第11章　一元线性回归

PowerPoint

第11章　一元线性回归

11.1　变量间关系的度量
11.2　一元线性回归
11.3　利用回归方程进行估计和预测
11.4　残差分析

学习目标

1. 相关关系的分析方法
2. 一元线性回归的基本原理和参数的最小二乘估计
3. 回归直线的拟合优度
4. 回归方程的显著性检验
5. 利用回归方程进行估计和预测
6. 用 **Excel** 进行回归

11.1 变量间关系的度量

11.1.1 变量间的关系
11.1.2 相关关系的描述与测度
11.1.3 相关系数的显著性检验

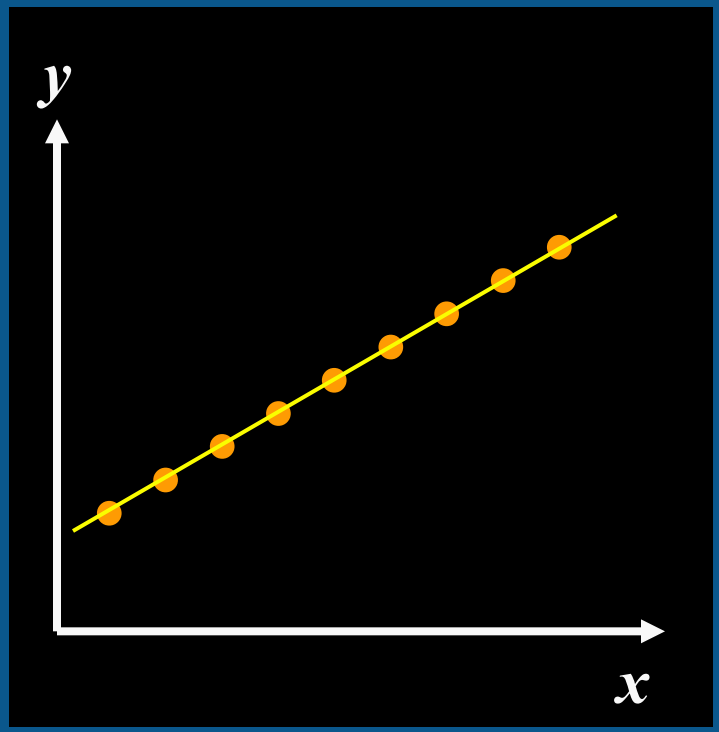
作者：贾俊平，中国人民大学统计学院

变量间的关系

作者：贾俊平，中国人民大学统计学院

函数关系

1. 是一一对应的确定关系
2. 设有两个变量 x 和 y，变量 y 随变量 x 一起变化，并完全依赖于 x，当变量 x 取某个数值时，y 依确定的关系取相应的值，则称 y 是 x 的函数，记为 $y = f(x)$，其中 x 称为自变量，y 称为因变量
3. 各观测点落在一条线上



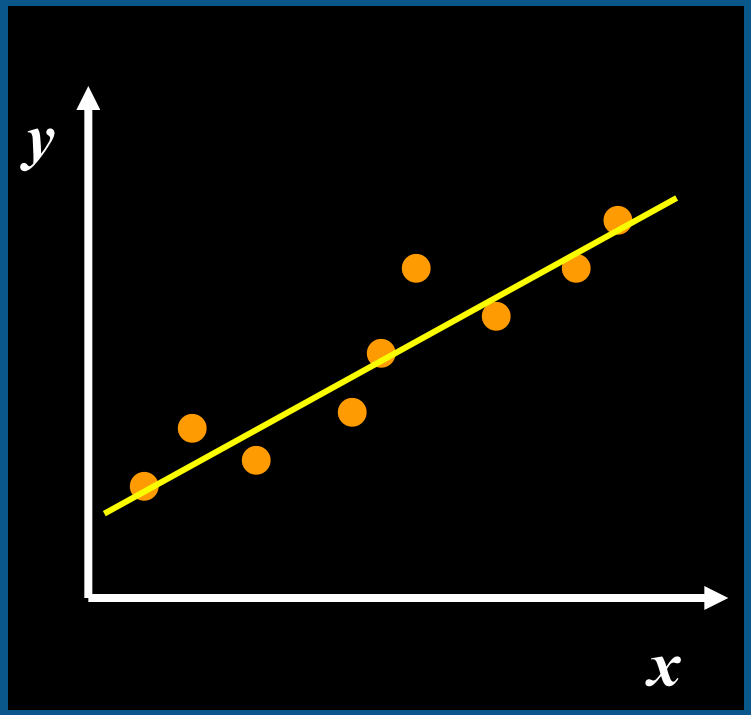
函数关系
(几个例子)

- 某种商品的销售额y与销售量x之间的关系可表示为 $y = px$ (p 为单价)
- 圆的面积S与半径R之间的关系可表示为 $S=\pi R^2$
- 企业的原材料消耗额y与产量x_1、单位产量消耗x_2、原材料价格x_3之间的关系可表示为

$$y = x_1 x_2 x_3$$

相关关系
(correlation)

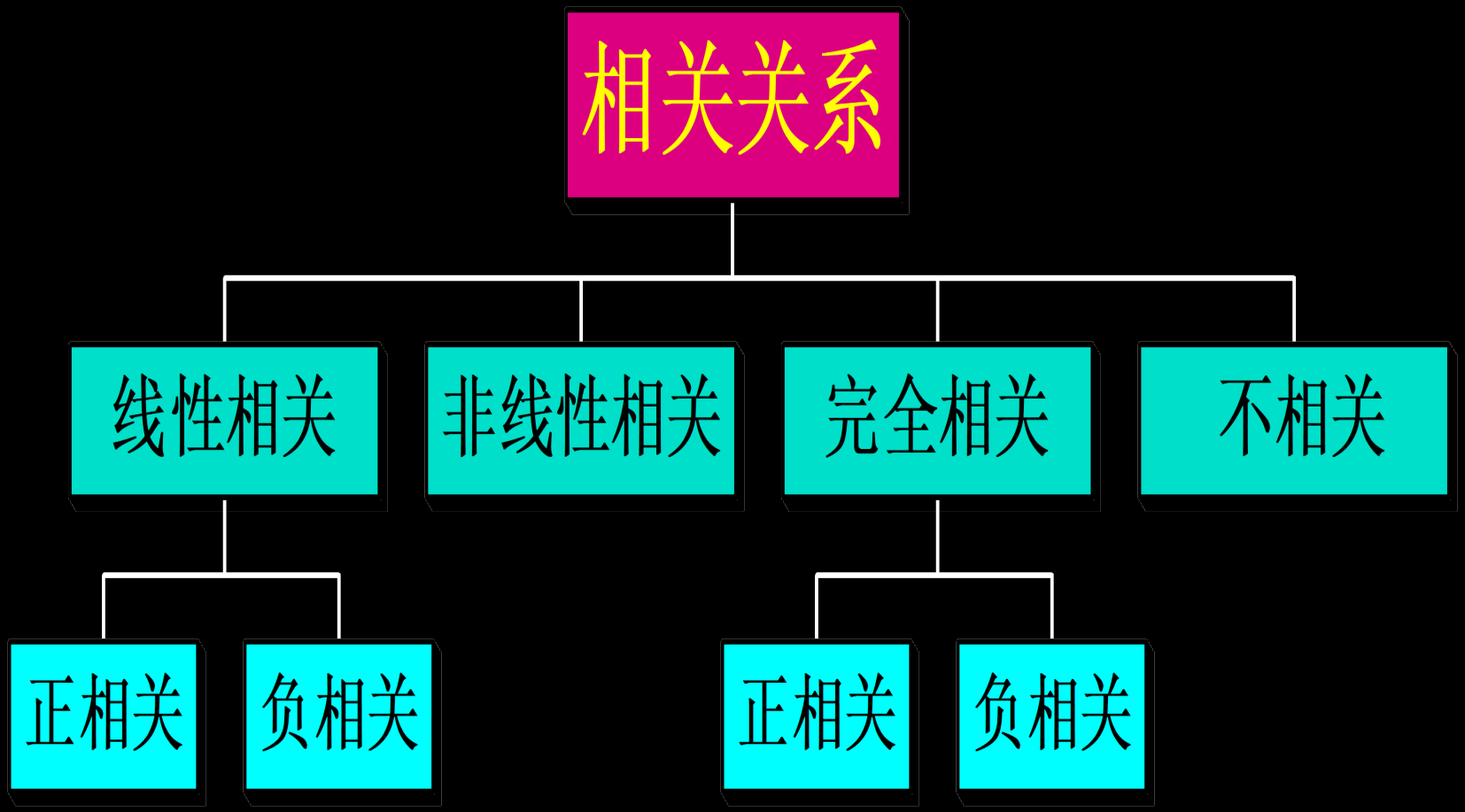
1. 变量间关系不能用函数关系精确表达
2. 一个变量的取值不能由另一个变量唯一确定
3. 当变量 x 取某个值时，变量 y 的取值可能有几个
4. 各观测点分布在直线周围



相关关系
(几个例子)

- 父亲身高y与子女身高x之间的关系
- 收入水平y与受教育程度x之间的关系
- 粮食单位面积产量y与施肥量x_1、降雨量x_2、温度x_3之间的关系
- 商品的消费量y与居民收入x之间的关系
- 商品销售额y与广告费支出x之间的关系

相关关系
(类型)



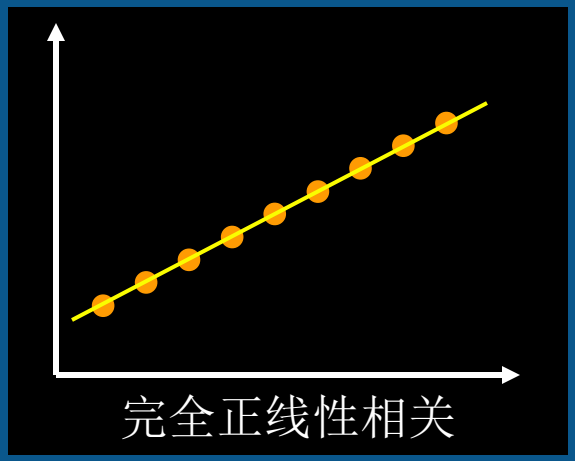
相关关系的描述与测度
(散点图)

