第 8 章　一元线性回归

内容提要

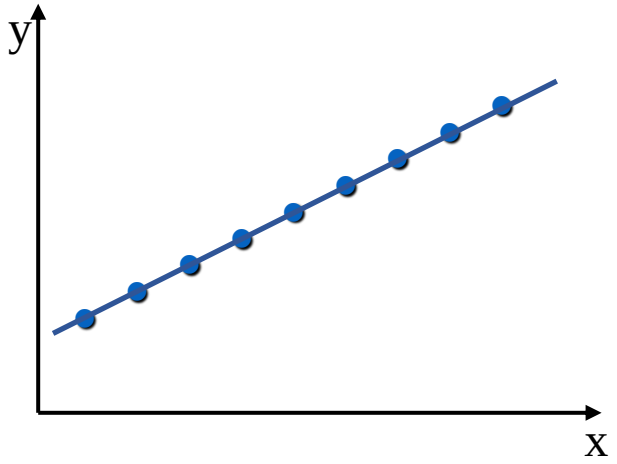
8.1 变量之间关系的度量

- 变量之间的关系
- 相关关系的描述与测度
- 相关系数的显著性检验

变量之间的关系：函数关系

定义

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依据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。



性质

函数关系是一一对应的**确定关系**

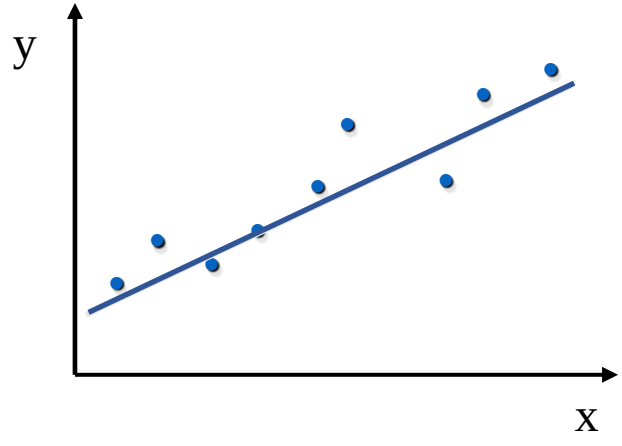
图示

各个观测点落在一条直线上。

例子

- 商品销售额y与销售量x的关系：y=px
- 圆的面积S与半径R的关系：S= πR^2

变量之间的关系：相关关系



描述

一个变量的取值y不能由另一个变量x唯一确定，即当变量x取某个值时，变量y的取值可能有多个。

性质

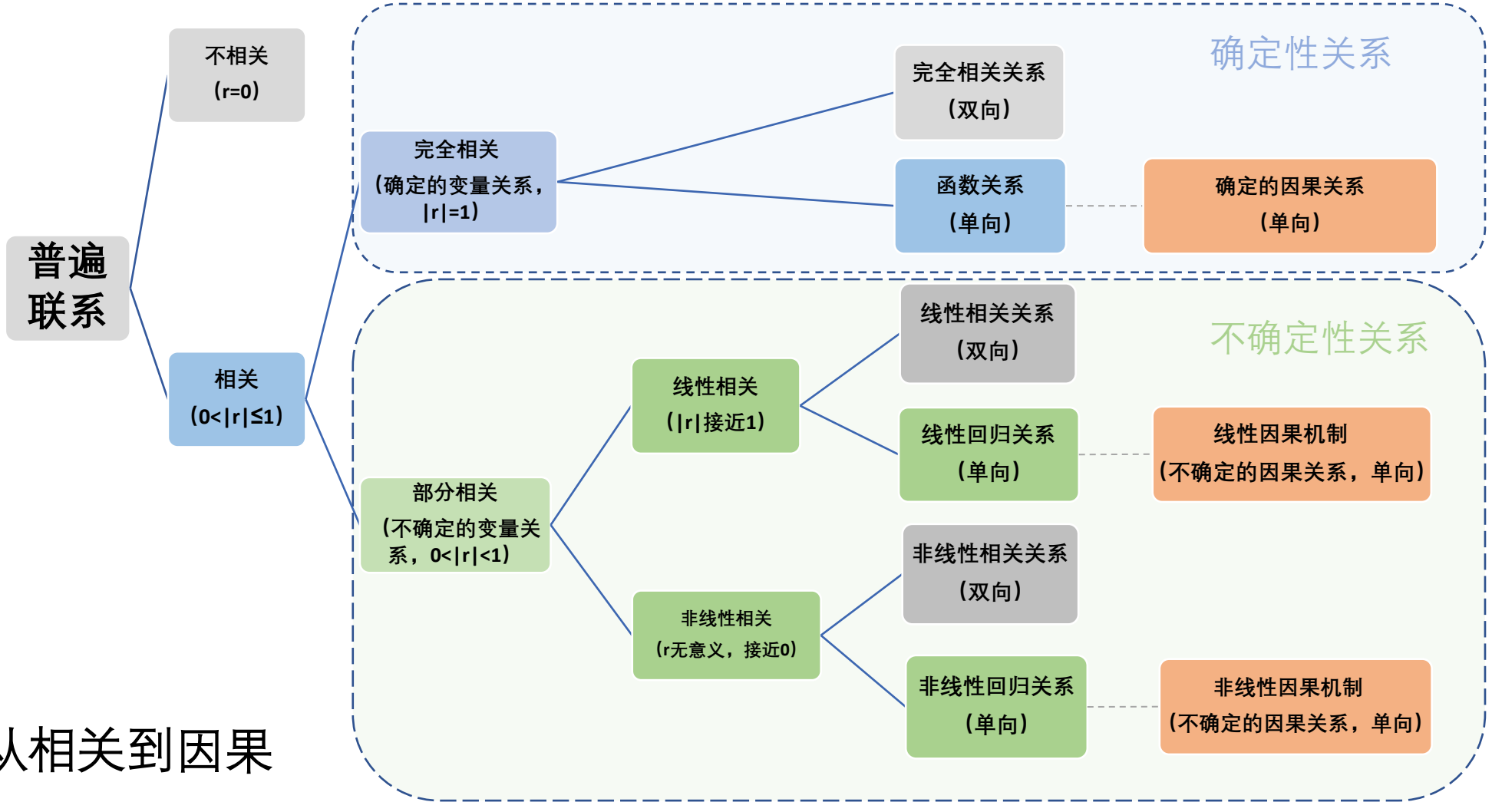
相关关系是**不确定的数量关系**

图示

各个观测点分布在直线周围。

例子

- 子女身高y与父母身高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
- 商品消费y与居民收入x之间的关系
- 商品销售额y与广告支出x之间的关系


普遍联系
不相关
(r=0)
相关
(0<|r|≤1)
确定性关系
完全相关
(确定的变量关系，|r|=1)
完全相关关系
(双向)
函数关系
(单向)
确定的因果关系
(单向)
不确定性关系
部分相关
(不确定的变量关系，0<|r|<1)
线性相关
(|r|接近1)
线性相关关系
(双向)
线性回归关系
(单向)
线性因果机制
(不确定的因果关系，单向)
非线性相关
(r无意义，接近0)
非线性相关关系
(双向)
非线性回归关系
(单向)
非线性因果机制
(不确定的因果关系，单向)

从相关到因果

相关分析的主要问题与假定

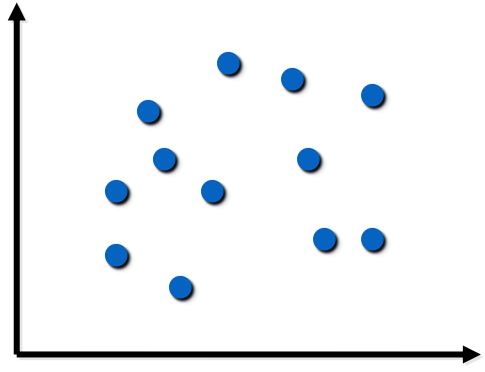
主要问题

- 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？ **线性vs非线性**
- 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？**显著性**

