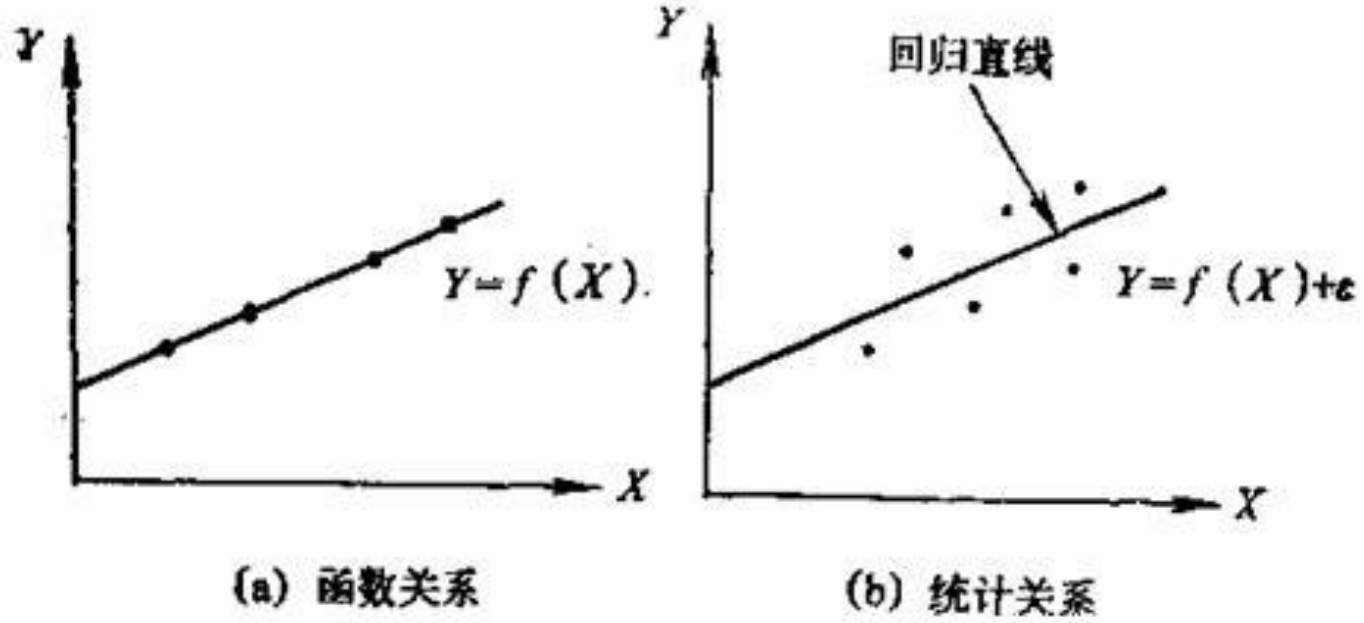
变量之间的统计关系

数学表达

- 如果把固体废物产生作为因变量Y，
- 把GDP、食物热量、经纬度和三产比例作为自变量，并分别用X_1、X_2、X_3和X_4来表示，
- 那么它们之间的关系可以表达为：
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon = f(X) + \varepsilon$
- ε为随机扰动误差项

变量之间的两种关系

- X与Y变量之间的两种关系



回归模型

- 回归分析所研究的变量之间的关系：
  - 是一种非确定性的统计关系
- 回归模型：
  - 回归分析当中用到的对变量之间统计关系进行定量描述的数学模型
- 例如：
  - $Y = f(X) + \varepsilon$
  - $f(X)$称为回归函数
  - ε称为随机扰动误差项，或简称误差项

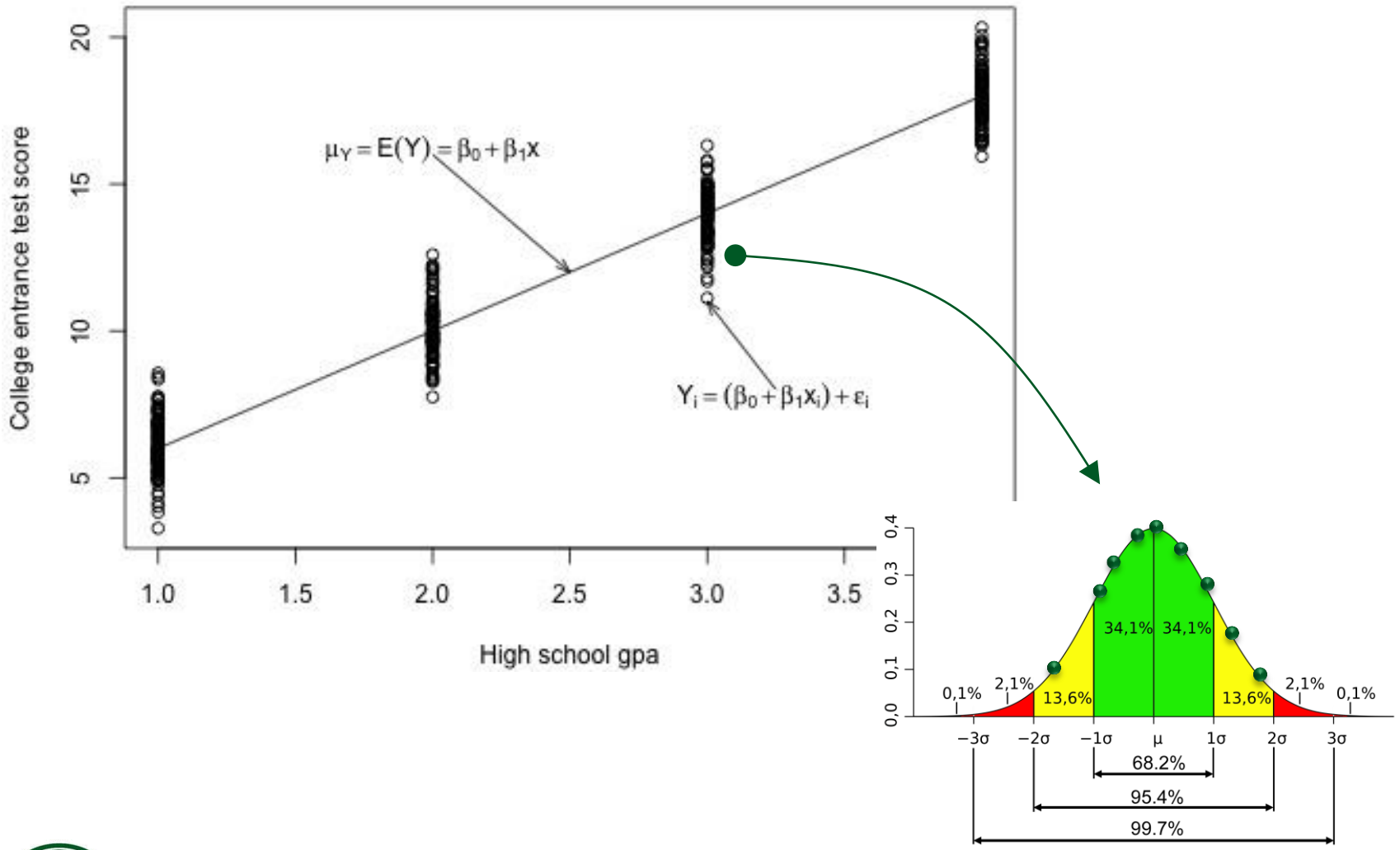
回归模型

- 随机扰动误差项ε，一般有四个来源：
  - 被省略掉而没有纳入方程，但实际上又影响着因变量Y的种种因素；
    - 这些因素被省略有多种原因
    - 在回归分析时尚不知道，难于观测，难于量化，为了简化模型，等等
  - 变量的观测误差；
  - 模型的设定误差（结构误差，用非线性近似线性）；
  - 随机误差
- 回归函数：
  - 总体回归函数
  - 样本回归函数