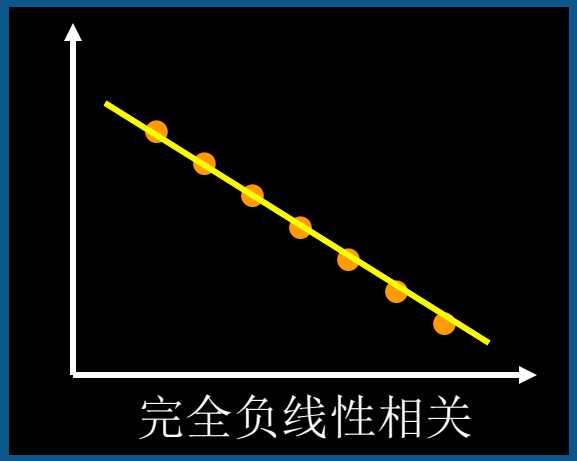
作者：贾俊平，中国人民大学统计学院

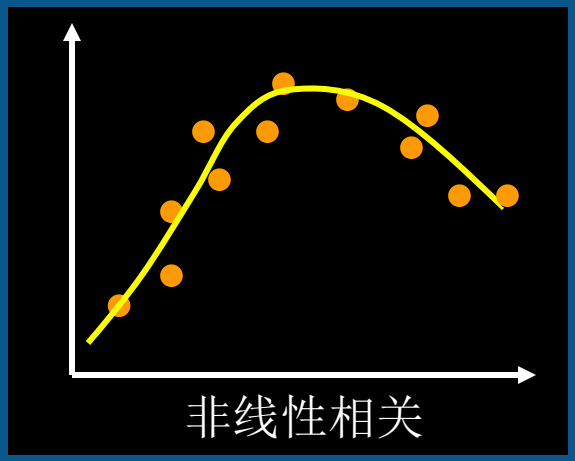
相关分析及其假定

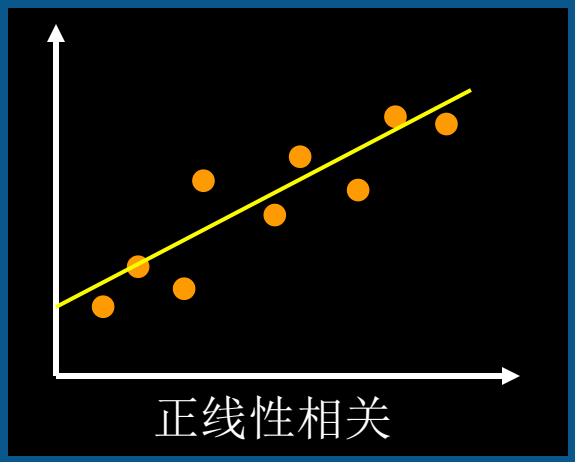
1. 相关分析要解决的问题
   - 变量之间是否存在关系？
   - 如果存在关系，它们之间是什么样的关系？
   - 变量之间的关系强度如何？
   - 样本所反映的变量之间的关系能否代表总体变量之间的关系？
2. 为解决这些问题，在进行相关分析时，对总体有以下两个主要假定
   - 两个变量之间是线性关系
   - 两个变量都是随机变量

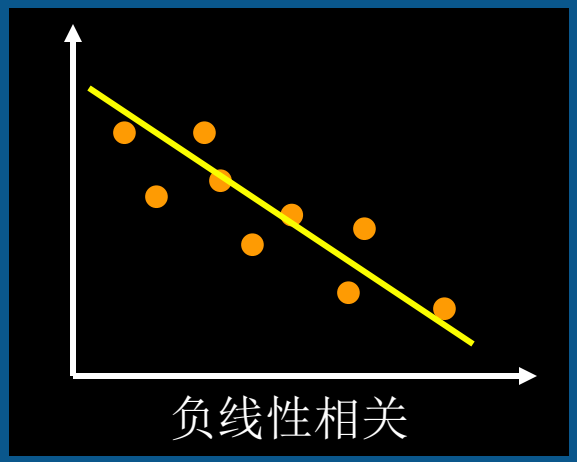
散点图
(scatter diagram)

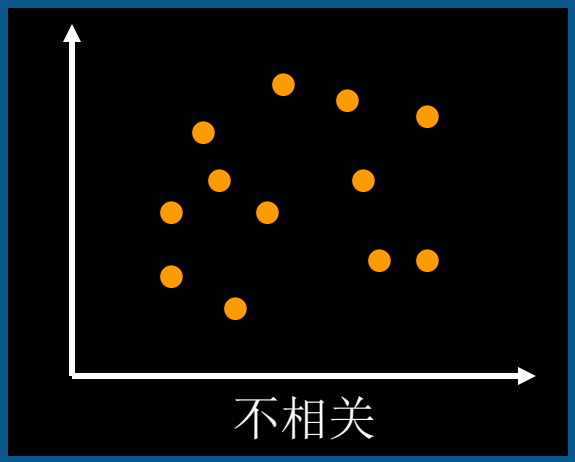












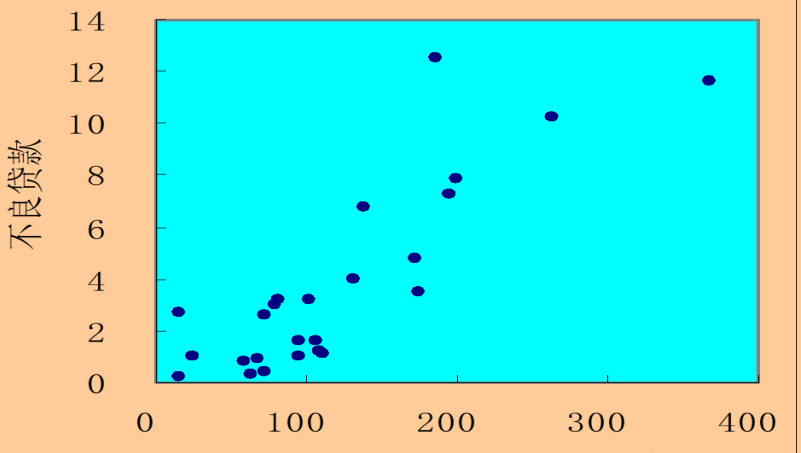
散点图
(例题分析)

【例】一家大型商业银行在多个地区设有分行，其业务主要是进行基础设施建设、国家重点项目建设、固定资产投资等项目的贷款。近年来，该银行的贷款额平稳增长，但不良贷款额也有较大比例的增长，这给银行业务的发展带来较大压力。为弄清楚不良贷款形成的原因，管理者希望利用银行业务的有关数据做些定量分析，以便找出控制不良贷款的办法。下面是该银行所属的25家分行的有关业务数据

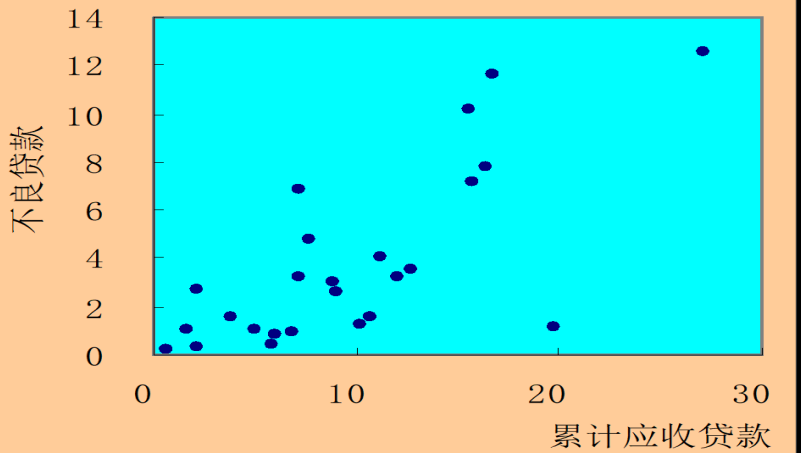
散点图
(例题分析)

	A	B	C	D	E	F
1	分行编号	不良贷款（亿元）	各项贷款余额（亿元）	本年累计应收贷款（亿元）	贷款项目个数（个）	本年固定资产投资额（亿元）
2	1	0.9	67.3	6.8	5	51.9
3	2	1.1	111.3	19.8	16	90.9
4	3	4.8	173.0	7.7	17	73.7
5	4	3.2	80.8	7.2	10	14.5
6	5	7.8	199.7	16.5	19	63.2
7	6	2.7	16.2	2.2	1	2.2
8	7	1.6	107.4	10.7	17	20.2
9	8	12.5	185.4	27.1	18	43.8
10	9	1.0	96.1	1.7	10	55.9
11	10	2.6	72.8	9.1	14	64.3
12	11	0.3	64.2	2.1	11	42.7
13	12	4.0	132.2	11.2	23	76.7
14	13	0.8	58.6	6.0	14	22.8
15	14	3.5	174.6	12.7	26	117.1
16	15	10.2	263.5	15.6	34	146.7
17	16	3.0	79.3	8.9	15	29.9
18	17	0.2	14.8	0.6	2	42.1
19	18	0.4	73.5	5.9	11	25.3
20	19	1.0	24.7	5.0	4	13.4
21	20	6.8	139.4	7.2	28	64.3
22	21	11.6	368.2	16.8	32	163.9
23	22	1.6	95.7	3.8	10	44.5
24	23	1.2	109.6	10.3	14	67.9
25	24	7.2	196.2	15.8	16	39.7
26	25	3.2	102.2	12.0	10	97.1

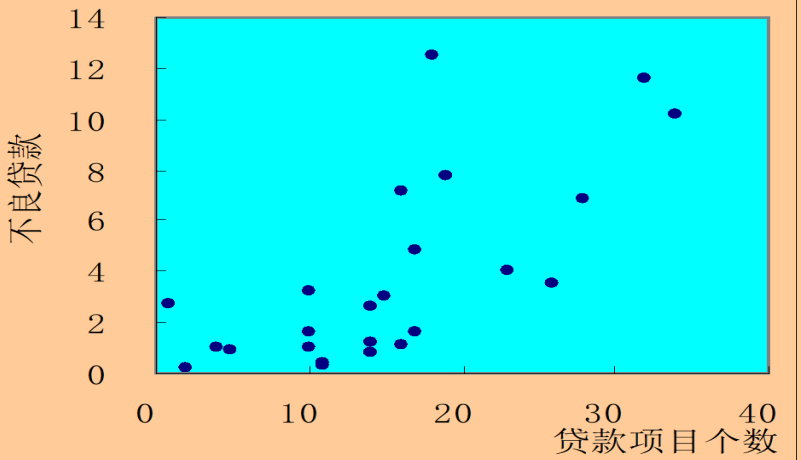
散点图
(不良贷款对其他变量的散点图)