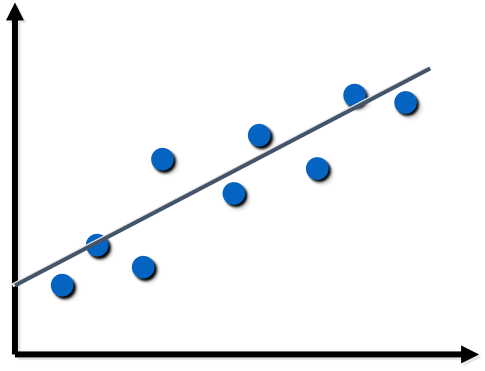
假定

- 两个变量之间是**线性关系**
- 两个变量都是**随机变量**

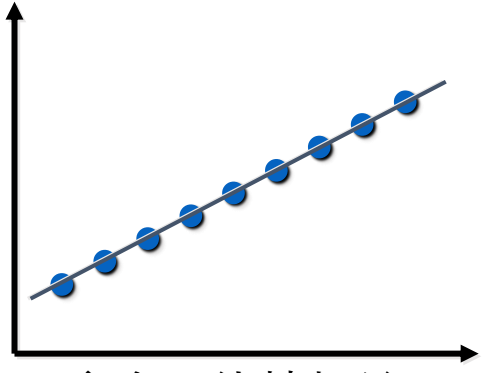
相关关系的描述与测度：散点图



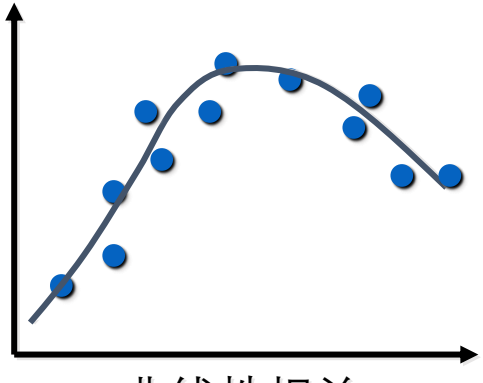
不相关



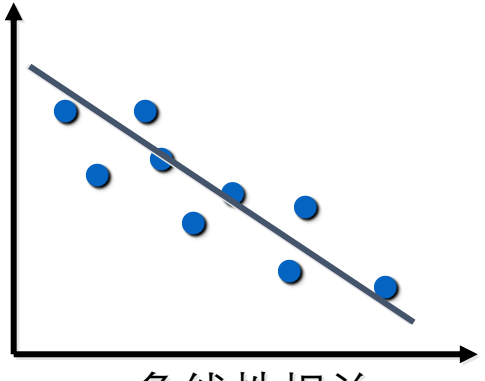
正线性相关



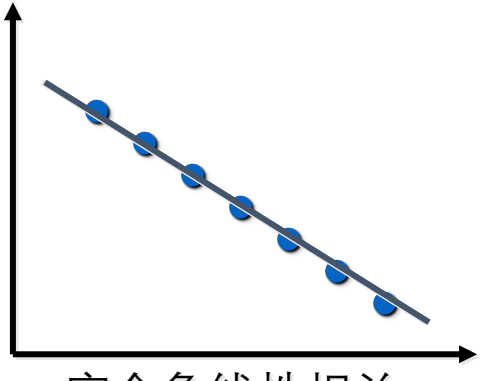
完全正线性相关



非线性相关



负线性相关



完全负线性相关

相关关系的描述与测度：相关系数(一)

定义

相关系数(correlation coefficient)，是根据样本数据计算的，度量两个变量之间**线性关系强度**的**统计量**。

分类

- 总体相关系数：ρ
- 样本相关系数：$\mathbf{r}$(Pearson相关系数)

公式

- 总体相关系数

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

- **样本相关系数**

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

相关关系的描述与测度：相关系数(二)

性质

- r的**取值范围**$[-1,1]$。
- r具有**对称性**。即$r_{xy}=r_{yx}$
- r的数值大小**与原点和尺度无关**。
- r是x与y之间线性关系的度量，**不能用于描述非线性关系**。
- r是x与y之间线性关系的度量，**不一定意味着因果关系**。

r 的取值区间、含义与关系类型

取值	含义	关系类型
$r=-1$	完全负相关	确定性关系
$-1<r\le-0.8$	高度负相关	不确定性关系
$-0.8<r\le-0.5$	中度负相关	
$-0.5<r\le-0.3$	低度负相关	
$-0.3<r<0$	弱负相关(视为不相关)	
$r=0$	不相关	确定性关系
$0<r<0.3$	弱正相关(视为不相关)	不确定性关系
$0.3\le r<0.5$	低度正相关	
$0.5\le r<0.8$	中度正相关	
$0.8\le r<1$	高度正相关	
$r=1$	完全正相关	确定性关系

相关系数的显著性检验

总体相关系数与样本相关系数

- 总体相关系数：ρ（未知）
- 样本相关系数：r(依据样本计算)
- r是ρ的近似估计，受抽样波动的影响

r的抽样分布

- 当$\rho \to -1$时，r呈现出**右偏分布**
- 当$\rho \to 1$时，r呈现出**左偏分布**
- 当$\rho \to 0$，且$n \to +\infty$时,r接近**正态分布**

r的显著性检验

- 检验方式：t检验
- 检验内容：是否存在线性相关关系

r的显著性检验步骤

- 提出假设：

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

- 计算检验统计量

$$t = |r|\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2)$$

- 确定显著性水平α，做出统计决策

若$|t| > t_{\alpha/2}$，拒绝H_0

若$|t| < t_{\alpha/2}$，不能拒绝H_0

例子：背景与数据

例子

某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。

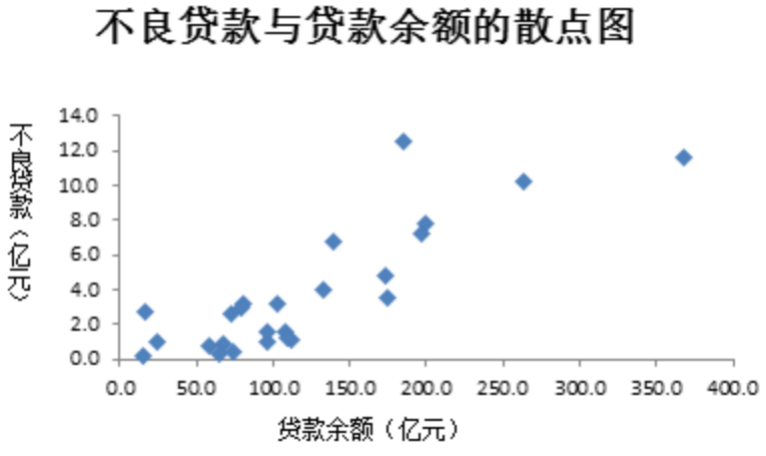
近年来，该银行的贷款额平稳增长，但**不良贷款**额也有较大比例的增长，这给银行业务的发展带来较大压力。