总体回归函数

- 每当X取某一给定值时，都会有一个总体与之对应。该总体：
  - 由所有Y的相应的可能值组成；
  - 具有一定的概率分布（条件概率分布）
- 总体回归函数（PRF）：
  - 从Y的各个条件概率分布中求出的各个条件均值（条件数学期望值）所组成的一条Y对X的回归线（直线或者曲线）
- PRF的物理意义：
  - PRF代表的是Y之条件均值，反映了在X取定值的条件下，因变量Y的平均变化状态。

总体回归函数


College entrance test score
20
15
10
5
$\mu_Y = E(Y) = \beta_0 + \beta_1 x$
$Y_i = (\beta_0 + \beta_1 x_i) + \varepsilon_i$
1.0
1.5
2.0
2.5
3.0
3.5
High school gpa
0,4
0,3
0,2
0,1
0,0
34,1%
34,1%
13,6%
13,6%
2,1%
2,1%
0,1%
0,1%
−3σ
−2σ
−1σ
μ
1σ
2σ
3σ
68.2%
95.4%
99.7%

总体回归函数

- 理论上，总体回归函数的形式
  - 根据定义，总体回归函数 PRF 可以记作：
  - $f(X) = E(Y|X)$
- 实践中，
  - 受各种因素的限制，在回归分析中，人们无法知道总体回归函数的确切形式
  - 为了研究X与Y之间的数量变化规律，可以对总体回归函数的形式做必要的、合理的假设

总体回归函数

- 总体回归函数的形式：
  - 假设X与Y之间是线性关系，那么，
    - $f(X) = \beta_0 + \beta_1 X_1$，该式又被称为一元线性回归方程
    - $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ 被称为一元线性回归模型
- “元”：
  - 是指回归模型中变量的个数，包括因变量Y在内
  - 如果回归模型中含有**一个自变量**，则称为一元回归模型
  - 如果回归模型中含有**多个自变量**，则称为多元回归模型

样本回归函数

- 实践中，在X取某定值的条件下，不可能掌握Y的所有可能值（总体），往往，
  - 只能掌握Y的一组样本值
  - 可以直接得到的只能是样本回归函数：

样本回归函数

- 回归分析的过程在本质上就是：
  - 根据样本回归函数（Sample Regression Function, SRF）来估计未知的总体回归函数（Population Regression Function, PRF）
  - 由于通常只知道总体中的一个样本，所以只能获得近似的估计结果。

PRF 与 SRF

举例

- 假设要研究某类企业的产量和污染物排放量之间的关系，以产量为自变量X，排污量为因变量Y。
- 一方面：
  - 随着产量的增加，污染物排放量必然也会相应增加
- 另一方面：
  - 除了产量之外，影响排污量的因素还有很多，如企业的不同技术水平、管理水平等等
- 因此，在研究该类企业的排污量与产量之间的数量变化规律时：
  - 在给定产量（X）的条件下所得到的排污量（Y）值将是围绕某些中心值上下波动的统计值
  - 与给定X值相对应的Y的所有可能

PRF 与 SRF

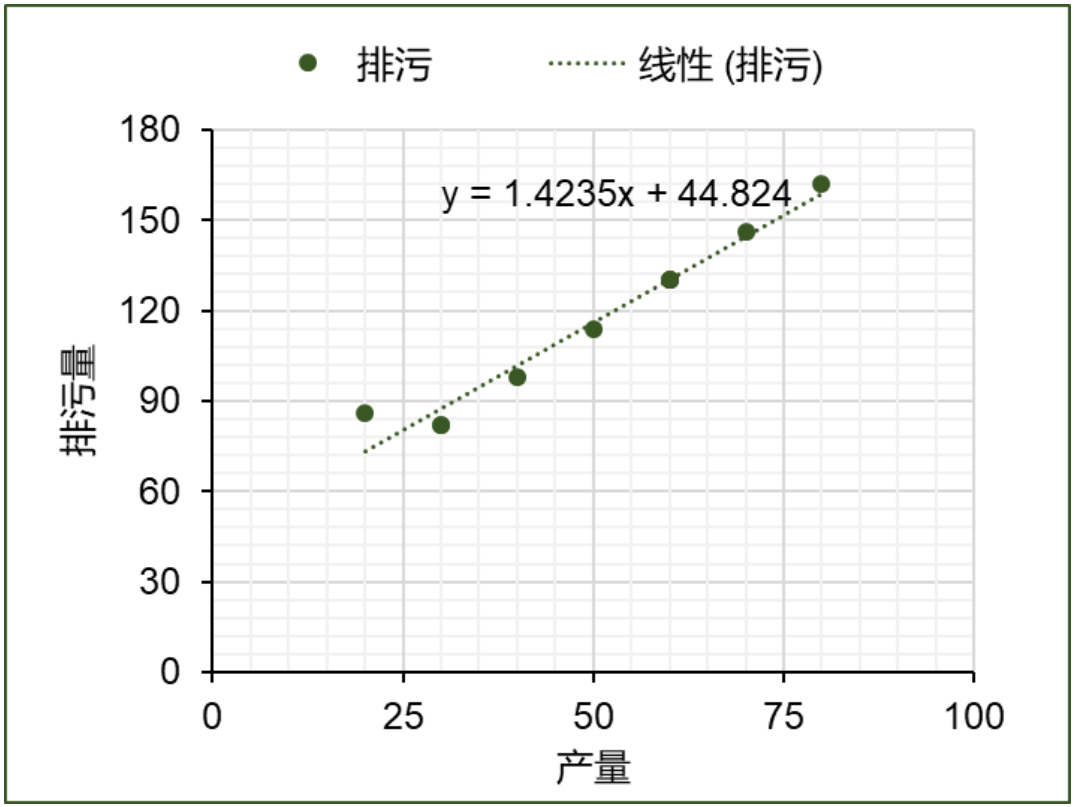
- 总体已知的情况：
  - 全部61个企业的产量、排污量及其10个分组是已知的，并作为统计数据列于表中。根据这些数据就可以获得PRF。

	X_1 30	X_2 20	X_3 60	X_4 80	X_5 40	X_6 50	X_7 60	X_8 30	X_9 70	X_{10} 60
Y_1	73	79	139	159	102	129	130	84	142	133
Y_2	91	46	128	173	80	97	135	88	140	136
Y_3	70	71	127	155	87	96	137	82	146	132
Y_4	75	50	138	169	98	108	122	96	137	120
Y_5	85	84	118	146	90	140	126	90	148	143
Y_6	98		115	170	110			69	154	116
Y_7			145		119			65	155	
合计	492	330	910	972	686	570	650	574	1022	780
Y的条件概率	1/6	1/5	1/7	1/6	1/7	1/5	1/5	1/7	1/7	1/6
Y的条件均值	82	86	130	162	98	114	130	82	146	130

PRF 与 SRF

- 总体已知的情况：
  - 全部61个企业的产量、排污量及其10个分组是已知的，并作为统计数据列于表中。根据这些数据就可以获得PRF。

	X_1 30	X_2 20	X_3 60	X_4 80	X_5 40	X_6 50	X_7 60	X_8 30	X_9 70	X_{10} 60
Y的条件均值	82	86	130	162	98	114	130	82	146	130



PRF 与 SRF

- 总体未知的情况：
  - 从Y的总体里随机抽取两个样本，样本1和样本2。
  - 用两组样本数据可以得到两条样本回归线：SRF_1和SRF_2。
  - 无论是SRF_1还是SRF_2，都是对总体回归线PRF的近似。

样本1

编号	X_i	Y_i
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

样本2

编号	X_i	Y_i
1	30	75
2	20	71
3	60	118
4	80	173
5	40	90
6	50	96
7	60	126
8	30	96
9	70	140
10	60	136