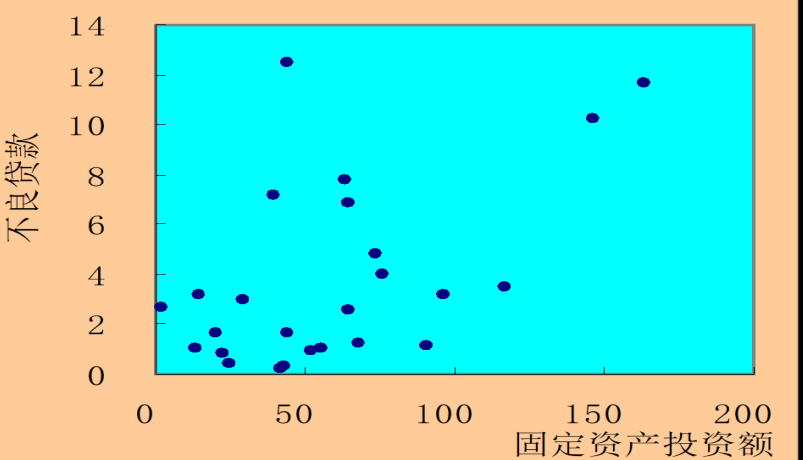
不良贷款与贷款余额的散点图



不良贷款与累计应收贷款的散点图



不良贷款与贷款项目个数的散点图



不良贷款与固定资产投资额的散点图

相关关系的描述与测度
(相关系数)

作者：贾俊平，中国人民大学统计学院

相关系数 (correlation coefficient)

1. 度量变量之间关系强度的一个统计量
2. 对两个变量之间线性相关强度的度量称为简单相关系数
3. 若相关系数是根据总体全部数据计算的，称为总体相关系数，记为ρ
4. 若是根据样本数据计算的，则称为样本相关系数，简称为相关系数，记为 r
   - 也称为线性相关系数(linear correlation coefficient)
   - 或称为Pearson相关系数 (Pearson's correlation coefficient)

相关系数
(计算公式)

- 样本相关系数的计算公式

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\cdot\sum(y-\bar{y})^2}}$$

或化简为 $$r=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-\left(\sum x\right)^2}\cdot\sqrt{n\sum y^2-\left(\sum y\right)^2}}$$

相关系数的性质

性质1：r 的取值范围是 [-1,1]

- $|r|=1$，为完全相关
  - $r=1$，为完全正相关
  - $r=-1$，为完全负正相关
- $r=0$，不存在*线性*相关关系
- $-1\le r<0$，为负相关
- $0<r\le 1$，为正相关
- $|r|$越趋于1表示关系越强；$|r|$越趋于0表示关系越弱

相关系数的性质

性质2：r具有对称性。即x与y之间的相关系数和y与x之间的相关系数相等，即$r_{xy} = r_{yx}$

性质3：r数值大小与x和y原点及尺度无关，即改变x和y的数据原点及计量尺度，并不改变r数值大小

性质4：仅仅是x与y之间线性关系的一个度量，它不能用于描述非线性关系。这意为着，r=0只表示两个变量之间不存在线性相关关系，并不说明变量之间没有任何关系

性质5：r虽然是两个变量之间线性关系的一个度量，却不一定意味着x与y一定有因果关系

相关系数的经验解释

1. $|r|\geq 0.8$时，可视为两个变量之间高度相关
2. $0.5\leq|r|<0.8$时，可视为中度相关
3. $0.3\leq|r|<0.5$时，视为低度相关
4. $|r|<0.3$时，说明两个变量之间的相关程度极弱，可视为不相关
5. 上述解释必须建立在对相关系数的显著性进行检验的基础之上

相关系数
(例题分析)

用Excel计算相关系数

	A	B	C	D	E	F
1		不良贷款	各项贷款余额	各项贷款余额	各项贷款余额	固定资产投资额
2	不良贷款	1				
3	各项贷款余额	0.843571	1			
4	各项贷款余额	0.731505	0.678772	1		
5	各项贷款余额	0.700281	0.848416	0.585831	1	
6	固定资产投资额	0.518518	0.779702	0.472431	0.746646	1

相关系数的显著性检验

作者：贾俊平，中国人民大学统计学院

相关系数的显著性检验
(检验的步骤)

1. 检验两个变量之间是否存在线性相关关系
2. 等价于对回归系数 β_1的检验
3. 采用R.A.Fisher提出的 t 检验
4. 检验的步骤为
   - 提出假设：H_0：$\rho = 0$；H_1：$\rho \neq 0$
   - 计算检验的统计量：$t = |r|\sqrt{\dfrac{n-2}{1-r^2}} \sim t(n-2)$
   - 确定显著性水平α，并作出决策
     - 若$|t|>t_{\alpha/2}$，拒绝H_0
     - 若$|t|<t_{\alpha/2}$，不拒绝H_0

相关系数的显著性检验 (例题分析)

- 对不良贷款与贷款余额之间的相关系数进行显著性检验(α=0.05)

1. 提出假设：H_0：$\rho = 0$；H_1：$\rho \neq 0$
2. 计算检验的统计量

$$t = |0.8436| \sqrt{\frac{25-2}{1-0.8436^2}} = 7.5344$$

3. 根据显著性水平α=0.05，查t分布表得$t_{\alpha/2}(n\text{-}2)$=2.069
   - 由于$|t|$=7.5344>$t_{\alpha/2}$(25-2)=2.069，拒绝H_0，不良贷款与贷款余额之间存在着显著的正线性相关关系

相关系数的显著性检验
(例题分析)

各相关系数检验的统计量

	A	B	C	D	E
1		不良贷款	各项贷款余额	累计应收贷款	贷款项目个数
2	各项贷款余额	7.533515			
3	累计应收贷款	5.145188	4.432870		
4	贷款项目个数	4.704564	7.686824	3.466726	
5	固定资产投资额	2.908224	5.971918	2.570663	5.382848

11.2 一元线性回归

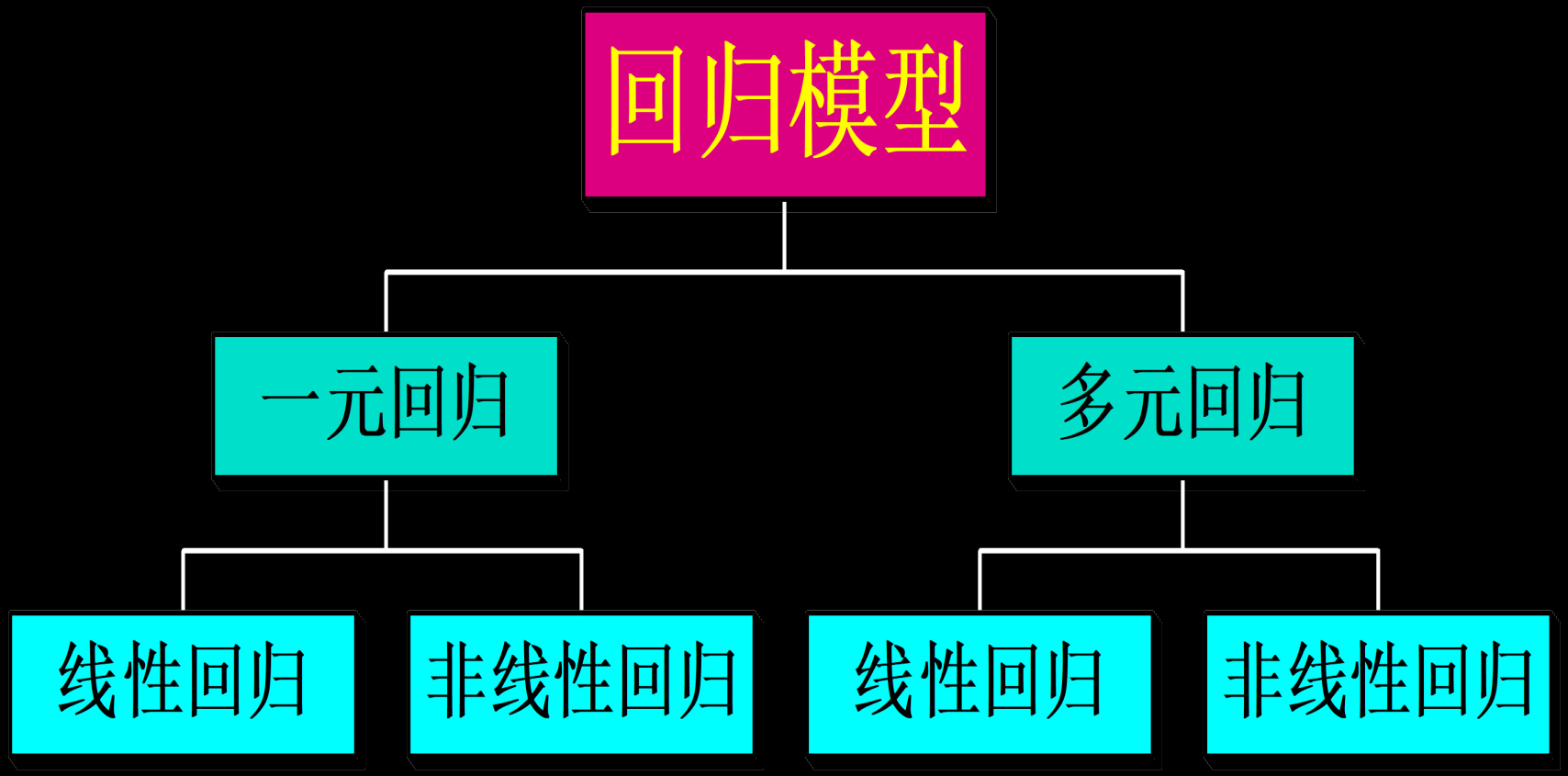
11.2.1 一元线性回归模型
11.2.2 参数的最小二乘估计
11.2.3 回归直线的拟合优度
11.2.4 显著性检验

作者：贾俊平，中国人民大学统计学院

什么是回归分析？
(Regression)

1. 从一组样本数据出发，确定变量之间的数学关系式
2. 对这些关系式的可信程度进行各种统计检验，并从影响某一特定变量的诸多变量中找出哪些变量的影响显著，哪些不显著
3. 利用所求的关系式，根据一个或几个变量的取值来预测或控制另一个特定变量的取值，并给出这种预测或控制的精确程度

回归模型的类型



一元线性回归模型

作者：贾俊平，中国人民大学统计学院

一元线性回归

1. 涉及一个自变量的回归
2. 因变量y与自变量x之间为线性关系
   - 被预测或被解释的变量称为因变量(dependent variable)，用y表示
   - 用来预测或用来解释因变量的一个或多个变量称为自变量(independent variable)，用x表示
3. 因变量与自变量之间的关系用一个线性方程来表示

回归模型
(regression model)

1. 回答“变量之间是什么样的关系？”
2. 方程中运用
   - 1 个数值型因变量(响应变量)
     - 被预测的变量
   - 1 个或多个数值型或分类型自变量 (解释变量)
     - 用于预测的变量
3. 主要用于预测和估计