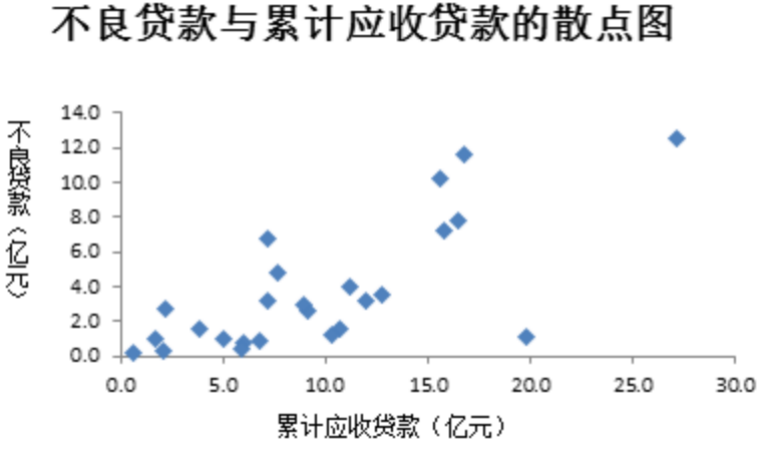
管理者希望知道：

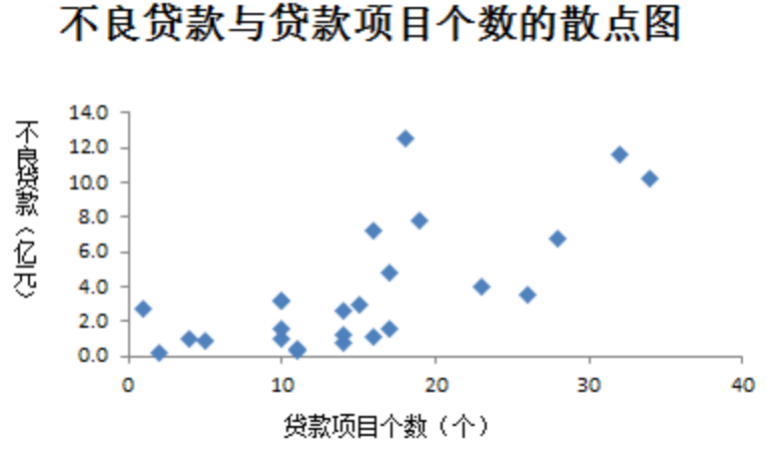
- **不良贷款**是否与**贷款余额、累计应收贷款、贷款项目个数、固定资产投资额**等因素有关？
- 如果有关，它们之间是**什么样的关系**？
- **关系强度**如何？

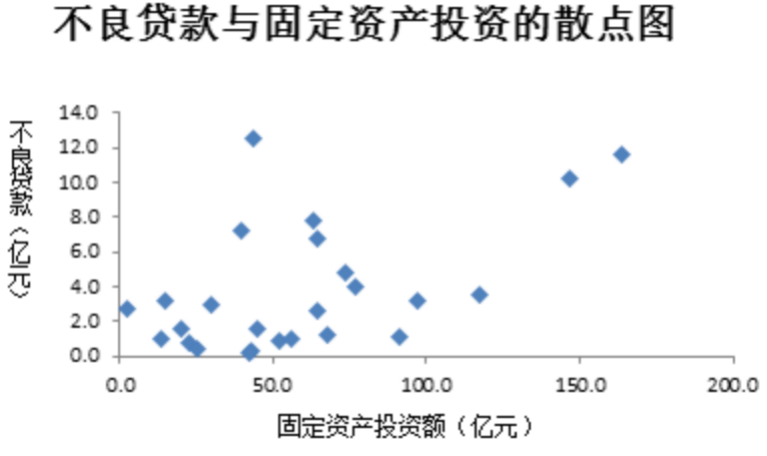
	A	B	C	D	E	F
1	分行编号	不良贷款（亿元）	各项贷款余额（亿元）	本年累计应收贷款（亿元）	贷款项目个数（个）	本年固定资产投资额（亿元）
2	1	0.9	67.3	6.8	5	51.9
3	2	1.1	111.3	19.8	16	90.9
4	3	4.8	173.0	7.7	17	73.7
5	4	3.2	80.8	7.2	10	14.5
6	5	7.8	199.7	16.5	19	63.2
7	6	2.7	16.2	2.2	1	2.2
8	7	1.6	107.4	10.7	17	20.2
9	8	12.5	185.4	27.1	18	43.8
10	9	1.0	96.1	1.7	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

例子：散点图


不良贷款与贷款余额的散点图
不良贷款（亿元）
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
0.0
50.0
100.0
150.0
200.0
250.0
300.0
350.0
400.0
贷款余额（亿元）


不良贷款与累计应收贷款的散点图
不良贷款（亿元）
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
0.0
5.0
10.0
15.0
20.0
25.0
30.0
累计应收贷款（亿元）


不良贷款与贷款项目个数的散点图
不良贷款（亿元）
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
0
10
20
30
40
贷款项目个数（个）


不良贷款与固定资产投资的散点图
不良贷款（亿元）
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
0.0
50.0
100.0
150.0
200.0
固定资产投资额（亿元）

例子：相关系数

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

例子：相关系数的显著性检验

检验：不良贷款与贷款余额之间的相关系数显著性（$\alpha = 0.05$）

步骤

- 提出假设：$H_0: \rho = 0$；$H_1: \rho \neq 0$
- 计算检验统计量

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

- 根据显著性水平$\alpha = 0.05$，查表得到临界值：$t_{\alpha/2}(n-2) = 2.069$。由于$t = 7.5344 > 2.069$，故拒绝$H_0$，即不良贷款与贷款余额之间存在着显著的正相关关系。

	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

8.2 一元线性回归

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

什么是回归分析

回归分析解决的问题

- **确定变量关系**

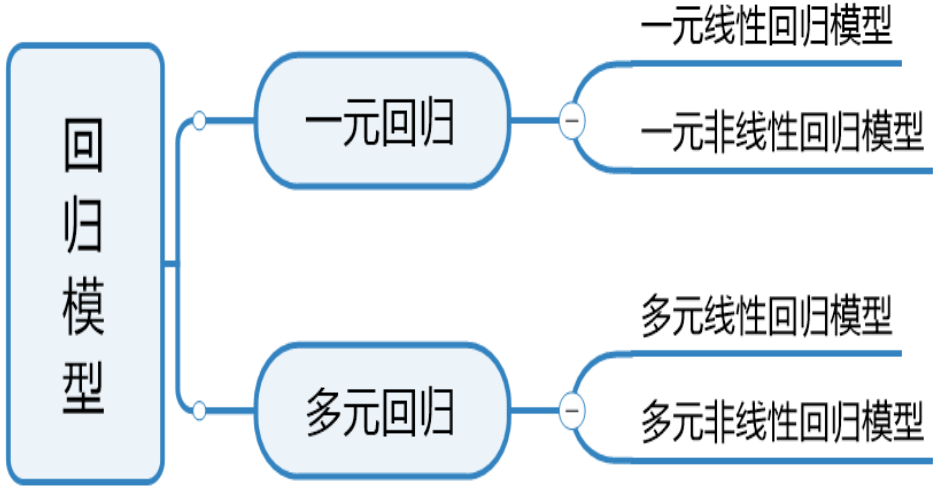
  从一组样本数据出发，确定变量之间的数学关系式

- **统计检验**

  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

- **预测**

  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



一元线性回归

定义

回归中只涉及一个自变量x，且因变量y与自变量x之间为线性关系的回归。

说明

- 自变量x：解释变量
- 因变量y：被解释变量
- 一元：一个解释变量x
- 线性：因变量y与自变量x之间呈现**直线关系**。

一元线性回归模型

基本假定

假定1：因变量y与自变量x之间具有**线性**关系。

假定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。

假定3：误差项ε是一个**期望值为0**的随机变量，即$E(\varepsilon)=0$。

假定4：对于所有的x值， ε 的**方差 σ^2 都相同。**

假定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$\varepsilon \sim N(0，\sigma^2)$。

