第12章　多元线性回归

PowerPoint

第12章　多元线性回归

12.1　多元线性回归模型
12.2　回归方程的拟合优度
12.3　显著性检验
12.4　多重共线性
12.5　利用回归方程进行估计和预测
12.6　变量选择与逐步回归

学习目标

1. 回归模型、回归方程、估计的回归方程
2. 回归方程的拟合优度
3. 回归方程的显著性检验
4. 多重共线性问题及其处理
5. 利用回归方程进行估计和预测
6. 变量选择与逐步回归
7. 用 **Excel** 进行回归分析

12.1 多元线性回归模型

12.1.1 多元回归模型与回归方程
12.1.2 估计的多元回归方程
12.1.3 参数的最小二乘估计

作者：贾俊平，中国人民大学统计学院

多元回归模型与回归方程

作者：贾俊平，中国人民大学统计学院

多元回归模型
(multiple regression model)

1. 一个因变量与两个及两个以上自变量的回归
2. 描述因变量 y 如何依赖于自变量 x_1，x_2，…，x_k 和误差项 ε 的方程，称为多元回归模型
3. 涉及 k 个自变量的多元回归模型可表示为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

- β_0，β_1，β_2，…，β_k是参数
- ε是被称为误差项的随机变量
- y 是x_1，x_2，…，x_k 的线性函数加上误差项ε
- ε包含在y里面但不能被k个自变量的线性关系所解释的变异性

多元回归模型
(基本假定)

1. 误差项ε是一个期望值为0的随机变量，即$E(\varepsilon)=0$
2. 对于自变量x_1，x_2，…，x_k的所有值，ε的方差σ^2都相同
3. 误差项ε是一个服从正态分布的随机变量，即$\varepsilon \sim N(0,\sigma^2)$，且相互独立

多元回归方程
(multiple regression equation)

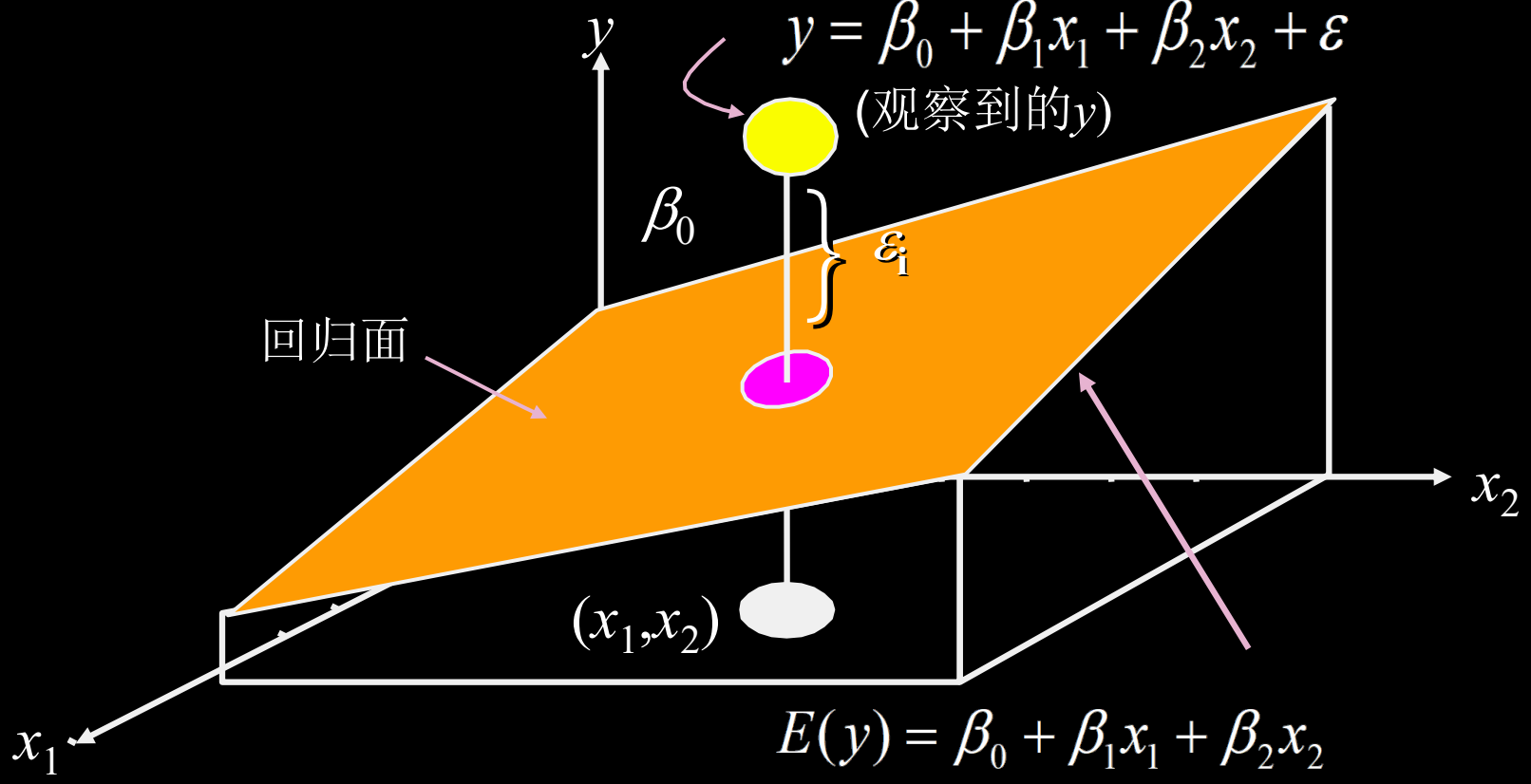
1. 描述因变量 y 的平均值或期望值如何依赖于自变量 x_1， x_2， …， x_k的方程
2. 多元线性回归方程的形式为