一元线性回归模型

1. 描述因变量 y 如何依赖于自变量 x 和误差项 ε 的方程称为*回归模型*
2. 一元线性回归模型可表示为

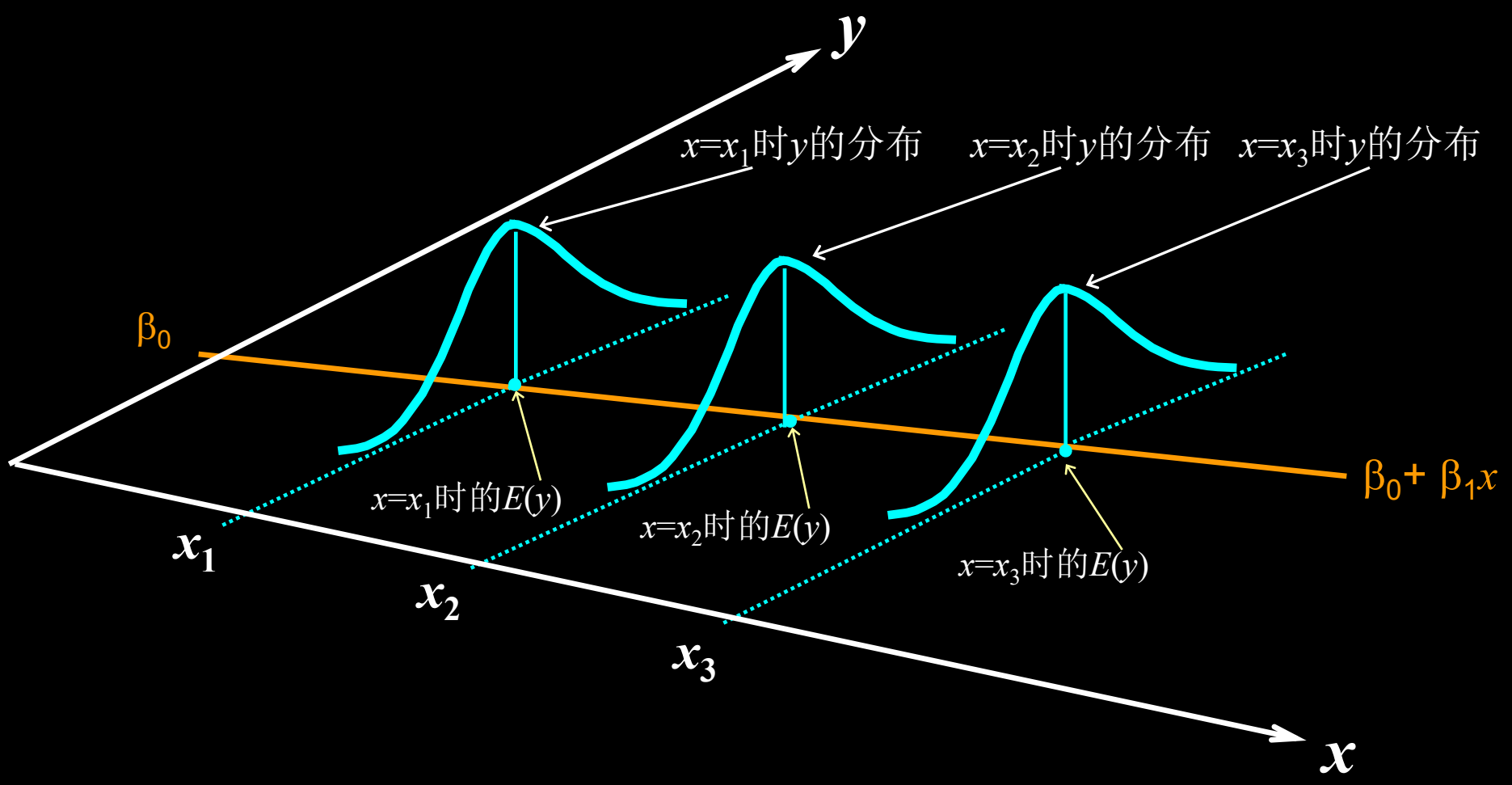
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- y 是 x 的线性函数(部分)加上误差项
- 线性部分反映了由于 x 的变化而引起的 y 的变化
- 误差项 ε 是随机变量
  - 反映了除 x 和 y 之间的线性关系之外的随机因素对 y 的影响
  - 是不能由 x 和 y 之间的线性关系所解释的变异性
- β_0 和 β_1 称为模型的参数

一元线性回归模型
(基本假定)

1. 因变量x与自变量y之间具有线性关系
2. 在重复抽样中，自变量x的取值是固定的，即假定x是非随机的
3. 误差项ε是一个期望值为0的随机变量，即$E(\varepsilon)=0$。对于一个给定的 x 值，y 的期望值为$E(y)=\beta_0+\beta_1 x$
4. 对于所有的 x 值，ε的方差σ^2 都相同
5. 误差项ε是一个服从正态分布的随机变量，且相互独立。即$\varepsilon \sim N(0, \sigma^2)$
   - 独立性意味着对于一个特定的 x 值，它所对应的ε与其他 x 值所对应的ε不相关
   - 对于一个特定的 x 值，它所对应的 y 值与其他 x 所对应的 y 值也不相关

一元线性回归模型
(基本假定)



回归方程
(regression equation)

1. 描述 y 的平均值或期望值如何依赖于 x 的方程称为回归方程
2. 一元线性回归方程的形式如下

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

- 方程的图示是一条直线，也称为直线回归方程
- β_0 是回归直线在 y 轴上的截距，是当 x=0 时 y 的期望值
- β_1 是直线的斜率，称为回归系数，表示当 x 每变动一个单位时，y 的平均变动值

估计的回归方程
(estimated regression equation)

1. 总体回归参数 β_0和 β_1是未知的，必须利用样本数据去估计
2. 用样本统计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$代替回归方程中的未知参数β_0和 β_1，就得到了*估计的回归方程*
3. 一元线性回归中估计的回归方程为

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

其中：$\hat{\beta}_0$是估计的回归直线在 y 轴上的截距，$\hat{\beta}_1$是直线的斜率，它表示对于一个给定的 x 的值，$\hat{y}$ 是 y 的估计值，也表示 x 每变动一个单位时， y 的平均变动值

参数的最小二乘估计

作者：贾俊平，中国人民大学统计学院

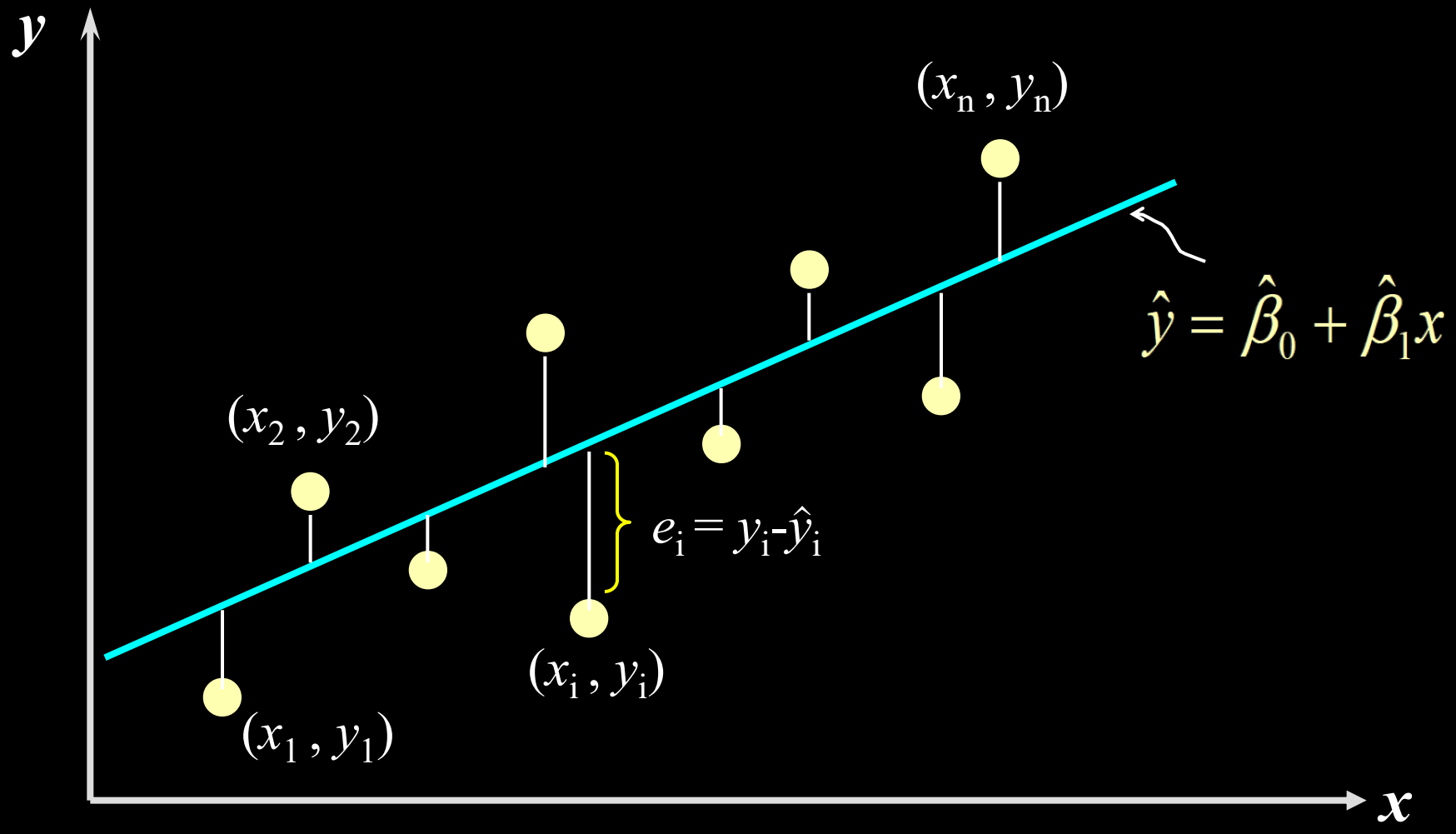
最小二乘估计
(method of least squares)

1. 德国科学家Karl Gauss(1777—1855)提出用最小化图中垂直方向的误差平方和来估计参数
2. 使因变量的观察值与估计值之间的误差平方和达到最小来求得$\hat{\beta}_0$和 $\hat{\beta}_1$的方法。即

$$\sum_{i=1}^{n}(y_i-\hat{y})^2=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2=\text{最小}$$