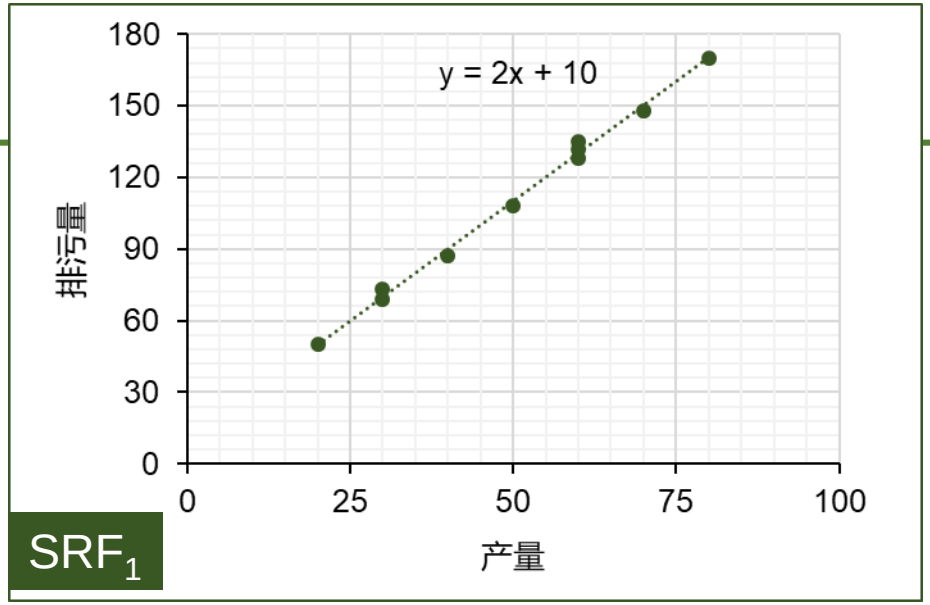
PRF 与 SRF

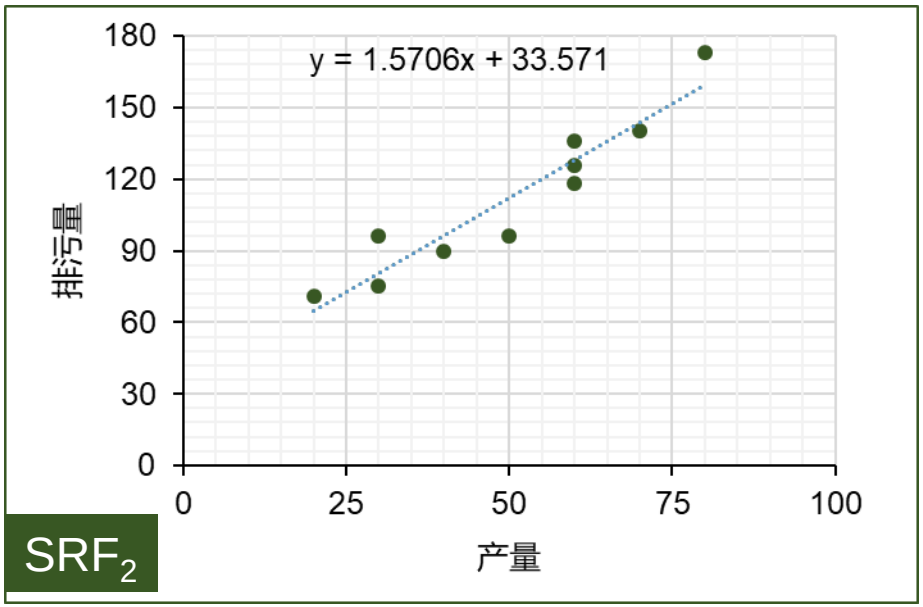
样本1

X_i	Y_i
30	73
20	50
60	128
80	170
40	87
50	108
60	135
30	69
70	148
60	132

样本2

X_i	Y_i
30	75
20	71
60	118
80	173
40	90
50	96
60	126
30	96
70	140
60	136





PRF 与 SRF

- 回归分析的重要任务：
  - 根据样本回归函数SRF尽可能精确地估计总体回归函数PRF
  - 要达到这一目的，就要使未知回归参数的估计量尽可能地接近其真值
- 如何做到这一点?

回归分析

- 基本概念
- 一元回归分析
  - 回归参数的估计
  - 回归模型的检验
  - 回归预测
- 多元回归分析
- 回归函数的形式

回归参数的估计方法

- 回归参数的估计方法
  - 最小平方法
  - 高斯提出的
  - 又称为最小二乘法

- 以一元线性回归模型为例讨论最小平方法

最小平方法

- 一元线性回归模型的一般形式：
  - $Y = E(Y|X) + \varepsilon$
  - $E(Y|X) = \beta_0 + \beta_1 X$
  - Y为因变量，X为自变量，β_0、β_1为回归参数，ε为随机扰动误差项

- 对Y和X分别进行n次独立观测，就可以取得n对观测值：
  - $(X_i, Y_i), i = 1, 2, \ldots, n$，则有，
  - $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, 2, \ldots, n$
  - $E(Y_i|X_i) = \beta_0 + \beta_1 X_i \qquad i = 1, 2, \ldots, n$

最小平方法

- ε_i实际上是无法观测的，不可能事先确定
  - 必须要对它作出若干基本假设，才能做进一步分析
- 高斯对此作了以下4条假设：
  - $E(\varepsilon_i) = 0$；
  - $Var(\varepsilon_i) = \sigma^2$；
  - $Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$；
  - 自变量X_i为非随机变量。
  - 满足以上4条假设的回归模型称为标准回归模型。

最小平方方法

- 采用最小平方方法估计未知回归参数β_0、 β_1
  - 估计准则：
  - 求β_0、 β_1的估计量$\widehat{\beta_0}$、 $\widehat{\beta_1}$，使得随机扰动误差项ε_i的平方和S最小。

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

  - 估计方法：
  - 对S分别求β_0、 β_1的偏导数，并使之为零。

最小平方方法

- 此时获得的估计量称为，
  - 未知回归参数的最小平方估计量（LSE）
  - β_0、 β_1的LSE：

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}[(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)]}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2}$$

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$$

最小平方法

- 利用样本1的数据说明计算$\widehat{\beta_1}$和$\widehat{\beta_0}$的方法：

$$\widehat{\beta_1} = \frac{61800 - (500)(1100)/10}{28400 - (500)^2/10} = 2$$

$$\widehat{\beta_0} = \frac{1}{10}(1100 - 2 \times 500) = 10$$

- 用10个样本企业数据拟合所得的产量—排污量回归直线方程：

$$\widehat{Y_i} = 10 + 2X_i$$

企业	产量X_i	排污量Y_i	$X_i Y_i$	X_i^2
1	30	73	2190	900
2	20	50	1000	400
3	60	128	7680	3600
4	80	170	13600	6400
5	40	87	3480	1600
6	50	108	5400	2500
7	60	135	8100	3600
8	30	69	2070	900
9	70	148	10360	4900
10	60	132	7920	3600
合计	500	1100	61800	28400

回归分析

- 基本概念
- 一元回归分析
  - 回归参数的估计
  - 回归模型的检验
  - 回归预测
- 多元回归分析
- 回归函数的形式

对回归模型的检验

- 在采用回归模型时，把自变量X与因变量Y之间的关系假设为：
  - 线性关系
- 这种假设是否适当？
  - 还需通过统计检验

对回归模型的检验

- 回归分析中的假设检验包含两个内容：
  - 检验变量之间的总体线性关系是否显著，检验自变量与因变量之间的关系能否用一个适当的回归模型表示；
  - 检验回归参数，检验回归模型中的每一个自变量对因变量的影响程度是否显著；
- 两种显著性统计检验在次序上不能颠倒：
  - 只有当回归模型所代表的变量之间的线性关系通过检验后，进一步检验模型中的个别回归参数才有意义。

总体线性关系的检验

- 以一元线性回归模型为例，讨论变量之间总体线性关系的显著性检验。
- 两种方法：
  - 方差分析
  - 考察残差图

方差分析

- 方差分析：
  - 对回归分析的初步计算结果进行总结的常用方法
  - Analysis of Variance （ANOVA）
  - 方差分析表的形式

Source of Variance 变差来源 SV	Degree of Freedom 自由度 df	Sum of Squares 平方和 SS	Mean Square 均方 MS	统计量 F
回归 Regression	1	$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$	MSR= SSR/1	F=MSR/MSE
残差 Residual	n-2	$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$	MSE= SSE/(n-2)	
总离差 Total	n-1	$SST = SSR + SSE = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$	——	——