一元线性回归模型

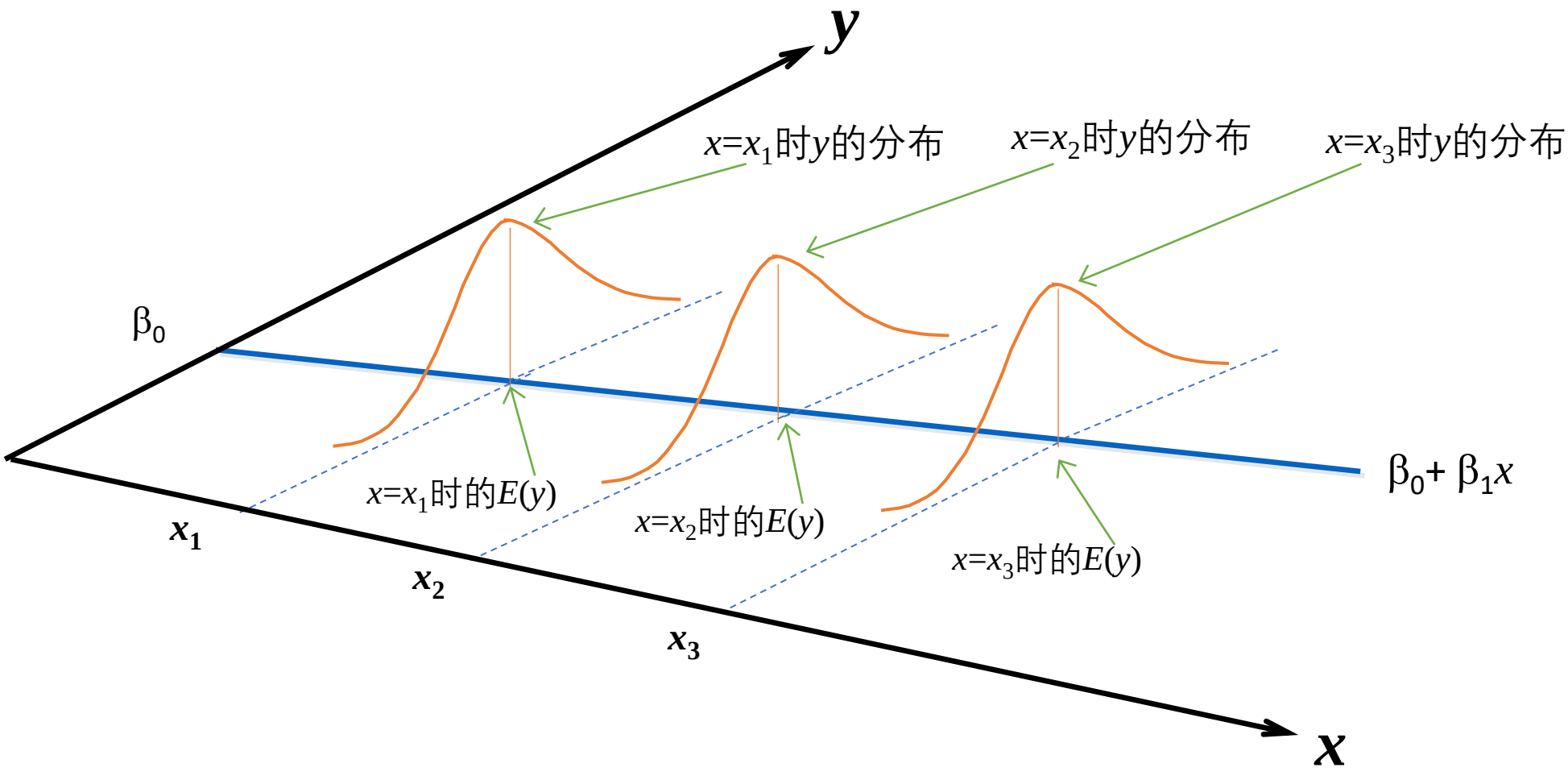
$$y=\beta_0+\beta_1 x+\varepsilon$$

说明

- y由线性部分$\beta_0+\beta_1 x$和误差项ε构成
- $\beta_0+\beta_1 x$：表示x的变化引起y的变化
- ε：表示除了x之外的所有**随机因素**对y的影响
- β_0和β_1：模型的参数

问题

- y与x之间是何种关系？确定性 vs不确定性
- 如何体现？


y
x=x1时y的分布
x=x2时y的分布
x=x3时y的分布
β0
β0+ β1x
x=x1时的E(y)
x=x2时的E(y)
x=x3时的E(y)
x1
x2
x3
x

一元线性回归方程

回归方程的定义

描述因变量y的期望值如何依赖于自变量x的方程。

一元线性回归方程（总体）

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

说明

- 方程图示为一条直线：总体回归线
- β_0：总体回归线的截距
- β_1：总体回归线的斜率

估计的回归方程（样本）

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

说明

- 总体回归参数β_0和β_1未知
- 样本回归参数$\hat{\beta}_0$和$\hat{\beta}_1$是对总体参数的估计
- $\hat{\beta}_0$：样本回归线的截距
- β_1：样本回归线的斜率

参数的最小二乘估计

参数估计方法：最小二乘法（Least Squares）

提出：卡尔·高斯（1777-1855）

目标函数

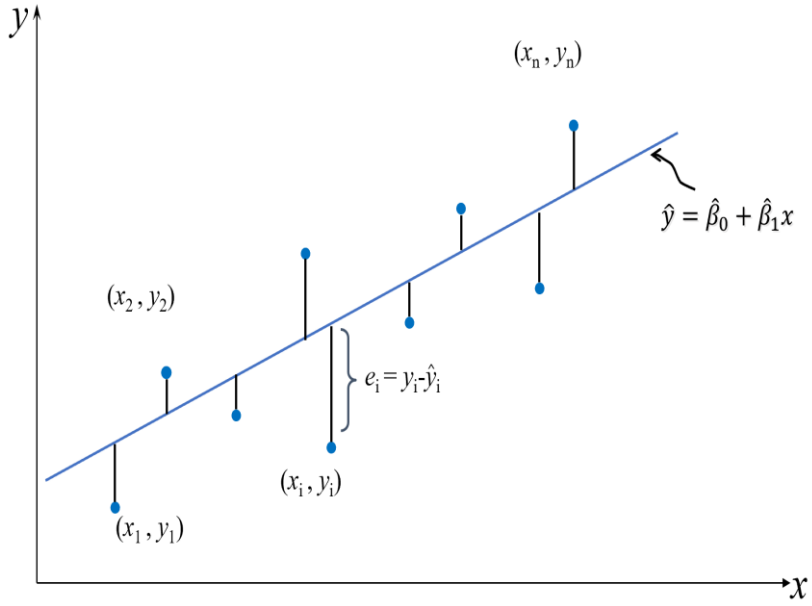
$$\min_{(\beta_0,\beta_1)} Q = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

原理

寻找最小化残差平方和的参数β_0, β_1的取值。

几何含义

寻找所有的点到直线的距离平方和最小的直线。



参数的最小二乘估计：求解

目标函数：

$$\min_{(\beta_0,\beta_1)} Q = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

求解：

$$F.O.C\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \end{cases} \Rightarrow \begin{cases} \hat{\beta}_1 = \dfrac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