3. 用最小二乘法拟合的直线来代表x与y之间的关系与实际数据的误差比其他任何直线都小

Karl Gauss的最小化图



最小二乘法
($\hat{\beta}_0$和$\hat{\beta}_1$的计算公式)

- 根据最小二乘法，可得求解$\hat{\beta}_0$和 $\hat{\beta}_1$的公式如下

$$\begin{cases} \left.\dfrac{\partial Q}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2 = 0 \\ \left.\dfrac{\partial Q}{\partial \beta_1}\right|_{\beta_1=\hat{\beta}_1} = -2\sum_{i=1}^{n}x_i(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2 = 0 \end{cases}$$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n}x_i y_i - \left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

估计方程的求法
(例题分析)

【例】求不良贷款对贷款余额的回归方程

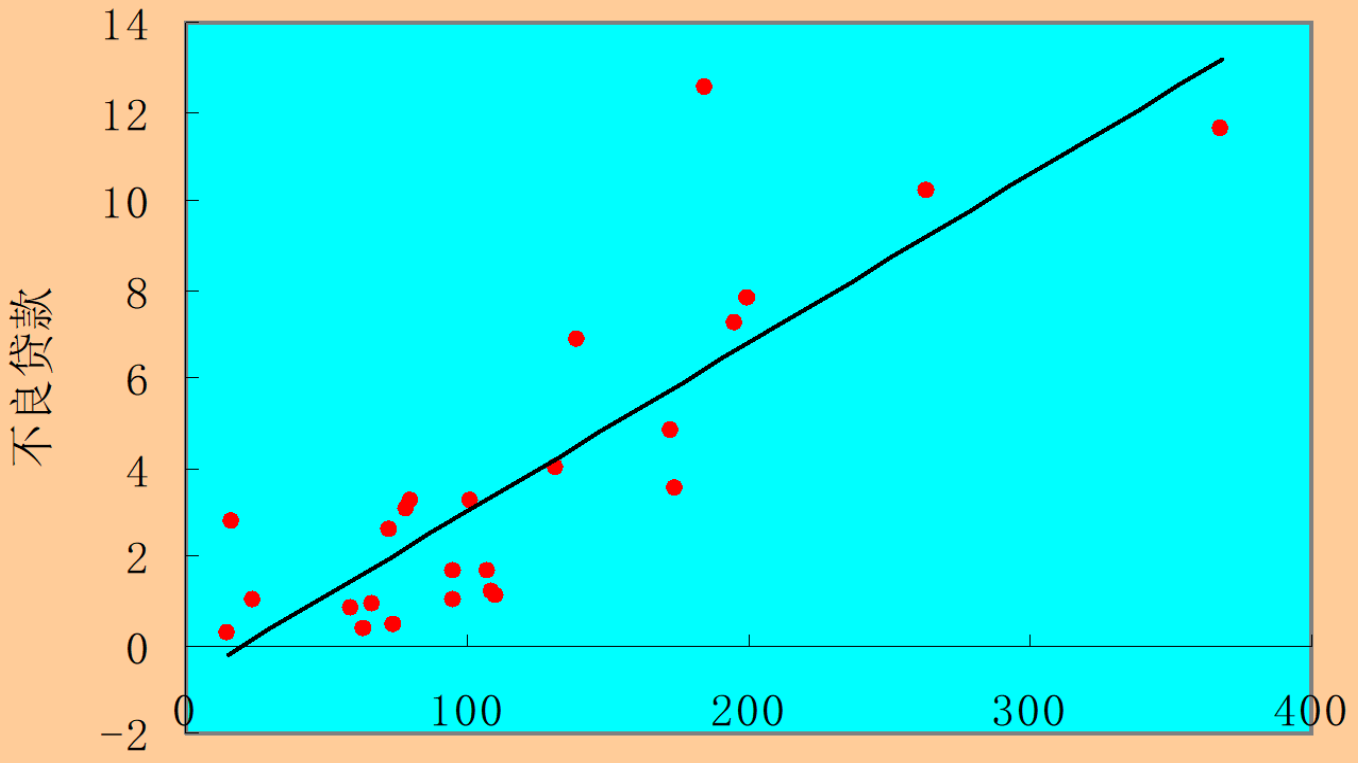
$$\begin{cases} \hat{\beta}_1 = \dfrac{25\times 17080.14 - 3006.7\times 93.2}{25\times 516543.37 - (3006.7)^2} = 0.037895 \\ \hat{\beta}_0 = 3.728 - 0.037895\times 120.268 = -0.8295 \end{cases}$$

回归方程为：$\hat{y} = -0.8295 + 0.037895\,x$

回归系数 $\hat{\beta}_1$=0.037895 表示，贷款余额每增加1亿元，不良贷款平均增加0.037895亿元

估计方程的求法
(例题分析)

不良贷款对贷款余额回归方程的图示



不良贷款对贷款余额的回归直线

用Excel进行回归分析

第1步：选择【工具】下拉菜单

第2步：选择【数据分析】选项

第3步：在分析工具中选择【回归】，选择【确定】

第4步：当对话框出现时

在【Y值输入区域】设置框内键入Y的数据区域

在【X值输入区域】设置框内键入X的数据区域

在【置信度】选项中给出所需的数值

在【输出选项】中选择输出区域

在【残差】分析选项中选择所需的选项

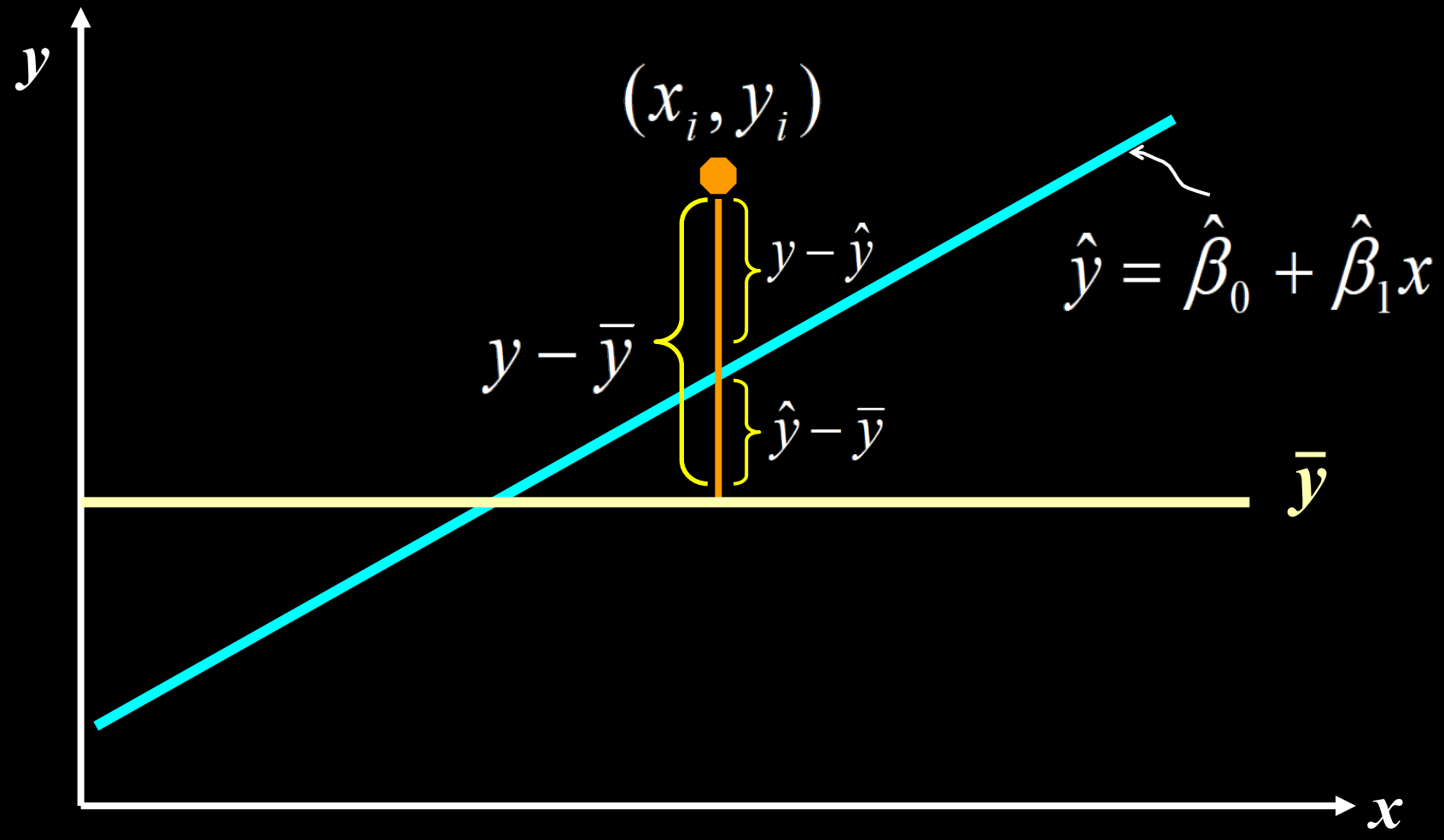
回归直线的拟合优度

作者：贾俊平，中国人民大学统计学院

变差

1. 因变量 y 的取值是不同的，y 取值的这种波动称为变差。变差来源于两个方面
   - 由于自变量 x 的取值不同造成的
   - 除 x 以外的其他因素(如x对y的非线性影响、测量误差等)的影响
2. 对一个具体的观测值来说，变差的大小可以通过该实际观测值与其均值之差 $y-\bar{y}$ 来表示

误差的分解
(图示)



误差平方和的分解
(三个平方和的关系)

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y})^2$$

总平方和 (*SST*)

回归平方和 (*SSR*)

残差平方和 (*SSE*)

$$SST = SSR + SSE$$

误差平方和的分解
(三个平方和的意义)

1. 总平方和(***SST*—total sum of squares**)
   - 反映因变量的 n 个观察值与其均值的总误差
2. 回归平方和(***SSR*—sum of squares of regression**)
   - 反映自变量 x 的变化对因变量 y 取值变化的影响，或者说，是由于 x 与 y 之间的线性关系引起的 y 的取值变化，也称为可解释的平方和
3. 残差平方和(***SSE*—sum of squares of error**)
   - 反映除 x 以外的其他因素对 y 取值的影响，也称为不可解释的平方和或剩余平方和

判定系数R^2
(coefficient of determination)

1. 回归平方和占总误差平方和的比例

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}$$

2. 反映回归直线的拟合程度
3. 取值范围在 [0 , 1] 之间
4. $R^2 \to 1$，说明回归方程拟合的越好；$R^2 \to 0$，说明回归方程拟合的越差
5. 判定系数等于相关系数的平方，即$R^2 = r^2$

判定系数
(例题分析)