方差分析

- *F*统计量的特征：
  - *F*值的高低直接反映因变量Y与自变量X之间线性关系的强弱；
  - *F*统计量服从自由度为(1, n-2)的*F*分布。
- 建立*F*检验的决策规则：
  - 设H_0：$\beta_1 = 0$； H_1：$\beta_1 \neq 0$
  - 若$F \leq F_\alpha(1, n-2)$，则接受$\beta_1 = 0$，即线性关系不显著
  - 若$F > F_\alpha(1, n-2)$，则接受$\beta_1 \neq 0$，即线性关系显著
  - $F_\alpha(1, n-2)$是*F*统计量的临界值，α是事先确定的显著水平，而$(1-\alpha)$表示检验的可靠程度。例如取α =0.01，由*F*分布表可以查到$F_{0.01}(1, n-2)$的值，而检验的可靠程度为99%。

方差分析

- 仍以样本1当中的产量、排污量数据为例。

企业	产量X_i	排污量Y_i			
1	30	73	70	1600	9
2	20	50	50	3600	0
3	60	128	130	400	4
4	80	170	170	3600	0
5	40	87	90	400	9
6	50	108	110	0	4
7	60	135	130	400	25
8	30	69	70	1600	1
9	70	148	150	1600	4
10	60	132	130	400	4
合计	500	1100	1100	13600	60

方差分析表

变差来源 SV	自由度 df	平方和 SS	均方 MS	统计量 F
回归	1	13600	13600	1813.3
残差	8	60	7.5	
总离差	9	13640	——	——

- 比较1813.3与$F_\alpha(1,8)$的大小就可以确定总体线性关系是否显著。

$$F_{0.1}(1,8) = 3.46$$

考察残差图

- 残差
  - $e_i = \left(Y_i - \widehat{Y_i}\right) \sim N(0, \sigma^2)$

- 对σ^2的估计
  - 最小平方估计量：对总离差进行分解

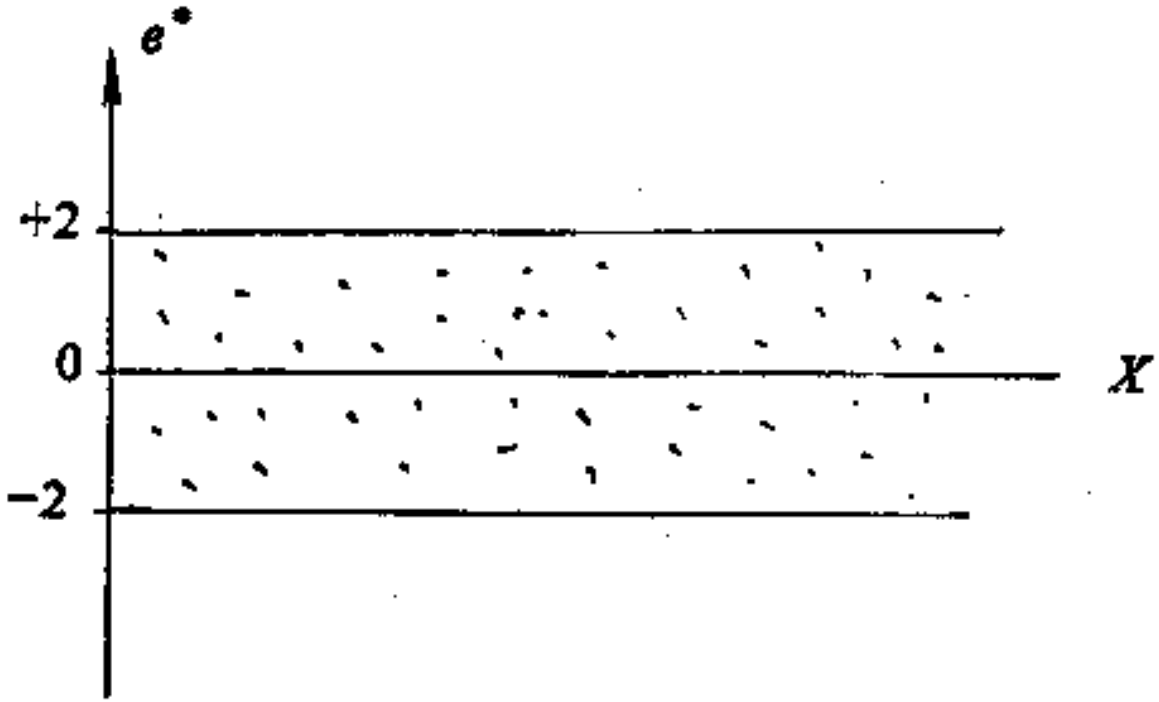
$$LSE\left(\widehat{\sigma^2}\right) = \frac{SSE}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{Y_i}\right)^2 = MSE$$

考察残差图

- 残差的标准化：
  - $e_i^* = (e_i/\sigma) \sim N(0,1)$
  - $P(|e_i^*| < 2) = 0.9545$，即e_i^*的值落在区间$(-2,+2)$上的概率应为95.45%
- 绘制标准化残差图：
  - X_i为横坐标，e_i^*为纵坐标
  - 将数据点(X_i, e_i^*)，$i = 1, \ldots, n$，画在平面图上

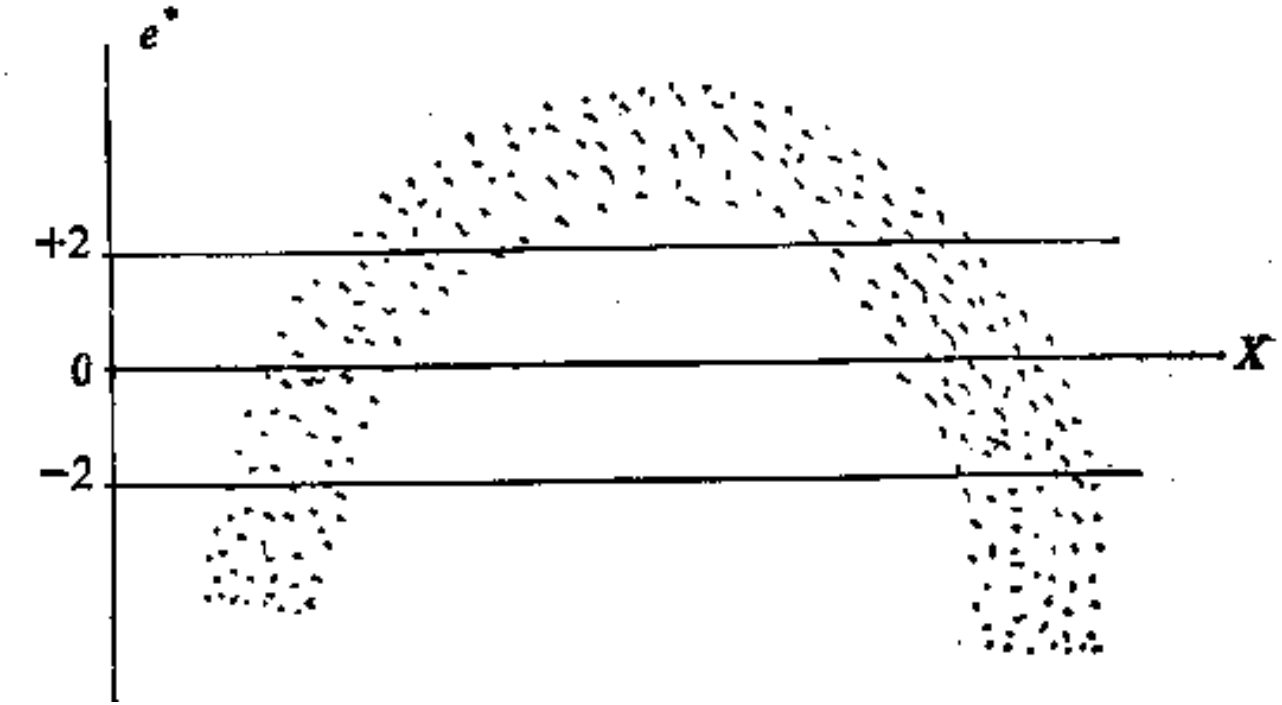
考察残差图

- 考察标准化残差图时，常遇到的4种情况：
  - （1）回归方程拟合良好，绝大多数数据点都在(-2,+2)水平带状区间内，且不带有任何系统趋势，完全随机的散布。