参数的最小二乘估计：例子

求解：不良贷款对贷款余额的回归方程

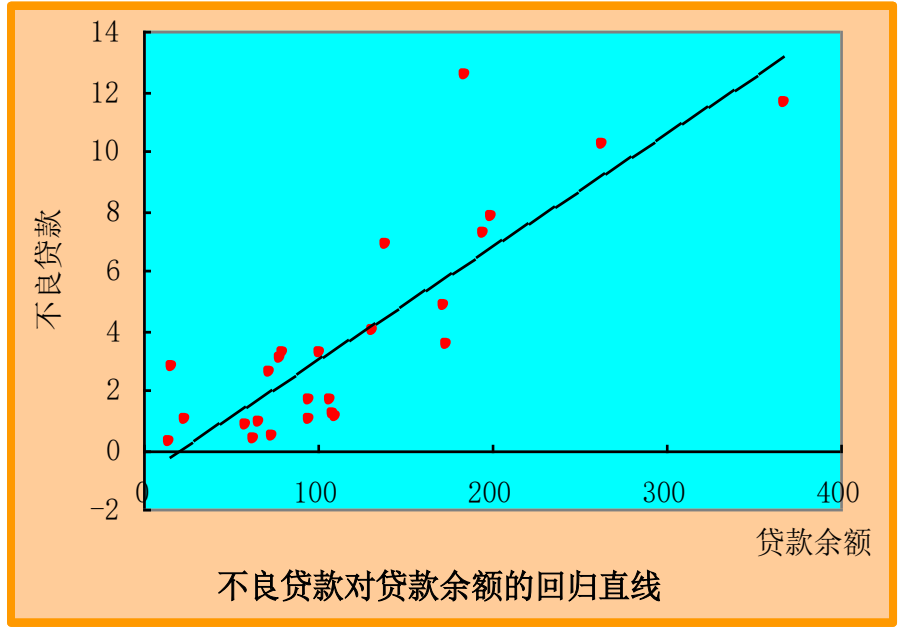
$$\begin{cases}\hat{\beta}_1 = \dfrac{25 \times 17080.14 - 3006.7 \times 93.2}{25 \times 516543.37 - (3006.7)^2} = 0.037895 \\ \hat{\beta}_0 = 3.728 - 0.037895 \times 120.268 = -0.8295\end{cases}$$

回归方程

$$\hat{y} = -0.8295 + 0.03795x$$

经济含义

贷款余额每增加1亿元，不良贷款平均增加0.037895亿元。

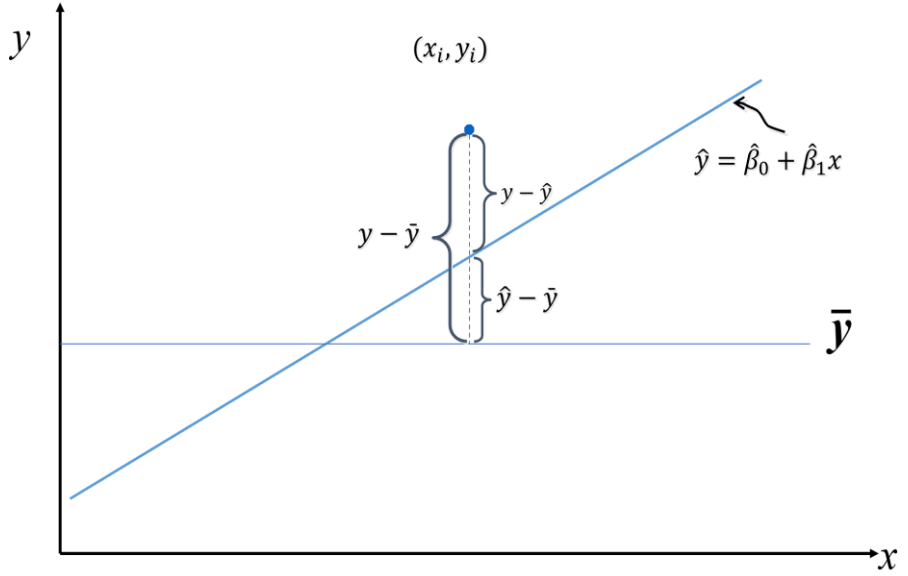


不良贷款对贷款余额的回归直线

回归直线的拟合优度：判定系数

方差分解：

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{总平方和 (SST)}}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{回归平方和 (SSR)}}+\underbrace{\sum_{i=1}^{n}(y_i-\hat{y})^2}_{\text{残差平方和 (SSE)}}$$



判定系数R^2：

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

说明：

- $R^2 \to 1$，说明回归方程拟合的越
- $R^2 \to 0$，说明回归方程拟合的越差
- 对于**一元回归**，判定系数等于相关系数的平方，即$R^2 = r^2$

回归直线的拟合优度：判定系数的计算

例子：

计算不良贷款对贷款余额回归的判定系数，并解释其意义

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

经济含义：

在不良贷款取值的变动中，有71.16%是由贷款余额所决定的。

回归直线的拟合优度：估计的标准误差

公式

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

作用：度量实际观测值y_i与回归估计值$\hat{y}_i$之间的差异程度。

例子：计算不良贷款对贷款余额回归的估计标准误差，并解释其意义

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{SST - SSR}{n-2}} = \sqrt{\frac{312.6504 - 222.4860}{25-2}} = 1.9799(\text{亿元})$$

经济含义：根据贷款余额估计不良贷款时，估计值的平均误差是1.9799亿元。

显著性检验

“显著性”的含义

某个(某些)自变量(系数)的影响是否**显著不同于零**。

目的：回归模型（系数）是否反映了自变量X和因变量Y之间的真实关系。

步骤

- 提出假设　　H_0:系数等于零（原假设）　H_1:系数不等于零（备择假设）
- 以样本数据为基础，计算检验统计量(F统计量，或t统计量)
- 根据给定的**显著性水平**（如α=0.05），查表获得临界值，比较检验统计量与临界值的大小：
  - 检验统计量>临界值，系数显著，拒绝H_0；
  - 检验统计量<临界值,系数不显著，不能拒绝H_0 。

类型

- 线性关系检验（回归方程的整体显著性）：**多个变量系数**的显著性的**联合检验**，通常采用**F检验**。
- 回归系数检验（单个变量系数的显著性）：**单个变量系数**的显著性检验，通常采用**t检验**。

显著性检验：线性关系检验

含义

检验自变量x与因变量y之间的**线性关系是否显著**（回归方程的**整体显著性检验**）

实质

所有自变量X的系数**是否同时为零**

方式： F统计量

步骤

- **提出假设**

  H_0: $\beta_1 = 0$ (线性关系不显著)

  H_1: $\beta_1 \neq 0$ (线性关系显著)

- **计算检验统计量**

$$F = \frac{SSR/_1}{SSE/_{(n-2)}} = \frac{MSR}{MSE} \sim F(1, n-2)$$

- **确定显著性水平α，并依据分子自由度1和分母自由度n-2，查找临界值F_α**
- **做出统计决策**

  若$F > F_\alpha$，拒绝H_0

  若$F < F_\alpha$，不能拒绝H_0

显著性检验：线性关系检验的例子

检验：不良贷款与贷款余额之间线性关系

- **提出假设**

  H_0: $\beta_1 = 0$ (线性关系不显著)

  H_1: $\beta_1 \neq 0$ (线性关系显著)

	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			

- **计算检验统计量**

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

- **确定显著性水平$\alpha = 0.05$, 并依据分子自由度1和分母自由度n-2，查表知，临界值$F_\alpha = 4.28$**
- **做出统计决策**

  统计量F大于临界值F_α, 拒绝H_0，不得不接受H_1，即不良贷款与贷款余额之间的线性关系显著

显著性检验：回归系数检验

含义

检验某个自变量x与因变量y的**影响是否显著**（单个变量系数的**显著性**）

实质

某个自变量X的系数**是否为零**

方式：t统计量

说明：

在一元线性回归中，回归系数检验与线性关系的检验等效：$t_{n-2}^2 = F_{1,n-2}$

步骤

- **提出假设**

  H_0: $\beta_1 = 0$ (影响不显著)

  H_1: $\beta_1 \neq 0$ (影响显著)

- **计算检验统计量**

  $$\mathrm{t} = \frac{\widehat{\beta}_1}{s_{\widehat{\beta}_1}} \sim t(n-2)$$

- **确定显著性水平α, 并依据自由度，查找临界值$t_{\alpha/2}$**

- **做出统计决策**

  若$|t| > t_{\alpha/2}$, 拒绝H_0

  若$|t| < t_{\alpha/2}$, 不能拒绝H_0

显著性检验：回归系数检验的例子

检验：贷款余额对不良贷款的影响是否显著

- **提出假设**

  H_0: $\beta_1 = 0$ (影响不显著)