【例】计算不良贷款对贷款余额回归的判定系数，并解释其意义

$$R^2 = \frac{SSR}{SST} = \frac{222.4860}{312.6504} = 0.7116 = 71.16\%$$

判定系数的实际意义是：在不良贷款取值的变差中，有71.16%可以由不良贷款与贷款余额之间的线性关系来解释，或者说，在不良贷款取值的变动中，有71.16%是由贷款余额所决定的。也就是说，不良贷款取值的差异有2/3以上是由贷款余额决定的。可见不良贷款与贷款余额之间有较强的线性关系

估计标准误差
(standard error of estimate)

1. 实际观察值与回归估计值误差平方和的均方根
2. 反映实际观察值在回归直线周围的分散状况
3. 对误差项ε的标准差σ的估计，是在排除了x对y的线性影响后，y随机波动大小的一个估计量
4. 反映用估计的回归方程预测y时预测误差的大小
5. 计算公式为

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$$

注：例题的计算结果为**1.9799**

显著性检验

作者：贾俊平，中国人民大学统计学院

线性关系的检验

1. 检验自变量与因变量之间的线性关系是否显著
2. 将回归均方(*MSR*)同残差均方(*MSE*)加以比较，应用*F*检验来分析二者之间的差别是否显著
   - 回归均方：回归平方和*SSR*除以相应的自由度(自变量的个数*k*)
   - 残差均方：残差平方和*SSE*除以相应的自由度(*n*-*k*-1)

线性关系的检验
(检验的步骤)

1. 提出假设
   - H_0：$\beta_1=0$　线性关系不显著
2. 计算检验统计量F

$$F=\frac{SSR/1}{SSE/(n-2)}=\frac{MSR}{MSE}\sim F(1,n-2)$$

3. 确定显著性水平α，并根据分子自由度1和分母自由度n-2找出临界值F_α
4. 作出决策：若$F>F_\alpha$，拒绝H_0；若$F<F_\alpha$，不拒绝H_0

线性关系的检验
(例题分析)

1. 提出假设
   - H_0：$\beta_1=0$　不良贷款与贷款余额之间的线性关系不显著
2. 计算检验统计量F

$$F=\frac{SSR/1}{SSE/(n-2)}=\frac{222.48598/1}{90.164421/(25-2)}=56.753844$$

3. 确定显著性水平α=0.05，并根据分子自由度1和分母自由度25-2找出临界值F_α=4.28
4. 作出决策：若$F>F_\alpha$，拒绝H_0，线性关系显著

线性关系的检验
(方差分析表)

Excel 输出的方差分析表

	A	B	C	D	E	F
1	方差分析					
2		df	SS	MS	F	Significance F
3	回归分析	1	110252.7	110252.7	56.75384	1.18349E-07
4	残差	23	44680.88	1942.647		
5	总计	24	154933.6			

回归系数的检验

1. 检验 x 与 y 之间是否具有线性关系，或者说，检验自变量 x 对因变量 y 的影响是否显著
2. 理论基础是回归系数 $\hat{\beta}_1$的抽样分布
3. 在一元线性回归中，等价于线性关系的显著性检验
4. 采用*t*检验

回归系数的检验
(检验步骤)

1. 提出假设
   - H_0: $\beta_1 = 0$ (没有线性关系)
   - H_1: $\beta_1 \neq 0$ (有线性关系)
2. 计算检验的统计量

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \sim t(n-2)$$

3. 确定显著性水平α，并进行决策
   - $|t| > t_{\alpha/2}$，拒绝H_0；$|t| < t_{\alpha/2}$，不拒绝H_0

回归系数的检验 (例题分析)

- 对例题的回归系数进行显著性检验(α=0.05)

1. 提出假设
   - H_0：$\beta_1 = 0$
   - H_1：$\beta_1 \neq 0$
2. 计算检验的统计量

$$t = \frac{0.037895}{0.005030} = 7.533515$$

3. t=7.533515>$t_{\alpha/2}$=2.201，拒绝H_0，表明不良贷款与贷款余额之间有显著的线性关系

回归系数的检验
(例题分析)

☞ *P* 值的应用

	A	B	C	D	E
1		Coefficients	标准误差	t Stat	P-value
2	Intercept	-0.829521	0.723043	-1.147263	0.263068
3	X Variable	0.037895	0.005030	7.533515	0.000000

P=0.000000<α=0.05，拒绝原假设，不良贷款与贷款余额之间有显著的线性关系

回归分析结果的评价

- 建立的模型是否合适？或者说，这个拟合的模型有多“好”？要回答这些问题，可以从以下几个方面入手

1. 所估计的回归系数 $\hat{\beta}_1$ 的符号是否与理论或事先预期相一致
   - 在不良贷款与贷款余额的回归中，可以预期贷款余额越多，不良贷款也可能会越多，也就是说，回归系数的值应该是正的，在上面建立的回归方程中，我们得到的回归系数为正值，$\hat{\beta}_1 = 0.037895$
2. 如果理论上认为x与y之间的关系不仅是正的，而且是统计上显著的，那么所建立的回归方程也应该如此
   - 在不良贷款与贷款余额的回归中，二者之间为正的线性关系，而且，对回归系数的t检验结果表明而这之间的线性关系是统计上显著的

回归分析结果的评价

3. 回归模型在多大程度上解释了因变量y取值的差异？可以用判定系数R^2来回答这一问题
   - 在不良贷款与贷款余额的回归中，得到的R^2=71.16%，解释了不良贷款变差的2/3以上，说明拟合的效果还算不错
4. 考察关于误差项ε的正态性假定是否成立。因为我们在对线性关系进行F检验和回归系数进行t检验时，都要求误差项ε服从正态分布，否则，我们所用的检验程序将是无效的。ε正态性的简单方法是画出残差的直方图或正态概率图

Excel输出的部分回归结果

名称	计算公式
Adjusted R Square	$R_a^2 = 1-(1-R^2)\times\frac{n-1}{n-k-1}$
Intercept的抽样标准误差	$s_{\hat{\beta}_0} = s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$
Intercept95%的置信区间	$\hat{\beta}_0 \pm t_{\alpha/2}(n-2)s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$
斜率95%的置信区间	$\hat{\beta}_1 \pm t_{\alpha/2}(n-2)\frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

11.3 利用回归方程进行估计和预测

11.3.1 点估计

11.3.2 区间估计

作者：贾俊平，中国人民大学统计学院

利用回归方程进行估计和预测

1. 根据自变量 x 的取值估计或预测因变量 y 的取值
2. 估计或预测的类型
   - 点估计
     - y 的平均值的点估计
     - y 的个别值的点估计
   - 区间估计
     - y 的平均值的*置信区间*估计
     - y 的个别值的*预测区间*估计

点估计

作者：贾俊平，中国人民大学统计学院

点估计

1. 对于自变量 x 的一个给定值x_0，根据回归方程得到因变量 y 的一个估计值 $\hat{y}_0$
2. 点估计值有
   - y 的*平均值*的点估计
   - y 的*个别值*的点估计
3. 在点估计条件下，平均值的点估计和个别值的的点估计是一样的，但在区间估计中则不同

y 的平均值的点估计

- 利用估计的回归方程，对于自变量 x 的一个给定值 x_0，求出因变量 y 的平均值的一个估计值$E(y_0)$，就是平均值的点估计
  - 在前面的例子中，假如我们要估计贷款余额为100亿元时，所有分行不良贷款的平均值，就是平均值的点估计 。根据估计的回归方程得

$$E(y_0) = -0.8295 + 0.037895 \times 100 = 2.96(\text{亿元})$$

y 的个别值的点估计

- 利用估计的回归方程，对于自变量 x 的一个给定值 x_0，求出因变量 y 的一个个别值的估计值 $\hat{y}_0$，就是个别值的点估计
  - 例如，如果我们只是想知道贷款余额为72.8亿元的那个分行(这里是编号为10的那个分行)的不良贷款是多少，则属于个别值的点估计。根据估计的回归方程得

$$\hat{y}_0 = -0.8295 + 0.037895 \times 72.8 = 1.93(\text{亿元})$$

区间估计

作者：贾俊平，中国人民大学统计学院

区间估计

1. 点估计不能给出估计的精度，点估计值与实际值之间是有误差的，因此需要进行区间估计
2. 对于自变量 x 的一个给定值 x_0，根据回归方程得到因变量 y 的一个估计区间
3. 区间估计有两种类型
   - 置信区间估计(confidence interval estimate)
   - 预测区间估计(prediction interval estimate)

置信区间估计

1. 利用估计的回归方程，对于自变量 x 的一个给定值 x_0，求出因变量 y 的平均值的估计区间，这一估计区间称为*置信区间*(confidence interval)
2. $E(y_0)$ 在1-α置信水平下的置信区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)s_e\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

式中：s_e为估计标准误差

置信区间估计
(例题分析)

【例】求出贷款余额为100亿元时，不良贷款95%置信水平下的置信区间

解：根据前面的计算结果，已知n=25，$\hat{y}_0 = 2.96$

s_e=1.9799，$t_{\alpha/2}(25-2)=2.069$

置信区间为

$$2.96 \pm 2.069 \times 1.9799 \times \sqrt{\frac{1}{25} + \frac{(100-120.268)^2}{154933.5744}}$$

$$2.1141 \le E(y_0) \le 3.8059$$

当贷款余额为100亿元时，不良贷款的平均值在2.1141亿元到3.8059亿元之间

预测区间估计

1. 利用估计的回归方程，对于自变量 x 的一个给定值 x_0，求出因变量 y 的一个个别值的估计区间，这一区间称为*预测区间*(prediction interval)
2. y_0在1-α置信水平下的预测区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)S_e \sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

注意！#1

预测区间估计
(例题分析)

【例】求出贷款余额为72.8亿元的那个分行，不良贷款95%的预测区间

解：根据前面的计算结果，已知n=25，$\hat{y}_0 = 1.93$

s_e=1.9799，$t_{\alpha/2}(25-2)=2.069$

预测区间为

$$2.96 \pm 2.0687 \times 1.9799 \times \sqrt{1+\frac{1}{25}+\frac{(72.8-120.268)^2}{154933.5744}}$$