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- β_1，β_2，…，β_k称为偏回归系数
- β_i 表示假定其他变量不变，当 x_i 每变动一个单位时，y 的平均变动值

二元回归方程的直观解释

二元线性回归模型



估计的多元回归方程

作者：贾俊平，中国人民大学统计学院

估计的多元回归的方程
(estimated multiple regression equation)

1. 用样本统计量 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ 估计回归方程中的 参数 $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ 时得到的方程
2. 由最小二乘法求得
3. 一般形式为

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

- $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ 是 $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ 估计值
- $\hat{y}$ 是 y 的估计值

参数的最小二乘估计

作者：贾俊平，中国人民大学统计学院

参数的最小二乘法

1. 使因变量的观察值与估计值之间的离差平方和达到最小来求得 $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\cdots$, $\hat{\beta}_k$ 。即

$$Q(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \text{最小}$$

2. 求解各回归参数的标准方程如下

$$\begin{cases} \left.\dfrac{\partial Q}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0} = 0 \\ \left.\dfrac{\partial Q}{\partial \beta_i}\right|_{\beta_i=\hat{\beta}_i} = 0 \qquad (i = 1,2,\cdots, \ k) \end{cases}$$

参数的最小二乘法
(例题分析)

【例】一家大型商业银行在多个地区设有分行，为弄清楚不良贷款形成的原因，抽取了该银行所属的25家分行的有关业务数据。试建立不良贷款y与贷款余额x_1、累计应收贷款x_2、贷款项目个数x_3和固定资产投资额x_4的线性回归方程，并解释各回归系数的含义

12.2 回归方程的拟合优度

12.2.1 多重判定系数

12.2.2 估计标准误差

作者：贾俊平，中国人民大学统计学院

多重判定系数

作者：贾俊平，中国人民大学统计学院

多重判定系数
(multiple coefficient of determination)

1. 回归平方和占总平方和的比例
2. 计算公式为

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

3. 因变量取值的变差中，能被估计的多元回归方程所解释的比例

修正多重判定系数

(adjusted multiple coefficient of determination)