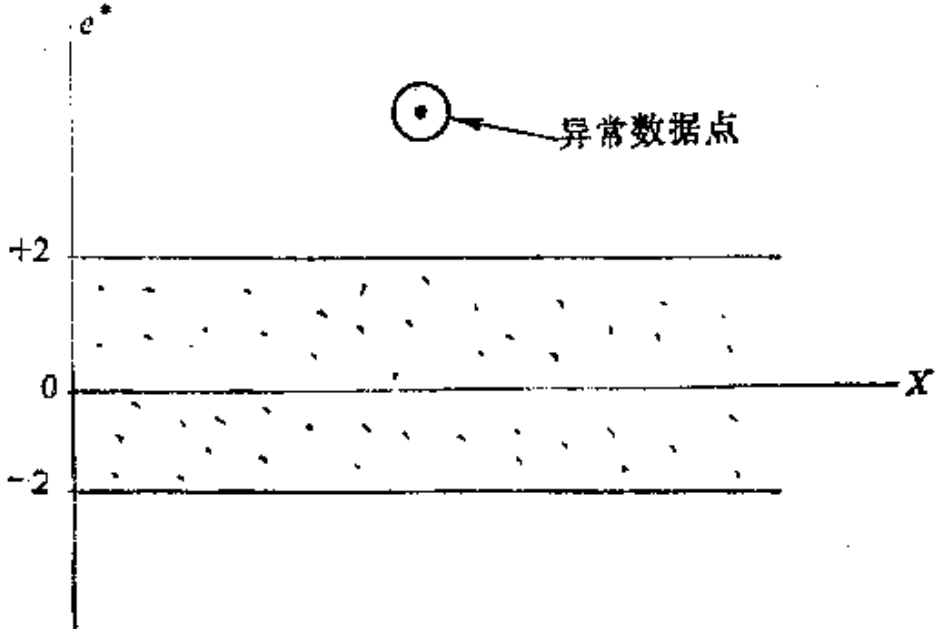
考察残差图

- 考察标准化残差图时，常遇到的4种情况：
  - （2）回归函数具有曲线形式，如果回归函数的形式应为曲线，但却采用了直线回归方程，标准残差图就会出现类似下图的形式。



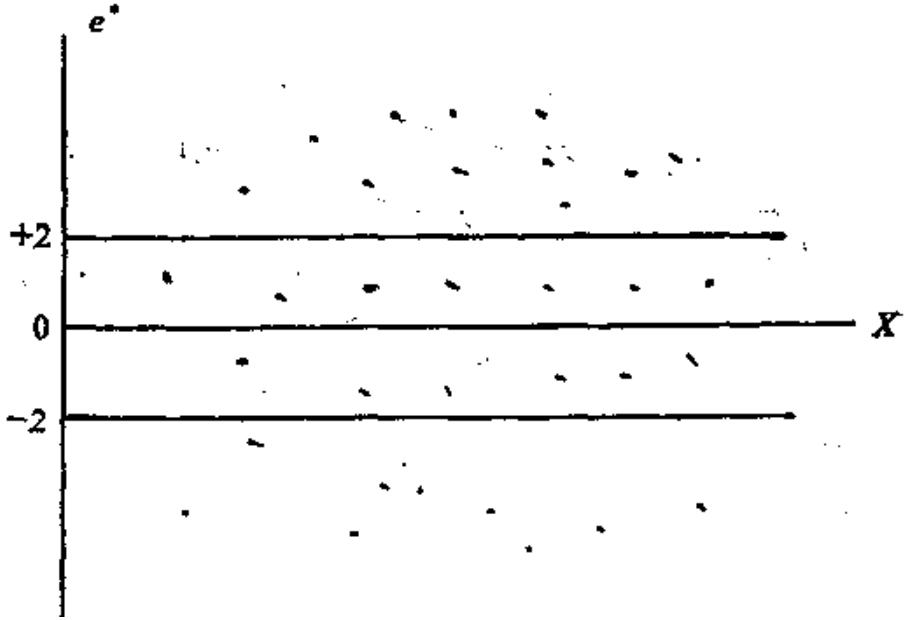
考察残差图

- 考察标准化残差图时，常遇到的4种情况：
  - （3）样本数据中有一个或数个异常点（outlier），如果由于错误或其它特殊因素的影响，样本数据中出现了个别异常点，那么在标准残差图里就会相应出现远离大多数数据点的一个或数个异常数据点。



考察残差图

- 考察标准化残差图时，常遇到的4种情况：
  - （4）回归方程拟合不充分：如果有许多数据点落在区间(-2,+2)的外面，就说明回归方程对数据的拟合是不充分的；原因可能是回归模型的函数形式选择不当，也可能是漏掉了一个（或数个）比较重要的自变量。



回归参数的检验

- 以一元回归模型为例，参数检验的任务
  - 对回归参数β_0、β_1进行统计检验，并求出其置信区间。
- 对β_1的检验：
  - β_1代表了X_i的单位变动对Y_i的影响程度，所以对β_1的检验首先考察这种影响程度是否与零有显著差异
  - 在一元回归模型中，对参数β_1的检验相当于检验总体线性关系

回归参数的检验

- 对β_1的检验过程：
  - 设H_0：$\beta_1=0$；H_1：$\beta_1 \neq 0$；
  - 检验统计量t：

$$t=\frac{\widehat{\beta_1}}{\sqrt{\widehat{\sigma^2}/S_{XX}}}\sim t_{n-2}$$

$$S_{XX}=\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$\widehat{\sigma^2}=\frac{SSE}{n-2}=\frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i-\widehat{Y_i}\right)^2$$

    - 服从自由度为(n-2)的t分布
  - 决策规则：
    - 若$|t| \leq t_{\alpha/2,n-2}$，则接受$\beta_1=0$，即X对Y的影响是不显著的；
    - 若$|t| > t_{\alpha/2,n-2}$，则接受$\beta_1 \neq 0$，即X对Y的影响是显著的

回归参数的检验

对β_0的检验过程

- 设H_0：$\beta_0 = \beta_0^*$；　H_1：$\beta_0 \neq \beta_0^*$
  - β_0^*是某个假设的特定值，如果$\beta_0^* = 0$，相当于检验回归直线是否过原点
- 检验统计量t：

$$t = \frac{\widehat{\beta_0} - \beta_0^*}{\sqrt{\widehat{\sigma^2}\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}} \sim t_{n-2}$$

  - 服从自由度为(n-2)的t分布
- 决策规则：
  - 若$|t| \leq t_{\alpha/2,n-2}$，则接受$\beta_0 = \beta_0^*$
  - 若$|t| > t_{\alpha/2,n-2}$，则接受$\beta_0 \neq \beta_0^*$

回归参数的检验

- 点估计量
  - 采用最小平方法获得的参数估计量是点估计量。
  - 由于样本与样本之间数据差异产生的波动，单一的参数点估计量很可能与参数的真值不同。
- 如何对点估计量的可靠性进行度量：
  - 通过点估计量围绕其真值变动的标准误差或者方差来度量
  - 根据点估计量的已知概率分布来确定它落在某一区间（置信区间）内的概率
  - 这一概率用$(1-\alpha)$来表示，又称置信水平

回归参数的检验

- 参数β_1的置信区间
  - 把参数β_1的标准误差估计量定义为：

$$S(\widehat{\beta_1}) = \sqrt{\widehat{\sigma^2}/S_{XX}}$$

  - 参数β_1的$100(1-\alpha)\%$置信区间为：

$$\widehat{\beta_1} \pm t_{\alpha/2,n-2} \cdot S(\widehat{\beta_1})$$

回归参数的检验

- 参数β_0的置信区间
  - 把参数β_0的标准误差估计量定义为：

$$S(\widehat{\beta_0}) = \sqrt{\widehat{\sigma^2}\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}$$