  H_1: $\beta_1 \neq 0$ (影响显著)

- **计算检验统计量**

	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

$$\mathrm{t} = \frac{\widehat{\beta}_1}{s_{\widehat{\beta}_1}} = \frac{0.037895}{0.005030} = 7.533515$$

- **确定显著性水平$\alpha = 0.05$, 并依据自由度23，查表知，临界值$t_{\alpha/2} = 2.201$**
- **做出统计决策**

  统计量t大于临界值$t_{\alpha/2}$, 拒绝H_0，不得不接受H_1，即贷款余额对不良贷款的影响显著

回归模型的评价：“好”模型的四个标准

标准1：符合预期

- **含义**：即回归系数符号与理论或事先预期一致。
- 例子：在不良贷款与贷款余额的回归中，我们预期：贷款余额越多，不良贷款也可能会越多，即回归系数的值应该是正的。在回归方程中，贷款余额的回归系数$\hat{\beta}_1 = 0.03795$，符号为正，符合预期。

标准2：统计显著

- **含义**：回归系数通过t检验和F检验。
- **例子**：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$\hat{\beta}_1$的t检验和F检验结果表明：不良贷款与贷款余额之间的线性关系是统计上显著的。

回归模型的评价：“好”模型的四个标准

标准3：解释程度高

- **含义**：模型具有较高的判定系数R^2。
- **例子**：在不良贷款与贷款余额的回归中，判定系数R2=71.16%，解释了不良贷款变差的2/3以上，说明模型的解释程度不错。

标准4：好的残差

- **含义**：残差e接近，或者服从正态分布。
- **依据**：$\varepsilon \sim N(0,\sigma^2)$,残差e是误差项 ε 的替代，也应服从正态分布
- **重要性**：模型的F检验和t检验均依赖于残差的正态分布假设。若残差不服从正态分布，则上述检验无效。
- **检验方法**：直方图，JB统计量
- **补救措施**：识别出未知的自变量，增大样本容量等

8.3 利用回归方程进行预测

- 点估计
- 区间估计

利用回归方程进行预测

含义

预测（Predict），是指通过自变量x的取值来预测因变量y的取值。

例子

根据不良贷款y与贷款余额x的回归方程，给定一个贷款余额x的取值，就可以得到一个不良贷款y的预测值。

类型

- 点估计：对于x的一个特定值x_0，根据回归方程得到y的一个**估计值**。
  - y的**平均值**的点估计
  - y的**个别值**的点估计
- 区间估计：对于x的一个特定值x_0，根据回归方程得到y的一个**估计区间**。
  - 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：平均值 vs 个别值

平均值的点估计

- 含义

  对于自变量x的一个给定值x_0，求因变量y的**平均值**的一个估计值$E(y_0)$。

- **说明：**x_0为给定值，**并非**样本中的真实值

- 例子

  给定贷款余额为100亿元时，估计所有分行不良贷款的平均值，即：

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

个别值的点估计

- 含义

  对于自变量x的一个给定值x_0，求因变量y的**个别值**的一个估计值$\hat{y}_0$。

- **说明：**x_0为样本中的**真实值**

- 例子

  给定某分行贷款余额为72.8亿元时，估计该分行不良贷款的个别值，即：

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

区间估计：置信区间

含义

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，即为**置信区间**(confidence interval)。

- **说明：平均值→置信区间**
- **公式**

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

例子

求解：当贷款余额为100亿元时，不良贷款95%置信水平下的置信区间。

解：已知$n = 25,\ \hat{y}_0 = 2.96,\ s_e = 1.9799$

$$t_{\alpha/2}(25 - 2) = 2.069$$

置信区间：

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

即 $2.1141 \leq E(y_0) \leq 3.8059$

含义：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

区间估计：预测区间

含义

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**个别值**的估计区间，即为**预测区间**(Prediction interval)。

- **说明：个别值→预测区间**
- **公式**

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

例子

求解：当某分行的贷款余额为72.8亿元时，不良贷款95%置信水平下的预测区间。

解：已知$n=25,\ \hat{y}_0=1.93,\ s_e=1.9799$

$$t_{\alpha/2}(25-2)=2.069$$

预测区间：

$$1.93 \pm 2.069 \times 1.9799 \times \sqrt{1+\frac{1}{25}+\frac{(72.8-120.268)^2}{154933.5744}}$$

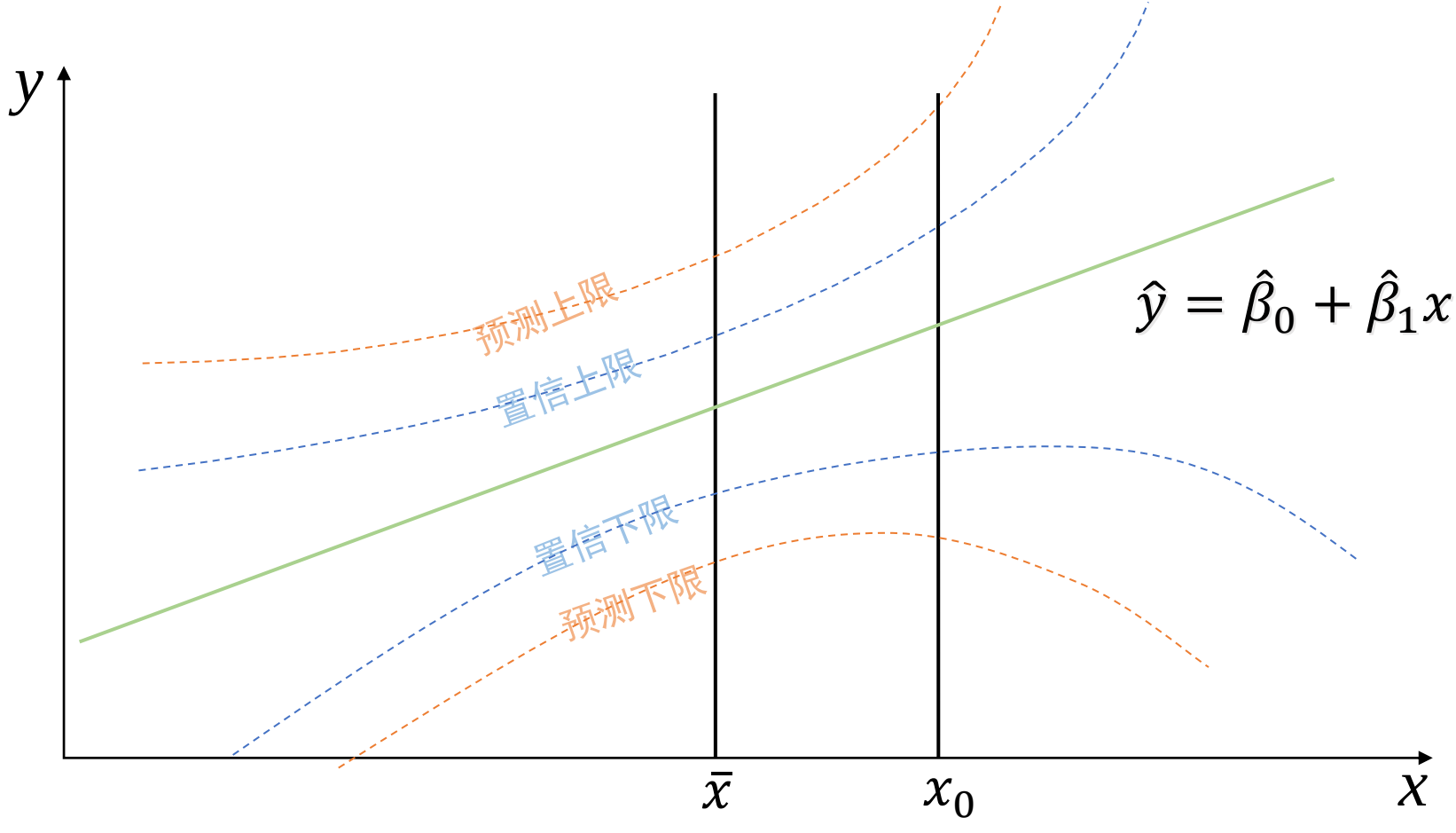
即 $-2.2766 \le \hat{y}_0 \le 6.1366$

含义：当某分行贷款余额为72.8亿元时，不良贷款的预测值在-2.2766亿元到6.1366亿元之间

区间估计：置信区间 vs 预测区间

	A	B	C	D	E	F	G	H
1	分行编号	不良贷款（y）	贷款余额（x）	预测 Y	置信区间		预测区间	
2					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