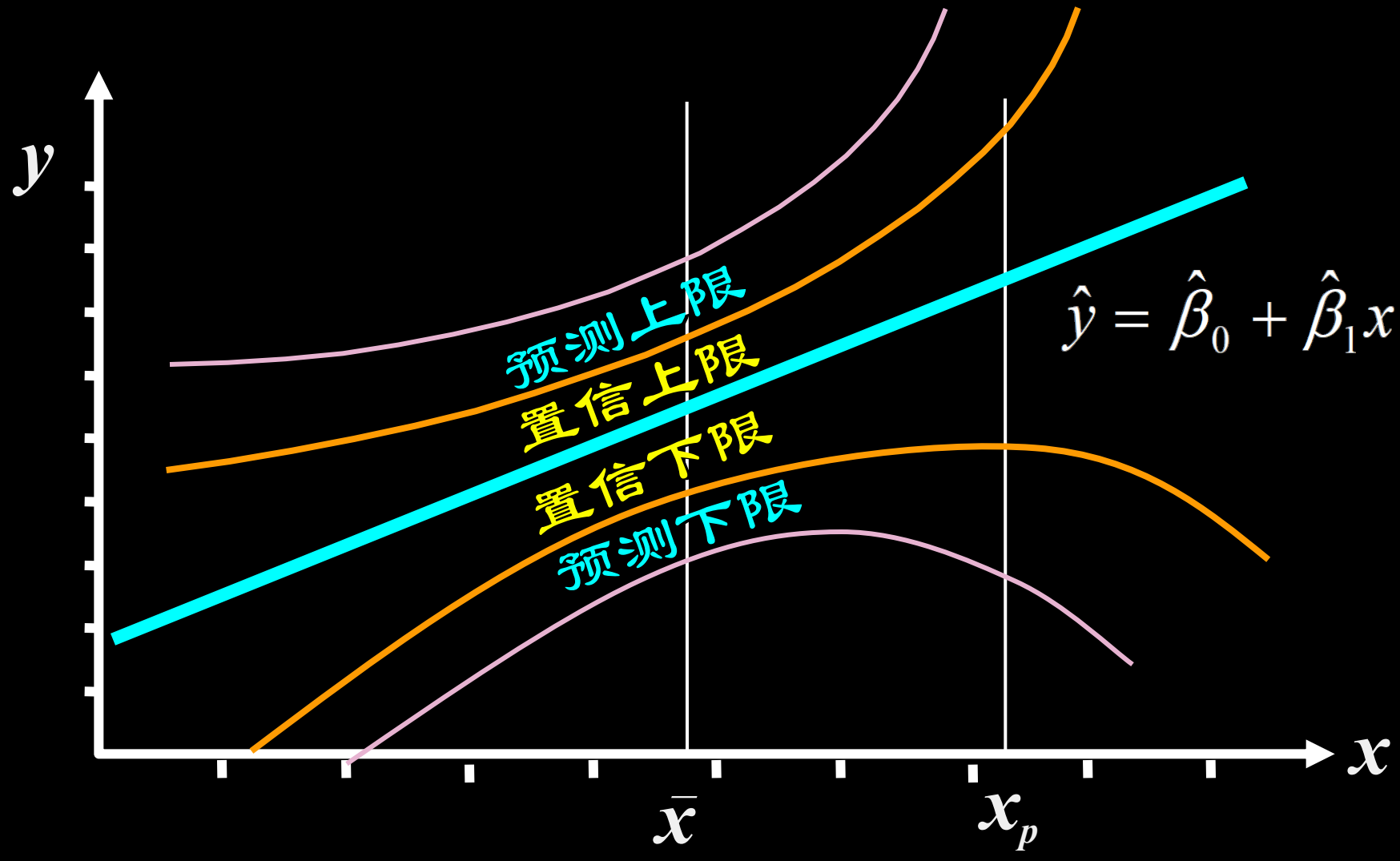
$$-2.2766 \le \hat{y}_0 \le 6.1366$$

贷款余额为72.8亿元的那个分行，其不良贷款的预测区间在-2.2766亿元到6.1366亿元之间

置信区间和预测区间
(例题分析)

	A	B	C	D	E	F	G	H
1	分行	不良贷款	贷款余额	预测 Y	置信区间		预测区间	
2	编号	(y)	(x)		置信下限	置信上限	预测下限	预测上限
3	1	0.9	67.3	1.7208	0.7333	2.7083	-2.4964	5.9380
4	2	1.1	111.3	3.3882	2.5636	4.2128	-0.7939	7.5702
5	3	4.8	173	5.7263	4.7401	6.7124	1.5094	9.9431
6	4	3.2	80.8	2.2324	1.3159	3.1489	-1.9687	6.4335
7	5	7.8	199.7	6.7381	5.5742	7.9019	2.4761	11.0000
8	6	2.7	16.2	-0.2156	-1.5737	1.1424	-4.5346	4.1034
9	7	1.6	107.4	3.2404	2.4102	4.0705	-0.9428	7.4235
10	8	12.5	185.4	6.1962	5.1328	7.2595	1.9606	10.4317
11	9	1.0	96.1	2.8122	1.9551	3.6692	-1.3764	7.0007
12	10	2.6	72.8	1.9292	0.9725	2.8859	-2.2809	6.1393
13	11	0.3	64.2	1.6033	0.5975	2.6092	-2.6182	5.8248
14	12	4.0	132.2	4.1802	3.3515	5.0088	-0.0027	8.3630
15	13	0.8	58.6	1.3911	0.3504	2.4319	-2.8388	5.6211
16	14	3.5	174.6	5.7869	4.7914	6.7824	1.5678	10.0059
17	15	10.2	263.5	9.1557	7.4547	10.8567	4.7170	13.5945
18	16	3.0	79.3	2.1755	1.2519	3.0991	-2.0271	6.3782
19	17	0.2	14.8	-0.2687	-1.6384	1.1010	-4.5913	4.0540
20	18	0.4	73.5	1.9557	1.0028	2.9087	-2.2535	6.1650
21	19	1.0	24.7	0.1065	-1.1821	1.3951	-4.1912	4.4041
22	20	6.8	139.4	4.4530	3.6099	5.2962	0.2673	8.6387
23	21	11.6	368.2	13.1233	10.4160	15.8306	8.2102	18.0364
24	22	1.6	95.7	2.7970	1.9387	3.6553	-1.3918	6.9858
25	23	1.2	109.6	3.3237	2.4969	4.1505	-0.8587	7.5062
26	24	7.2	196.2	6.6054	5.4671	7.7437	2.3504	10.8604
27	25	3.2	102.2	3.0433	2.2027	3.8839	-1.1419	7.2285

置信区间、预测区间、回归方程



11.4 残差分析

11.4.1 残差与残差图

11.4.2 标准化

作者：贾俊平，中国人民大学统计学院

残差与残差图

作者：贾俊平，中国人民大学统计学院

残差
(residual)

1. 因变量的观测值与根据估计的回归方程求出的预测值之差，用e表示

$$e_i = y_i - \hat{y}_i$$

2. 反映了用估计的回归方程去预测而引起的误差
3. 可用于确定有关误差项ε的假定是否成立
4. 用于检测有影响的观测值

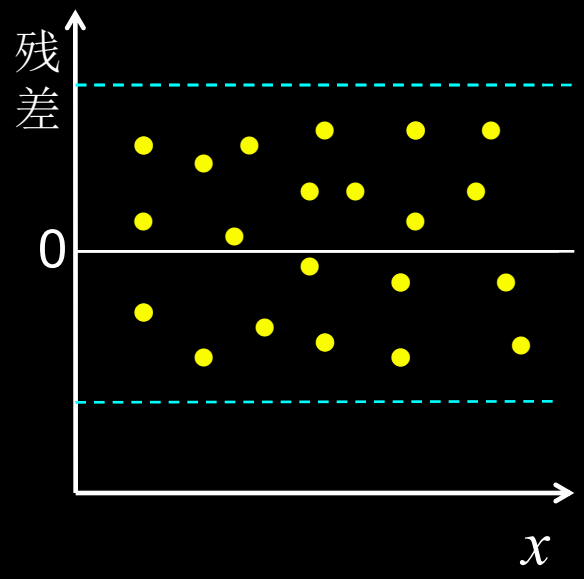
残差图
(residual plot)

1. 表示残差的图形
   - 关于x的残差图
   - 关于y的残差图
   - 标准化残差图
2. 用于判断误差ε的假定是否成立
3. 检测有影响的观测值

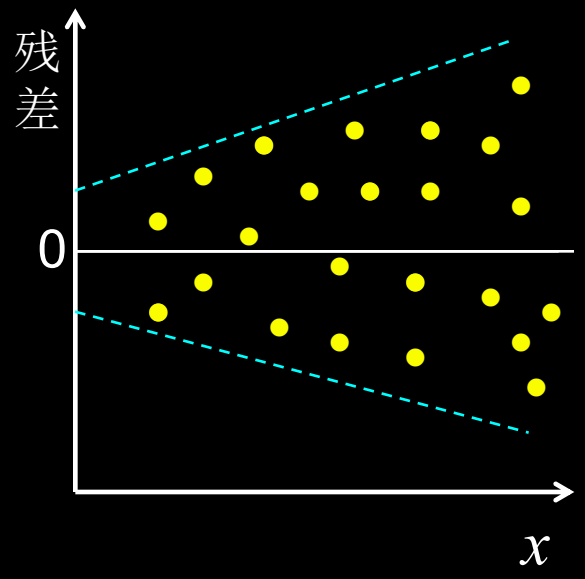
残差与标准化残差图
(例题分析)

	A	B	C	D	E	F	G
1	分行编号	不良贷款 y	贷款余额 x	预测 $\hat{y}_i$	残差 e_i	标准残差 z_{e_i}	杠杆率 h_i
2	1	0.9	67.3	1.7208	-0.8208	-0.4146	0.0581
3	2	1.1	111.3	3.3882	-2.2882	-1.1557	0.0405
4	3	4.8	173.0	5.7263	-0.9263	-0.4678	0.0579
5	4	3.2	80.8	2.2324	0.9676	0.4887	0.0501
6	5	7.8	199.7	6.7381	1.0619	0.5364	0.0807
7	6	2.7	16.2	-0.2156	2.9156	1.4726	0.1099
8	7	1.6	107.4	3.2404	-1.6404	-0.8285	0.0411
9	8	12.5	185.4	6.1962	6.3038	3.1838	0.0674
10	9	1.0	96.1	2.8122	-1.8122	-0.9153	0.0438
11	10	2.6	72.8	1.9292	0.6708	0.3388	0.0545
12	11	0.3	64.2	1.6033	-1.3033	-0.6583	0.0603
13	12	4.0	132.2	4.1802	-0.1802	-0.0910	0.0409
14	13	0.8	58.6	1.3911	-0.5911	-0.2985	0.0645
15	14	3.5	174.6	5.7869	-2.2869	-1.1550	0.0591
16	15	10.2	263.5	9.1557	1.0443	0.5274	0.1724
17	16	3.0	79.3	2.1755	0.8245	0.4164	0.0508
18	17	0.2	14.8	-0.2687	0.4687	0.2367	0.1118
19	18	0.4	73.5	1.9557	-1.5557	-0.7857	0.0541
20	19	1.0	24.7	0.1065	0.8935	0.4513	0.0989
21	20	6.8	139.4	4.4530	2.3470	1.1854	0.0424
22	21	11.6	368.2	13.1233	-1.5233	-0.7694	0.4368
23	22	1.6	95.7	2.7970	-1.1970	-0.6046	0.0439
24	23	1.2	109.6	3.3237	-2.1237	-1.0726	0.0407
25	24	7.2	196.2	6.6054	0.5946	0.3003	0.0772
26	25	3.2	102.2	3.0433	0.1567	0.0791	0.0421