  - 参数β_0的$100(1-\alpha)\%$置信区间为：

$$\widehat{\beta_0} \pm t_{\alpha/2,n-2} \cdot S(\widehat{\beta_0})$$

回归预测问题

- 在一元回归模型中，预测是指：
  - 根据自变量X的某一已知值X_0求因变量Y的相应值Y_0的过程
  - X_0可以是样本数据中的某个数据；
  - X_0也可以是样本数据值域范围内的某一个取值；
  - 可以预测Y_0的平均值，$E(Y_0|X_0)=\beta_0+\beta_1X_0$，简称为均值预测；
  - 可以预测与X_0相对应的总体个别值（或称为特定值）Y_0，$Y_0=\beta_0+\beta_1X_0+\varepsilon_0$，简称为个别值预测

均值预测

- $E(Y_0|X_0)$的点估计量：

$$\widehat{Y_0} = \widehat{\beta_0} + \widehat{\beta_1} X_0$$

- $E(Y_0|X_0)$的点估计量与真值之差是一种预测误差
- 把Y_0的估计标准误差定义为：

$$S(\widehat{Y_0}) = \sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}}\right]}$$

- $E(Y_0|X_0)$的$100(1-\alpha)\%$置信区间为：

$$\widehat{Y_0} - t_{\frac{\alpha}{2},n-2} \cdot S(\widehat{Y_0}) \leq Y_0 \leq \widehat{Y_0} + t_{\frac{\alpha}{2},n-2} \cdot S(\widehat{Y_0})$$

个别值预测

- Y变量的总体个别值Y_0^*与Y_0的平均值$E(Y_0|X_0)$有：
  - 相同的预测值：

$$\widehat{Y_0^*} = \widehat{\beta_0} + \widehat{\beta_1} X_0$$

  - 不同的预测误差：
    - Y_0^*的预测区间比$E(Y_0|X_0)$的置信区间要宽

- 把Y_0^*的估计标准误差定义为

$$S(\widehat{Y_0^*}) = \sqrt{\widehat{\sigma^2}\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}}\right]}$$

- Y_0^*的$100(1-\alpha)\%$预测区间为：

$$\widehat{Y_0^*} - t_{\frac{\alpha}{2},n-2} \cdot S(\widehat{Y_0^*}) \leq \widehat{Y_0^*} \leq \widehat{Y_0^*} + t_{\frac{\alpha}{2},n-2} \cdot S(\widehat{Y_0^*})$$

回归分析

- 基本概念
- 一元回归分析
  - 回归参数的估计
  - 回归模型的检验
  - 回归预测
- 多元回归分析
- 回归函数的形式

多元回归分析

- 多元回归分析：
  - 二元回归分析的延伸
- 多元线性回归模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\ldots + \beta_p X_p + \varepsilon$$

- 多元线性回归方程：

$$E(Y|X = X_1, X = X_2, \ldots\ldots, X = X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\ldots + \beta_p X_p$$

其中，β_0，β_1，β_2，……，β_p是（$p+1$）个未知的回归参数；
$X_1, X_2, \ldots\ldots, X_p$为$p$个自变量；
Y为因变量；
ε为未知的随机扰动误差项。

多元回归分析

- 通过对$X_1, X_2, \ldots\ldots, X_p, Y$进行$n$项独立的观测，并取得
  - 样本：（$X_{i1}, X_{i2}, \ldots\ldots, X_{ip}, Y_i$）$i = 1, \ldots\ldots, n$
  - 模型：$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots\ldots + \beta_p X_{ip} + \varepsilon_i$
  - 相应的回归方程：$E(Y|X = X_1, X = X_2, \ldots\ldots, X = X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\ldots + \beta_p X_p$
- 多元正态线性回归模型：
  - 假设ε_i是独立、正态的随机变量，$\varepsilon_i \sim N(0, \sigma^2)$；
  - 假设p个自变量为非随机的变量，并且自变量之间不存在完全的或者接近完全的线性相关性；
  - Y_i也是一个正态随机变量，其分布为：

$$Y_i \sim N(\beta_0 + \beta_1 Xi_1 + \beta_2 X_{i2} + \cdots\ldots + \beta_p X_{ip}, \sigma^2) \quad i = 1, \ldots, n$$

多元回归分析

- 用矩阵形式讨论多元正态线性回归模型
- 设$k = p + 1$，模型可表示为：$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\underset{n \times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \underset{n \times k}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \qquad \underset{k \times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \underset{n \times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

因变量向量　　设计矩阵　　待估参数向量　　误差向量

多元回归模型的参数估计

- 待估参数的最小平方估计量：

$$(LSE)\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

多元回归模型的检验

变量之间线性关系的检验：

- 方法1：利用方差分析表进行F检验

变差来源	自由度df	平方和SS	均方MS	统计量F
回归	p	$SSR = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2$	$MSR = SSR/p$	$F = MSR/MSE$
残差	$n-p-1$	$SSE = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$	$MSE = SSE/(n-p-1)$	
总离差	$n-1$	$SST = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2$		

- 假设：H_0：$\beta_1 = \beta_2 = \ldots = \beta_p = 0$；$H_1$：$\beta_1\beta_2 \ldots \beta_p$中至少一个不等于零；
- 若统计量$F \le F_\alpha(p, n-p-1)$，接受H_0，说明变量之间的线性关系是不显著的；
- 若统计量$F > F_\alpha(p, n-p-1)$，接受H_1，说明变量之间的线性关系是显著的。

多元回归模型的检验

- 变量之间线性关系的检验：
  - 方法2：利用多元测定系数
  - 定义：

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

  - 值域：

$$0 \le R^2 \le 1$$

  - R^2是评价多元回归模型对变量之间线性关系代表性的一个指标，它的值越大，说明拟合优度越好
  - 为了去除自变量个数对R^2的影响，对它进行调整：

$$R_a^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-p-1}$$

多元回归模型的检验

- 回归参数的检验：
  - 假设：H_0：$\beta_j = \beta_{j0}$（β_{j0}是β_j的某个假设值，常取$\beta_{j0} = 0$）；H_1：$\beta_j \neq \beta_{j0}$；
  - 记

$$S^2(\hat{\boldsymbol{\beta}}) = \widehat{\sigma^2}(\mathbf{X}'\mathbf{X})^{-1} = (MSE)\begin{bmatrix} a_{00} & a_{01} & \cdots & a_{0p} \\ a_{10} & a_{11} & \cdots & a_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p0} & a_{p1} & \cdots & a_{pp} \end{bmatrix}$$

  - 检验统计量

$$t = \frac{\hat{\beta}_j - \beta_{j0}}{S(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{\sigma}\sqrt{a_{jj}}} = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{(MSE)a_{jj}}}, \qquad j = 0,1,2,\ldots,p$$

  - 若$|t| \leq t_{\frac{\alpha}{2},n-p-1}$，则接受$H_0$；$|t| > t_{\frac{\alpha}{2},n-p-1}$，则接受$H_1$。

多元回归模型的检验

- 回归参数的检验：
  - β_j的$100(1-\alpha)\%$置信区间为：

$$\widehat{\beta_j} - t_{\frac{\alpha}{2},n-p-1}\sqrt{(MSE)a_{jj}} \le \beta_j \le \widehat{\beta_j} + t_{\frac{\alpha}{2},n-p-1}\sqrt{(MSE)a_{jj}}$$
$$j = 0,1,2,\dots,p$$

多元回归的预测

均值预测

- $E(Y_0|\boldsymbol{X_0})$的估计量为：

$$\widehat{Y_0} = \mathbf{X'}_0\widehat{\boldsymbol{\beta}} \qquad X_0 = \begin{bmatrix} 1 \\ X_{01} \\ X_{02} \\ \vdots \\ X_{0p} \end{bmatrix}$$