1. 用样本量*n*和自变量的个数*k*去修正R^2得到
2. 计算公式为

$$R_a^2 = 1 - (1 - R^2) \times \frac{n-1}{n-k-1}$$

3. 避免增加自变量而高估 R^2
4. 意义与 R^2类似
5. 数值小于R^2

估计标准误差 S_y

1. 对误差项ε的标准差σ的一个估计值
2. 衡量多元回归方程的拟合优度
3. 计算公式为

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

12.3 显著性检验

12.3.1 线性关系检验

12.3.2 回归系数检验和推断

作者：贾俊平，中国人民大学统计学院

线性关系检验

作者：贾俊平，中国人民大学统计学院

线性关系检验

1. 检验因变量与所有自变量之间的线性关系是否显著
2. 也被称为总体的显著性检验
3. 检验方法是将回归均方(*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
   - 如果是显著的，因变量与自变量之间存在线性关系
   - 如果不显著，因变量与自变量之间不存在线性关系

线性关系检验

1. 提出假设
   - H_0：$\beta_1=\beta_2=\ldots=\beta_k=0$　线性关系不显著
   - H_1：β_1，β_2，…β_k至少有一个不等于0
2. 计算检验统计量F

$$F=\frac{SSR/k}{SSE/(n-k-1)}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2/k}{\sum_{i=1}^{n}(y_i-\hat{y})^2/(n-k-1)}\sim F(k,n-k-1)$$

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

回归系数检验和推断

作者：贾俊平，中国人民大学统计学院

回归系数的检验

1. 线性关系检验通过后，对各个回归系数有选择地进行一次或多次检验
2. 究竟要对哪几个回归系数进行检验，通常需要在建立模型之前作出决定
3. 对回归系数检验的个数进行限制，以避免犯过多的第Ⅰ类错误(弃真错误)
4. 对每一个自变量都要单独进行检验
5. 应用 *t* 检验统计量

回归系数的检验 (步骤)

1. 提出假设
   - H_0： $\beta_i = 0$ (自变量 x_i 与因变量 y 没有线性关系)
   - H_1： $\beta_i \neq 0$ (自变量 x_i 与因变量 y 有线性关系)
2. 计算检验的统计量 t

$$t = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}} \sim t(n-k-1)$$

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

回归系数的推断
(置信区间)

- 回归系数在(1-α)%置信水平下的置信区间为

$$\hat{\beta}_i \pm t_{\alpha/2}(n-k-1)s_{\hat{\beta}_i}$$

回归系数的抽样标准差

$$s_{\hat{\beta}_i} = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$$

12.4 多重共线性

12.4.1 多重共线性及其所产生的问题
12.4.2 多重共线性的判别
12.4.3 多重共线性问题的处理

作者：贾俊平，中国人民大学统计学院

多重共线性及其产生的问题

作者：贾俊平，中国人民大学统计学院

多重共线性 (multicollinearity)

1. 回归模型中两个或两个以上的自变量彼此相关
2. 多重共线性带来的问题有
   - 可能会使回归的结果造成混乱，甚至会把分析引入歧途
   - 可能对参数估计值的正负号产生影响，特别是各回归系数的正负号有可能同预期的正负号相反

多重共线性的识别

作者：贾俊平，中国人民大学统计学院

多重共线性的识别

1. 检测多重共线性的最简单的一种办法是计算模型中各对自变量之间的相关系数，并对各相关系数进行显著性检验
2. 若有一个或多个相关系数显著，就表示模型中所用的自变量之间相关，存在着多重共线性