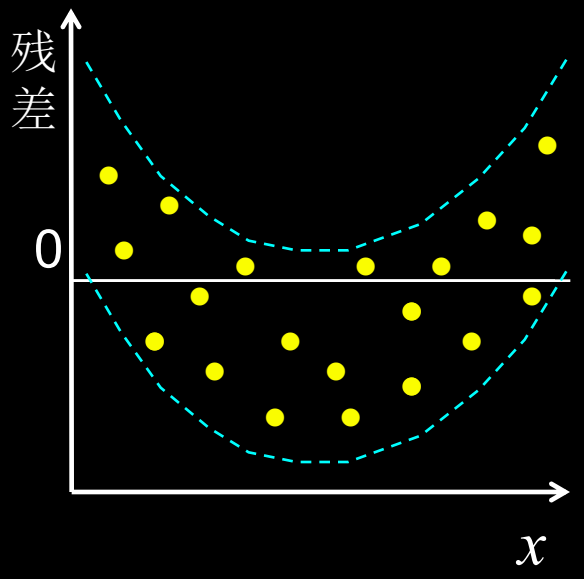
残差图
(形态及判别)



(a)满意模式

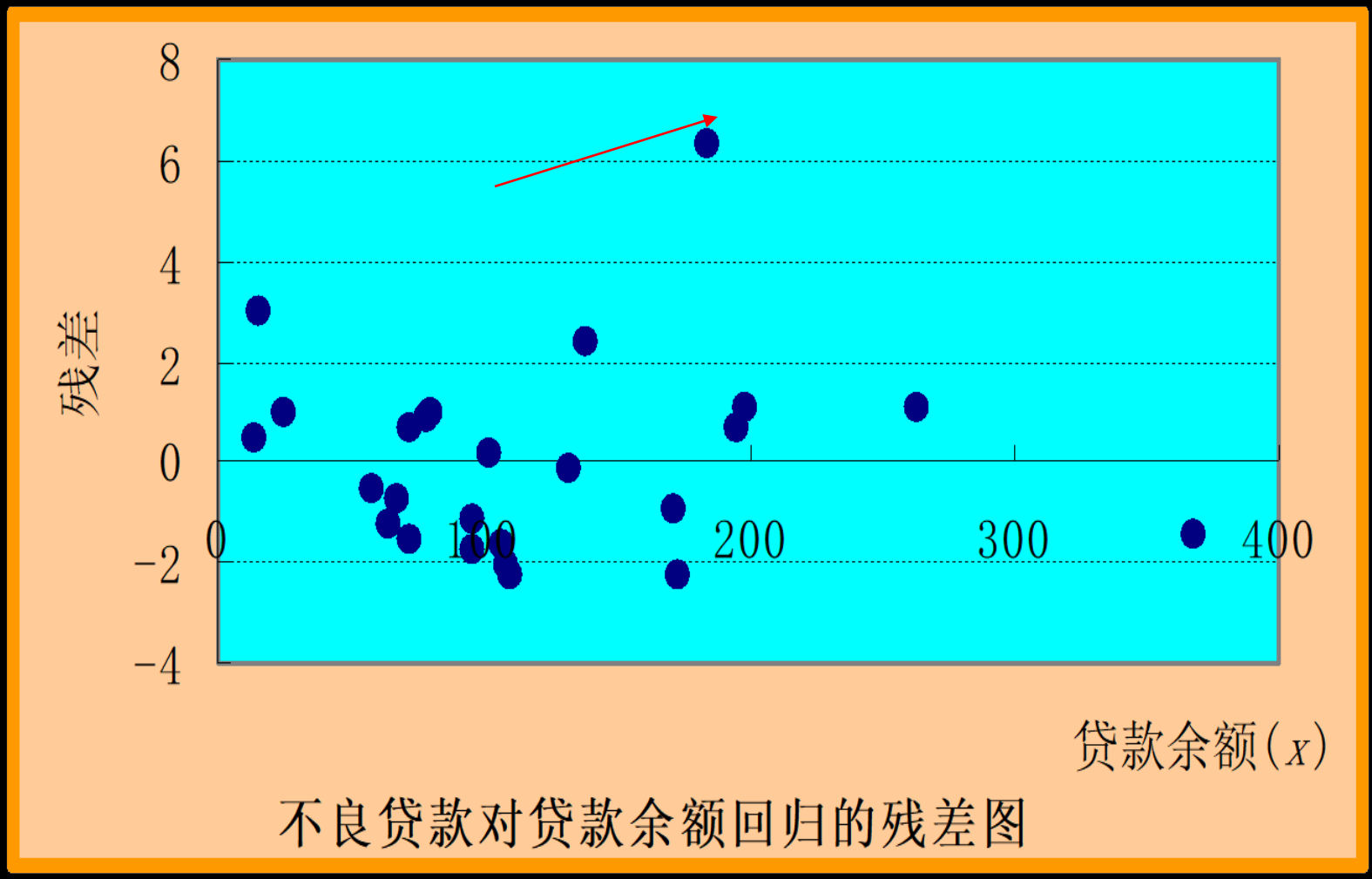


(b)非常数方差



(c)模型不合适

残差图
(例题分析)



不良贷款对贷款余额回归的残差图

标准化残差

作者：贾俊平，中国人民大学统计学院

标准化残差 (standardized residual)

1. 残差除以它的标准差
2. 也称为Pearson残差或半学生化残差(semi-studentized residuals)
3. 计算公式为 $z_{e_i} = \frac{e_i}{s_e} = \frac{y_i - \hat{y}_i}{s_e}$

注意：Excel给出的标准残差的计算公式为

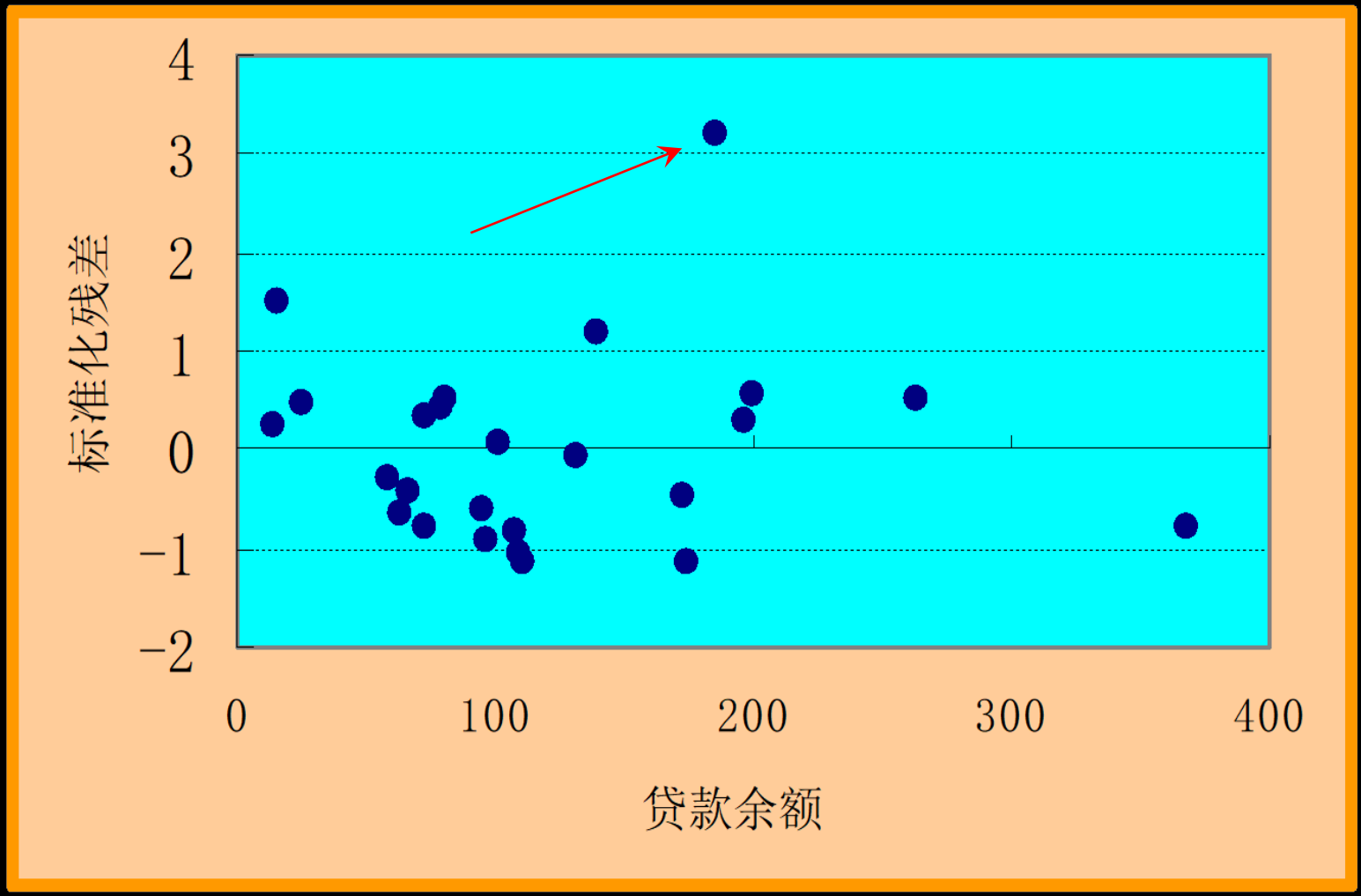
$$z_{e_i} = \frac{y_i - \hat{y}_i}{s_e \sqrt{1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}}$$

这实际上是学生化删除残差(studentized deleted residuals)

标准化残差图

- 用以直观地判断误差项服从正态分布这一假定是否成立
  - 若假定成立，标准化残差的分布也应服从正态分布
  - 在标准化残差图中，大约有95%的标准化残差在-2到+2之间

标准化残差图
(例题分析)



本章小结

1. 变量间关系的度量
2. 回归模型、回归方程与估计的回归方程
3. 回归直线的拟合优度
4. 回归分析中的显著性检验
5. 估计和预测
6. 用Excel 进行回归分析

结　束

THANKS