- $E(Y_0|\boldsymbol{X_0})$的$100(1-\alpha)\%$置信区间为：

$$\widehat{Y_0} - t_{\frac{\alpha}{2},n-p-1}\sqrt{(MSE)\mathbf{X_0'}(\mathbf{X'X})^{-1}\mathbf{X_0}} \leq E(Y_0|\mathbf{X_0}) \leq \widehat{Y_0} + t_{\frac{\alpha}{2},n-p-1}\sqrt{(MSE)\mathbf{X_0'}(\mathbf{X'X})^{-1}\mathbf{X_0}}$$

多元回归的预测

- 个别值预测：
  - 个别值Y_0的估计量为：

$$\widehat{Y_0} = \mathbf{X'}_0\widehat{\boldsymbol{\beta}} \qquad X_0 = \begin{bmatrix} 1 \\ X_{01} \\ X_{02} \\ \vdots \\ X_{0p} \end{bmatrix}$$

  - Y_0的$100(1-\alpha)\%$置信区间为：

$$\widehat{Y_0} - t_{\frac{\alpha}{2},n-p-1}\sqrt{(MSE)(1+\mathbf{X_0'}(\mathbf{X'X})^{-1}\mathbf{X_0})} \le Y_0$$
$$\le \widehat{Y_0} + t_{\frac{\alpha}{2},n-p-1}\sqrt{(MSE)(1+\mathbf{X_0'}(\mathbf{X'X})^{-1}\mathbf{X_0})}$$

多元回归模型计算举例

- 例1：某企业开发了新型环保产品，目前在国内有15个销售点。为了分析该产品的市场情况，决定以销售点所在地区的人口（万人）为自变量X_1，以市场地区的人均年收入（元）为自变量X_2，以环保产品的年销售量（套）为因变量Y，并已取得样本数据的初步计算结果。设Y与X_1、 X_2的关系为线性且满足多元 线性回归模型的各种假定，取$\alpha = 0.05$，试进行多元回归分析。

$$n = 15, \sum X_{i1} = 3626, \sum X_{i2} = 44428, \sum Y_i = 2259, \sum X_{i1}^2 = 1067614,$$

$$\sum X_{i2}^2 = 139063428, \sum Y_i^2 = 394107, \sum X_{i1} X_{i2} = 11419181,$$

$$\sum X_{i1} Y_i = 647107, \sum X_{i2} Y_i = 7096619$$

多元回归模型计算举例

- 解：根据题意，Y和X_1、X_2之间的样本模型为

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad i = 1,2,3,\dots,15$$

- 其矩阵形式为 $\mathbf{Y} = \mathbf{X}\ \boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- （1）估计回归参数

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{15,1} \\ X_{12} & X_{22} & \cdots & X_{15,2} \end{bmatrix} \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{15,1} & X_{15,2} \end{bmatrix} = \begin{bmatrix} n & \sum X_{i1} & \sum X_{i2} \\ \sum X_{i1} & \sum X_{i1}^2 & \sum X_{i1}X_{i2} \\ \sum X_{i2} & \sum X_{i1}X_{i2} & \sum X_{i2}^2 \end{bmatrix}$$

$$= \begin{bmatrix} 15 & 3626 & 44428 \\ 3626 & 1067614 & 11419181 \\ 44428 & 11419181 & 139063428 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{15,1} \\ X_{12} & X_{22} & \cdots & X_{15,2} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{15} \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_{i1}Y_i \\ \sum X_{i2}Y_i \end{bmatrix} = \begin{bmatrix} 2259 \\ 647107 \\ 7096619 \end{bmatrix}$$

多元回归模型计算举例

估计回归参数

- 待估参数向量的最小平方估计量为

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$$

$$= \begin{bmatrix} 1.2463484 & 2.1296642 \times 10^{-4} & -4.1567125 \times 10^{-4} \\ 2.1296642 \times 10^{-4} & 7.7329030 \times 10^{-6} & -7.0302518 \times 10^{-7} \\ -4.1567125 \times 10^{-4} & -7.0302518 \times 10^{-7} & 1.9771851 \times 10^{-7} \end{bmatrix} \times \begin{bmatrix} 2259 \\ 647107 \\ 7096619 \end{bmatrix}$$

$$= \begin{bmatrix} 3.4526 \\ 0.4960 \\ 0.0092 \end{bmatrix} = \begin{bmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \\ \widehat{\beta_2} \end{bmatrix}$$

- 因此，估计的回归方程为：

$$\hat{Y} = 3.4526 + 0.4960X_1 + 0.0092X_2$$

多元回归模型计算举例

- （2）建立方差分析表（ANOVA）

$$\mathbf{Y}'\mathbf{Y} = \sum Y_i^2 = 394107 \qquad n\bar{Y}^2 = 15 \times \left[\frac{1}{15}\sum Y_i\right]^2 = 15 \times \left[\frac{1}{15} \times 2259\right]^2 = 340205.4$$

$$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = [3.4526 \quad 0.4960 \quad 0.0092]\begin{bmatrix} 2259 \\ 647107 \\ 7096619 \end{bmatrix} = 394050.116$$

$$SST = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 = 394107 - 340205.4 = 53901.6$$

$$SSE = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = 394107 - 394050.116 = 56.884$$

$$SSR = SST - SSE = 53901.6 - 56.884 = 53844.716$$

变差来源	自由度	平方和	均方	统计量F
回归	2	SSR=53844.716	MSR=26922.358	F=5680
残差	12	SSE=56.884	MSE=4.740	
总离差	14	SST=53901.600	——	——

多元回归模型计算举例

- （3）检验总体线性关系的显著性
- 设 H_0：$\beta_1 = \beta_2 = 0$；H_1：β_1和β_2中至少有一个不等于零
- 因为检验统计量 $F = 5680 \geq F_{0.05}(2,12) = 3.89$，所以在$\alpha = 0.05$的水平上接受$H_1$，即该环保产品销售量$Y$跟销售地区人口$X_1$和销售地区人均年收入$X_2$之间存在的线性关系是显著的。
- （4）考察回归方程的拟合优度
- 计算多元测定系数和调整后的多元测定系数

$$R^2 = \frac{SSR}{SST} = \frac{53844.716}{53901.6} = 0.9989 \qquad R_a^2 = 1 - (1 - 0.9989)\left(\frac{15-1}{15-2-1}\right) = 0.9987$$

- 多元测定系数的值较高，说明该回归方程拟合良好。

多元回归模型计算举例

- （5）检验回归参数的显著性
- 检验H_0： $\beta_1=0$， H_1： $\beta_1 \neq 0$
- 因为：

$$a_{11} = 7.732903 \times 10^{-6} \qquad \widehat{\sigma^2} = MSE = 4.74 \qquad \widehat{\beta_1} = 0.496$$

$$t = \frac{\widehat{\beta_1} - 0}{\sqrt{\widehat{\sigma^2} a_{11}}} = \frac{0.496}{\sqrt{4.74 \times 7.732903 \times 10^{-6}}} = 81.926 > t_{0.025,12} = 2.179$$

- 所以在$\alpha = 0.05$的水平上接受H_1，即该环保产品销售量Y跟销售地区人口数X_1之间存在显著的线性关系。β_1的95%置信区间为：

$$0.496 - 2.179 \times 0.006054 \leq \beta_1 \leq 0.496 + 2.179 \times 0.006054$$

$$0.4828 \leq \beta_1 \leq 0.5092$$

多元回归模型计算举例

- （5）检验回归参数的显著性
- 检验H_0： $\beta_2=0$， H_1： $\beta_2 \neq 0$
- 因为：

$$a_{22} = 1.9771851 \times 10^{-7} \qquad \widehat{\sigma^2} = MSE = 4.74 \qquad \widehat{\beta_1} = 0.0092$$

$$t = \frac{\widehat{\beta_2} - 0}{\sqrt{\widehat{\sigma^2} a_{22}}} = \frac{0.0092}{\sqrt{4.74 \times 1.9771851 \times 10^{-7}}} = 9.508 > t_{0.025,12} = 2.179$$