多重共线性的识别

- 如果出现下列情况，暗示存在多重共线性
  - 模型中各对自变量之间显著相关
  - 当模型的线性关系检验(*F*检验)显著时，几乎所有回归系数的*t*检验却不显著
  - 回归系数的正负号与预期的相反
  - 容忍度(tolerance)与方差扩大因子(variance inflation factor，VIF)。
    - 某个自变量的容忍度等于1减去该自变量为因变量而其他k-1个自变量为预测变量时所得到的线性回归模型的判定系数，即$1-R_i^2$。容忍度越小，多重共线性越严重。通常认为容忍度小于0.1时，存在严重的多重共线性
    - 方差扩大因子等于容忍度的倒数，即$VIF=\frac{1}{1-R_i^2}$。显然，VIF越大多重共线性就越严重。一般认为VIF大于10则认为存在严重的多重共线性。

SPSS 输出结果

多重共线性
(例题分析)

【例】判别各自变量之间是否存在多重共线性

贷款余额、应收贷款、贷款项目、固定资产投资额之间的相关矩阵

	A	B	C	D	E
1		各项贷款余额	本年累计应收贷款	贷款项目个数	本年固定资产投资额
2	各项贷款余额	1			
3	本年累计应收贷款	0.678772	1		
4	贷款项目个数	0.848416	0.585831	1	
5	本年固定资产投资额	0.779702	0.472431	0.746646	1

多重共线性
(例题分析)

【例】判别各自变量之间是否存在多重共线性

相关系数的检验统计量

	A	B	C	D
1		各项贷款余额	本年累计应收贷款	贷款项目个数
2	各项贷款余额	1		
3	本年累计应收贷款	4.432870	1	
4	贷款项目个数	7.686824	3.466726	1
5	固定资产投资额	5.971918	2.570663	5.382848

多重共线性
(例题分析)

1. $t_{\alpha/2}(25-2)=2.069$，所有统计量$t > t_{\alpha/2}(25-2)=2.069$，所以均拒绝原假设，说明这4个自变量两两之间都有显著的相关关系
2. 由表Excel输出的结果可知，回归模型的线性关系显著(Significance-F＝1.03539E-06<α=0.05)。而回归系数检验时却有3个没有通过t检验(P-Value=0.074935，0.862853，0.067030>α=0.05) 。这也暗示了模型中存在多重共线性
3. 固定资产投资额的回归系数为负号(-0.029193)，与预期的不一致

多重共线性问题的处理

作者：贾俊平，中国人民大学统计学院

多重共线性
(问题的处理)

1. 将一个或多个相关的自变量从模型中剔除，使保留的自变量尽可能不相关
2. 如果要在模型中保留所有的自变量，则应
   - 避免根据 *t* 统计量对单个参数进行检验
   - 对因变量值的推断(估计或预测)的限定在自变量样本值的范围内

提示

1. 在建立多元线性回归模型时，不要试图引入更多的自变量，除非确实有必要
2. 在社会科学的研究中，由于所使用的大多数数据都是非试验性质的，因此，在某些情况下，得到的结果往往并不令人满意，但这不一定是选择的模型不合适，而是数据的质量不好，或者是由于引入的自变量不合适

12.5 利用回归方程进行估计和预测

软件应用

作者：贾俊平，中国人民大学统计学院

用SPSS求置信区间和预测区间

第 1 步：选择【分析】→【回归-线性】，进入主对话框。

第 2 步：在对话框中将因变量选入【因变量】，将所有自变量选入【自变量】，并在【方法】下选择【逐步】。

第 3 步：点击【选项】，并在【步进方法标准】下选中【使用 F 的概率】，并在【进入】框中输入增加变量所要求的显著性水平（隐含值为 0.05，一般不用改变）；在【删除】输入剔除变量所要求的显著性水平（隐含值为 0.10，一般不用改变）。点击【继续】回到主对话框。点击【确定】。

（注：需要预测时，点击【保存】，在【预测值】下选中【未标准化】（输出点预测值）；在【预测区间】下选中【均值】和【单值】（输出置信区间和预测区间）；在【置信区间】中选择所要求的置信水平（隐含值为 95%，一般不用改变）。需要残差分析时，在【残差】下选中所需的残差。需要输出标准化残差的直方图和正态概率图时，点击【绘制】，在【标准化残差图】下选中【直方图】和【正态概率图】。）