区间估计：置信区间 vs 预测区间



8.4 残差分析

- 残差与残差图
- 标准化

残差与残差图

残差

利用因变量的观测值y_i与根据估计的回归方程求出的预测值$\hat{y}_i$。

公式

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

含义

残差反映了用估计的回归方程去预测而引起的误差。

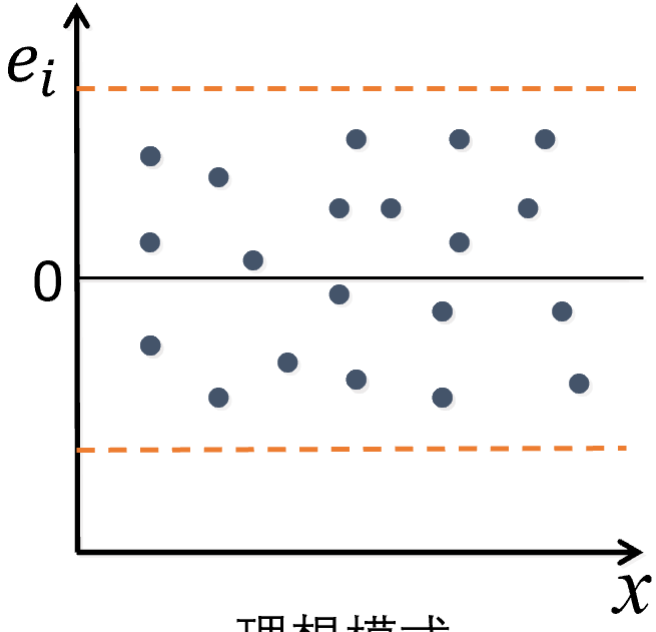
残差图

- 残差与X的XY散点图
- 残差与$\hat{y}_i$的XY散点图
- 标准化残差图

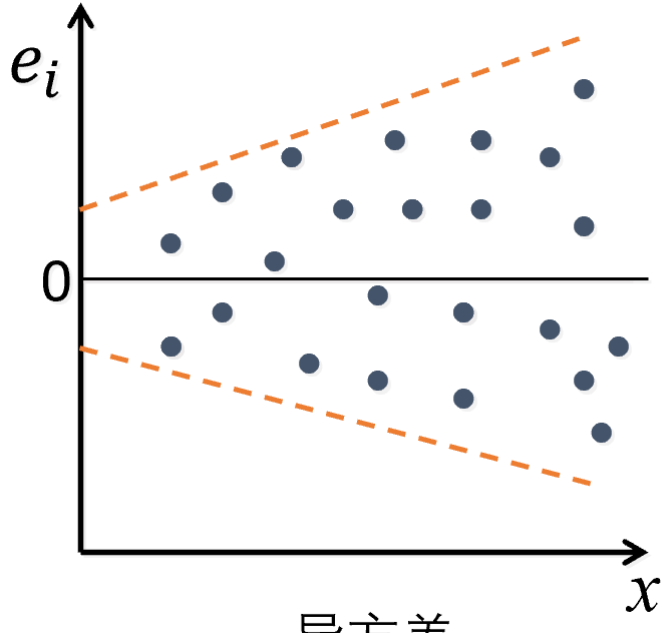
作用

用于判断误差ε的假定是否成立

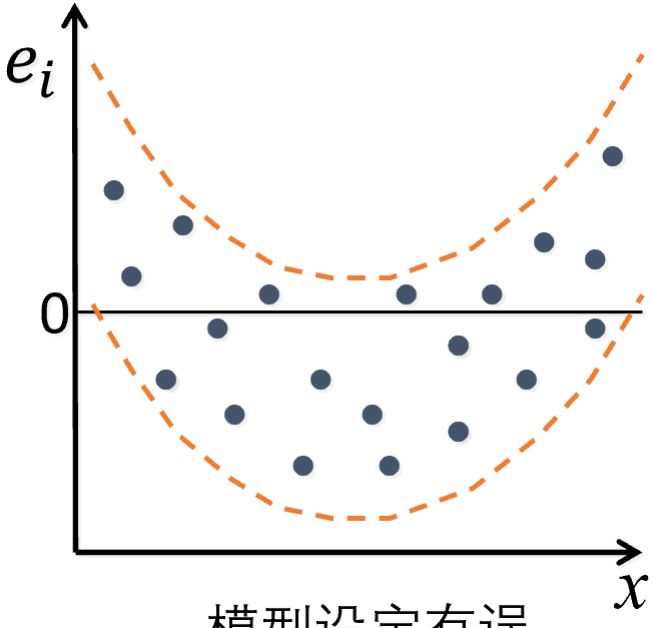
残差图：形态与判别



理想模式



异方差



模型设定有误

标准化残差与残差图

标准化残差

残差除以它的标准差。

公式

$$z_{e_i} = \frac{e_i}{s_e} = \frac{y_i - \hat{y}_i}{s_e}$$

Excel