- 所以在$\alpha = 0.05$的水平上接受H_1，即该环保产品销售量Y跟销售地区人均年收入X_2之间存在显著的线性关系。β_2的95%置信区间为：

$$0.0092 - 2.179 \times 0.000968 \leq \beta_2 \leq 0.0092 + 2.179 \times 0.000968$$

$$0.0071 \leq \beta_2 \leq 0.00113$$

多元回归模型计算举例

- （6）预测
- 该企业准备为它的产品在一个新地区开设销售点。已知新地区的人口为220（万人），人均年收入2500（元），试预测产品在新地区的年销售量Y的平均值$E(Y_0|\mathbf{X_0})$和个别值Y_0。
- 预测产品在新地区的年销售量Y的平均值$E(Y_0|\mathbf{X_0})$。
- 已知

$$\mathbf{X_0}=\begin{bmatrix}1\\220\\2500\end{bmatrix}$$

- $E(Y_0|\mathbf{X_0})$的估计值为

$$\widehat{Y_0}=X_0'\hat{\beta}=[1\quad 220\quad 2500]\begin{bmatrix}3.4526\\0.4960\\0.0092\end{bmatrix}=135.5726$$

多元回归模型计算举例

- （6）预测
- 预测产品在新地区的年销售量Y的平均值$E(Y_0|\mathbf{X_0})$。
- 因为

$$\mathbf{X_0'}(\mathbf{X'X})^{-1}\mathbf{X_0} = [1 \quad 220 \quad 2500]\,(\mathbf{X'X})^{-1}\begin{bmatrix}1\\220\\2500\end{bmatrix} = 0.09835$$

$$S(\hat{Y}_0) = \sqrt{(MSE)\mathbf{X_0'}(\mathbf{X'X})^{-1}\mathbf{X_0}} = \sqrt{4.74 \times 0.09835} = 0.6828$$

$$t_{0.025,12} = 2.179$$

- 所以，$E(Y_0|\mathbf{X_0})$的95%置信区间为

$$135.5726 - 2.179 \times 0.6828 \le E(Y_0|\mathbf{X_0}) \le 135.5726 + 2.179 \times 0.6828$$

$$134.09 \le E(Y_0|\mathbf{X_0}) \le 137.06$$

多元回归模型计算举例

- （6）预测
- 预测产品在新地区的年销售量Y的个别值Y_0。
- Y_0的点估计值仍然为

$$\hat{Y}_0 = X_0'\hat{\beta} = [1 \quad 220 \quad 2500]\begin{bmatrix}3.4526\\0.4960\\0.0092\end{bmatrix} = 135.5726$$

- 因为

$$S(\widehat{Y_0^*}) = \sqrt{(MSE)[1 + \mathbf{X_0'}(\mathbf{X'X})^{-1}\mathbf{X_0}]} = \sqrt{4.74 \times 1.09835} = 2.2817$$

- 所以， Y_0的95%置信区间为

$$135.5726 - 2.179 \times 2.2817 \le Y_0 \le 135.5726 + 2.179 \times 2.2817$$

$$130.60 \le Y_0 \le 140.54$$

多元回归模型计算举例

- 例2：已知历年实测的湖水中化学耗氧量COD的浓度与相应的环境自然经济资料。根据已知数据分析COD与环境、自然和经济状况之间的关系。

项目 年份	COD浓度 y（mg/l）	农业产量 x_1（亿斤）	工业总产值 x_2（亿元）	湖泊水位 x_3（米）
1960	2.50	0.25	4.00	3.17
1975	3.63	0.92	21.10	3.24
1976	3.15	0.87	29.10	3.02
1977	2.52	0.60	33.00	3.24
1978	4.06	0.63	37.50	2.63
1979	3.72	0.65	42.40	2.80
1980	2.82	0.42	49.25	3.65
1981	3.31	0.40	50.00	2.97

多元回归模型计算举例

- 解：
  - （1）建立回归模型：
  - 设：$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon$
  - 其中，b_0、 b_1、 b_2、 b_3是待估参数， ε是误差项。
  - （2）求回归方程：
  - 经计算获得各参数的估计量和回归方程（计算过程略）

$$\widehat{b_0} = 5.6609, \widehat{b_1} = 0.9613, \widehat{b_2} = 0.011, \widehat{b_3} = -1.0948$$

$$\hat{y} = 5.6609 + 0.9613 x_1 + 0.0011 x_2 - 1.0948 x_3$$

  - （3）显著性检验（给定$\alpha = 0.25$）
  - 方差分析

多元回归模型计算举例

- 方差分析表

方差来源	自由度	平方和	均方	F值
回归	3	1.6210	0.5373	3.1022
残差	4	0.6928	0.1732	
总计	7	2.3048		

- 因为$F = 3.1022 > F_{0.25}(3,4) = 2.05$，所以回归方程总体线性显著。
- 回归系数的显著性检验
- （计算过程略）在给定的置信水平下，x_1、x_3对y的影响显著，x_2对y的影响不显著。
- 将x_2从模型中剔除，在y与x_1、x_3之间重新进行回归分析。

多元回归模型计算举例

- （4）在y与x_1、 x_3之间建立回归方程
- （计算过程略）$\hat{y} = 6.1297 + 0.9411x_1 - 1.1241x_3$
- （5）再次进行显著性检验
- 方差分析

方差来源	自由度	平方和	均方	F值
回归	2	1.4099	0.7050	3.9387
残差	5	0.8948	0.1790	
总计	7	2.3048		

- 因为$F = 3.9387 > F_{0.25}(2,5) = 1.85$，所以新的回归方程显著。

多元回归模型计算举例

- （5）再次进行显著性检验
- 回归系数的显著性检验
- （计算过程略）在给定的置信水平下，x_1、x_3对y的影响显著，新的回归方程有实际意义。
- （6）课后练习：
- 基于该课程案例，给出完整的多元回归过程。

回归分析

- 基本概念
- 一元回归分析
  - 回归参数的估计
  - 回归模型的检验
  - 回归预测
- 多元回归分析
- 回归函数的形式

回归函数的形式

- 如果自变量与因变量之间的关系为某种曲线关系，但是回归分析中采用的回归函数是直线形式，那么
  - 模型对样本数据的代表性就很差。这一点会反映在：
    - 统计检验
    - 残差图
- 纠正办法之一：
  - 对样本数据进行某种变换，以使回归模型对于变换后的样本数据，其直线函数关系变得适宜；
  - transformation on the data

回归函数的形式

- 常用的变换方法：
  - 对数变换：
    - 半对数：只对变量X和Y中的一个进行对数变换。

$$Y = a_0 a_1^X \varepsilon \quad \Rightarrow \quad \lg Y = \lg a_0 + X \lg a_1 + \lg \varepsilon \quad \Rightarrow \quad Y' = \beta_0 + \beta_1 X + \varepsilon'$$

    - 双对数：对变量X和Y都进行对数变换。

$$Y = \beta_0 X^{\beta_1} \varepsilon \quad \Rightarrow \quad \lg Y = \lg \beta_0 + \beta_1 \lg X + \lg \varepsilon \quad \Rightarrow \quad Y' = \alpha + \beta_1 X' + \varepsilon'$$

  - 倒数变换

$$Y = \beta_0 + \frac{\beta_1}{X} + \varepsilon \quad \Rightarrow \quad Y = \beta_0 + \beta_1 X' + \varepsilon$$

回归函数的形式

- 常用的变换方法：
  - 对数和倒数联合变换：

$$Y = \frac{1}{a + b\exp(\beta_0 + \beta_1 X + \varepsilon)} \quad \Rightarrow \quad Y' = \ln\left[\left(\frac{1}{Y} - a\right)/b\right] \quad \Rightarrow \quad Y' = \beta_0 + \beta_1 X + \varepsilon$$

  - 多项式回归

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_m X^m + \varepsilon \quad \Rightarrow \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X^m + \varepsilon$$

回归分析总结

- 基本概念
  - 起源、研究任务、变量之间的统计关系、回归模型、回归函数、 PRF、SRF
- 回归参数估计
  - 最小平方法
- 回归模型检验
  - 总体线性检验：方差分析、残差图、多元测定系数
  - 参数检验：统计假设检验、置信区间
- 回归预测
  - 均值预测、个别值预测

中山大學
SUN YAT-SEN UNIVERSITY

谢谢！