置信区间和预测区间
(例题分析)

	不良贷款	贷款余额	累计应收贷款	贷款项目个数	固定资产投资额	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
1	0.9	67.3	6.8	5	51.9	1.23718	-0.18541	2.65976	-2.73660	5.21095
2	1.1	111.3	19.8	16	90.9	3.94465	1.66643	6.22287	-0.40937	8.29867
3	4.8	173.0	7.7	17	73.7	5.14051	3.85596	6.42507	1.21403	9.06699
4	3.2	80.8	7.2	10	14.5	3.00138	1.78174	4.22102	-0.90434	6.90711
5	7.8	199.7	16.5	19	63.2	7.84785	6.43467	9.26102	3.87743	11.81827
6	2.7	16.2	2.2	1	2.2	-0.09702	-1.62200	1.42796	-4.10860	3.91455
7	1.6	107.4	10.7	17	20.2	4.51985	3.01816	6.02155	0.51707	8.52263
8	12.5	185.4	27.1	18	43.8	9.39626	6.74006	12.05245	4.83309	13.95943
9	1.0	96.1	1.7	10	55.9	1.59121	0.16342	3.01901	-2.38443	5.56686
10	2.6	72.8	9.1	14	64.3	1.56664	0.32302	2.81027	-2.34664	5.47992
11	0.3	64.2	2.1	11	42.7	0.77305	-0.42179	1.96788	-3.12500	4.67109
12	4.0	132.2	11.2	23	76.7	4.02462	2.64760	5.40165	0.06693	7.98232
13	0.8	58.6	6.0	14	22.8	1.75068	0.40366	3.09771	-2.19667	5.69804
14	3.5	174.6	12.7	26	117.1	4.80854	3.19196	6.42512	0.76126	8.85582
15	10.2	263.5	15.6	34	146.7	8.04946	6.06303	10.03590	3.84077	12.25815
16	3.0	79.3	8.9	15	29.9	2.81606	1.62898	4.00313	-1.07962	6.71173
17	0.2	14.8	0.6	2	42.1	-1.54020	-3.24509	0.16470	-5.62356	2.54316
18	0.4	73.5	5.9	11	25.3	2.21590	1.18479	3.24701	-1.63512	6.06691
19	1.0	24.7	5.0	4	13.4	0.37443	-0.87232	1.62119	-3.53984	4.28871
20	6.8	139.4	7.2	28	64.3	4.15541	1.94297	6.36785	-0.16455	8.47537
21	11.6	368.2	16.8	32	163.9	11.88805	9.06656	14.70954	7.22672	16.54938
22	1.6	95.7	3.8	10	44.5	2.21887	1.03994	3.39779	-1.67434	6.11207
23	1.2	109.6	10.3	14	67.9	3.11264	2.22239	4.00289	-0.70308	6.92836
24	7.2	196.2	15.8	16	39.7	8.24653	6.33060	10.16246	4.07065	12.42240
25	3.2	102.2	12.0	10	97.1	2.15746	0.17604	4.13887	-2.04887	6.36378

12.6 变量选择与逐步回归

作者：贾俊平，中国人民大学统计学院

变量选择过程

1. 在建立回归模型时，对自变量进行筛选
2. 选择自变量的原则是对统计量进行显著性检验
   - 将一个或一个以上的自变量引入到回归模型中时，是否使得残差平方和(*SSE*)有显著地减少。如果增加一个自变量使*SSE*的减少是显著的，则说明有必要将这个自变量引入回归模型，否则，就没有必要将这个自变量引入回归模型
   - 确定引入自变量是否使*SSE*有显著减少的方法，就是使用*F*统计量的值作为一个标准，以此来确定是在模型中增加一个自变量，还是从模型中剔除一个自变量
3. 变量选择的方法主要有：向前选择、向后剔除、逐步回归、最优子集等