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

标准化残差图

- **作用**

  用以直观地判断误差项服从正态分布这一假定是否成立

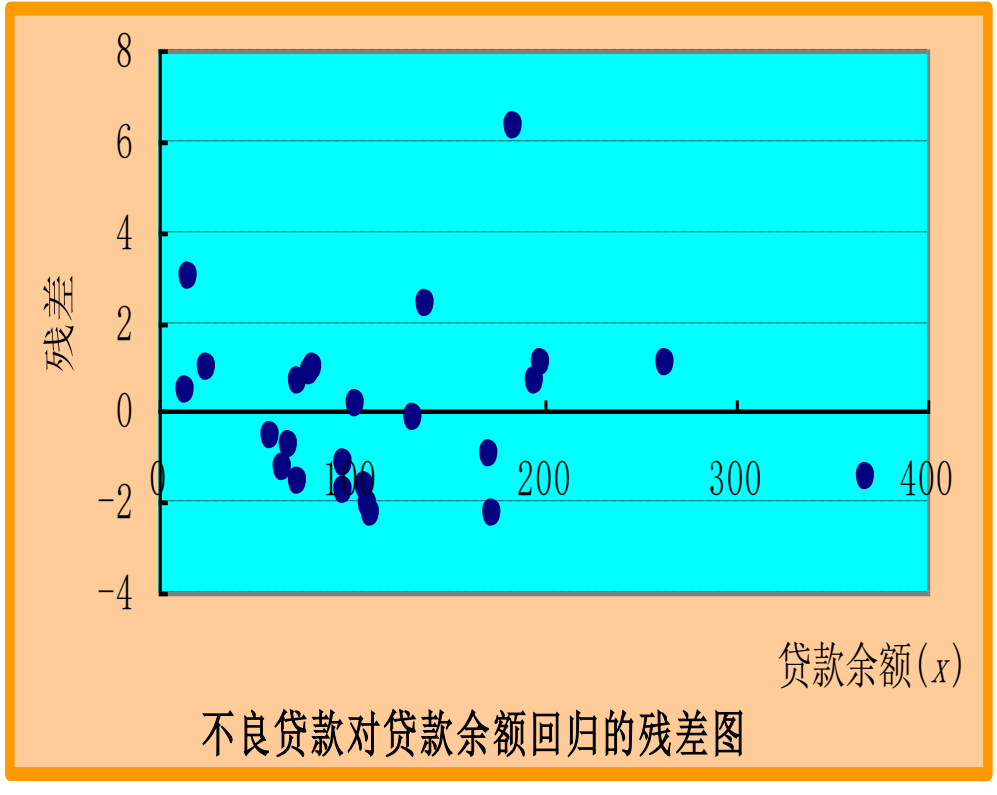
- **原理**

  若误差项服从正态分布，则标准化残差服从标准正态分布

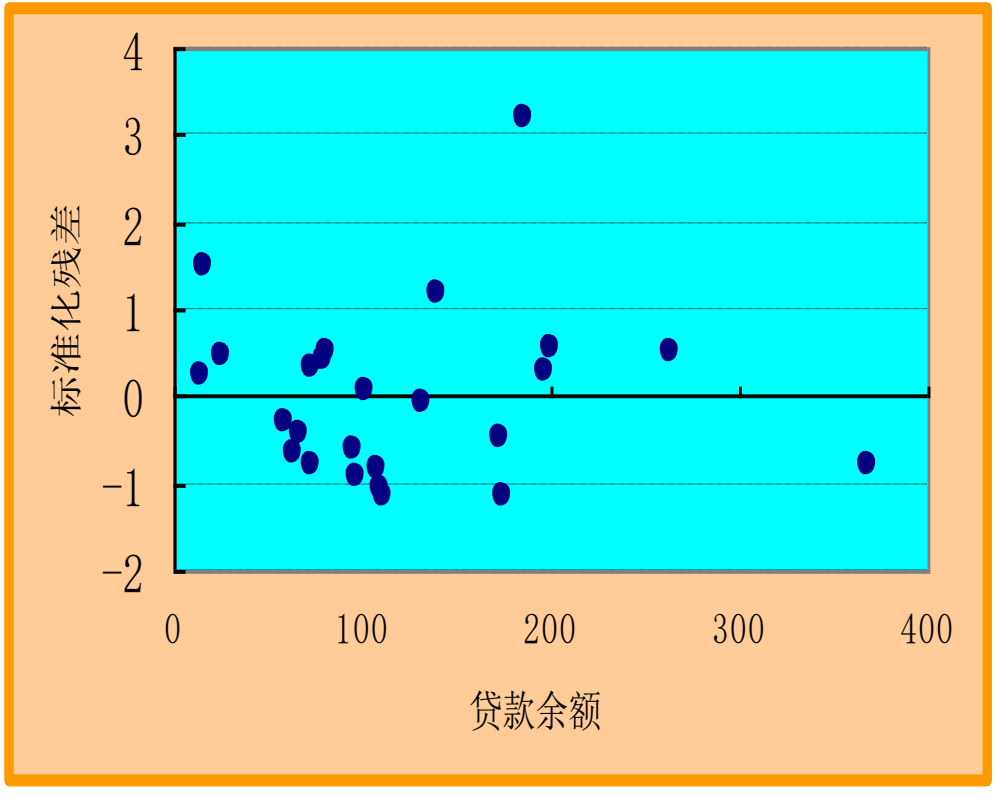
- **含义**

  依据经验法则，在标准化残差图中，大约有95%的标准化残差在-2到+2之间

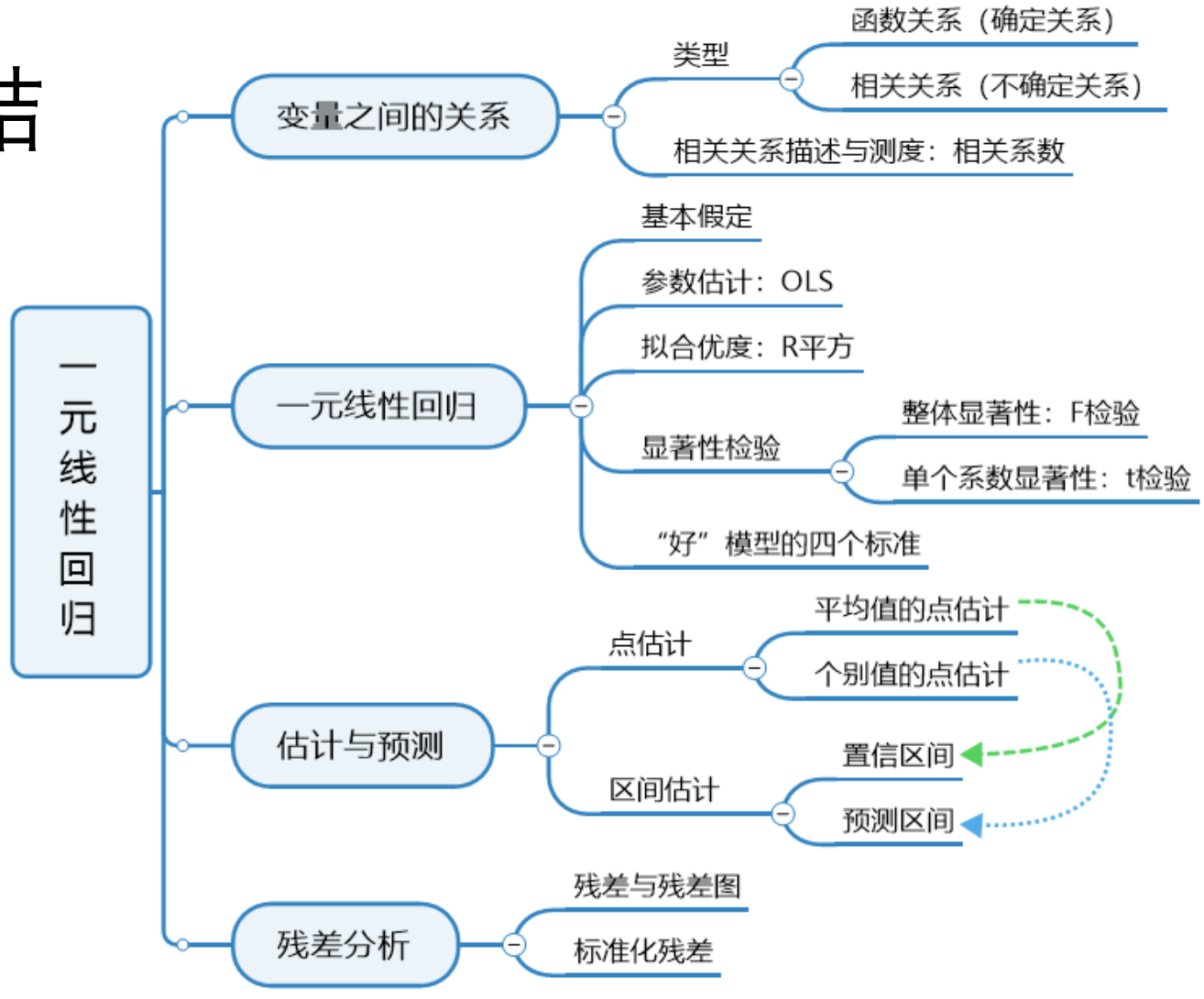
残差图 vs 标准化残差图



不良贷款对贷款余额回归的残差图



本章小结


一元线性回归
变量之间的关系
类型
函数关系（确定关系）
相关关系（不确定关系）
相关关系描述与测度：相关系数
一元线性回归
基本假定
参数估计：OLS
拟合优度：R平方
显著性检验
整体显著性：F检验
单个系数显著性：t检验
“好”模型的四个标准
估计与预测
点估计
平均值的点估计
个别值的点估计
区间估计
置信区间
预测区间
残差分析
残差与残差图
标准化残差