向前选择 (forward selection)

1. 从模型中没有自变量开始
2. 对*k*个自变量分别拟合对因变量的一元线性回归模型，共有*k*个，然后找出F统计量的值最高的模型及其自变量，并将其首先引入模型
3. 分别拟合引入模型外的*k*-1个自变量的线性回归模型
4. 如此反复进行，直至模型外的自变量均无统计显著性为止

向后剔除
(backward elimination)

1. 先对因变量拟合包括所有*k*个自变量的回归模型。然后考察*p*(*p*<*k*)个去掉一个自变量的模型(这些模型中每一个都有的*k*-1个自变量)，使模型的*SSE*值减小最少的自变量被挑选出来并从模型中剔除
2. 考察p-1个再去掉一个自变量的模型(这些模型中在每一个都有*k*-2个的自变量)，使模型的*SSE*值减小最少的自变量被挑选出来并从模型中剔除
3. 如此反复进行，一直将自变量从模型中剔除，直至剔除一个自变量不会使*SSE*显著减小为止

逐步回归
(stepwise regression)

1. 将向前选择和向后剔除两种方法结合起来筛选自变量
2. 在增加了一个自变量后，它会对模型中所有的变量进行考察，看看有没有可能剔除某个自变量。如果在增加了一个自变量后，前面增加的某个自变量对模型的贡献变得不显著，这个变量就会被剔除
3. 按照方法不停地增加变量并考虑剔除以前增加的变量的可能性，直至增加变量已经不能导致SSE显著减少
4. 在前面步骤中增加的自变量在后面的步骤中有可能被剔除，而在前面步骤中剔除的自变量在后面的步骤中也可能重新进入到模型中

逐步回归
(例题分析—SPSS输出结果)

输入／移去的变量[a]

模型	输入的变量	移去的变量	方法
1	各项贷款余额	.	步进（准则: F-to-enter 的概率 <= .050，F-to-remove 的概率 >= .100）。
2	本年固定资产投资额	.	步进（准则: F-to-enter 的概率 <= .050，F-to-remove 的概率 >= .100）。

a. 因变量: 不良贷款

模型汇总

模型	R	R 方	调整 R 方	标准 估计的误差
1	.844[a]	.712	.699	1.9799
2	.872[b]	.761	.739	1.8428

a. 预测变量: (常量), 各项贷款余额。
b. 预测变量: (常量), 各项贷款余额, 本年固定资产投资额。

逐步回归
(例题分析—SPSS输出结果)

Anova[c]

模型		平方和	df	均方	F	Sig.
1	回归	222.486	1	222.486	56.754	.000[a]
	残差	90.164	23	3.920		
	总计	312.650	24			
2	回归	237.941	2	118.971	35.034	.000[b]
	残差	74.709	22	3.396		
	总计	312.650	24			

a. 预测变量: (常量), 各项贷款余额。
b. 预测变量: (常量), 各项贷款余额, 本年固定资产投资额。
c. 因变量: 不良贷款

逐步回归
(例题分析—SPSS输出结果)

系数[a]

模型		非标准化系数		标准系数	t	Sig.
		B	标准 误差	试用版		
1	(常量)	-.830	.723		-1.147	.263
	各项贷款余额	.038	.005	.844	7.534	.000
2	(常量)	-.443	.697		-.636	.531
	各项贷款余额	.050	.007	1.120	6.732	.000
	本年固定资产投资额	-.032	.015	-.355	-2.133	.044

a. 因变量: 不良贷款

$$\hat{y} = -0.433 + 0.050x_1 - 0.032x_4$$

本章小结

1. 多元回归模型、回归方程、估计方程
2. 回归方程的拟合优度
3. 显著性检验
4. 多重共线性
5. 利用回归方程进行估计和预测
6. 变量选择与逐步回归
7. 虚拟自变量的回归

结　束

THANKS