《统计学》课程教案（“课程思政”版）

第十章 多元线性回归

教学章节	10.1 多元线性回归模型 10.2 拟合优度和显著性检验 10.3 多重共线性及其处理 10.4 利用回归方程进行预测 10.5 哑变量回归
教学目的	通过本章学习，应注意多元回归中可能存在的多重共线性问题；掌握变量选择的意义及逐步回归的应用；能熟练使用 Excel 和 SPSS 进行回归，并对结果做出合理解释。
主要内容	10.1 节：回归模型与回归方程；参数的最小二乘估计 10.2 节：模型的拟合优度；.显著性检验 10.3 节：多重共线性及其识别；变量选择及逐步回归 10.4 节：利用回归方程进行预测 10.5 节：在模型中引进哑变量；含有一个哑变量的回归
学习要点	10.1 节：▲概念：多元线性回归模型，估计的多元线性回归方程 ▲回归模型的基本假定 ▲偏回归系数的意义 ▲参数的最小二乘估计 ▲用 EXCEL 进行回归 10.2 节：▲概念：多重决定系数，调整的多重决定系数，多重相关系数 ▲多重决定系数的解释 ▲估计标准误差的解释 ▲线性关系检验的步骤 ▲回归系数检验的步骤 ▲线性关系检验与回归系数检验的区别 ▲回归系数的推断 10.3 节：▲概念：多重共线性 ▲多重共线性对回归模型的影响 ▲多重共线性的判别 ▲多重共线性的处理 ▲变量选择的意义 ▲向前选择的基本思想 ▲向后剔除的基本思想 ▲逐步回归的原理和应用 ▲用 SPSS 进行逐步回归 10.4 节：▲用 SPSS 进行预测 10.5 节：▲概念：哑变量 ▲引进哑变量的方法 ▲含有哑变量回归方程的解释
教学方法	直接讲陈法、示例法、提问，启示法。
教具	黑板 幻灯片
教学过程	1.导入新课 2.讲授新课 3.总结及布置练习

教

学

过

程

导入新课：

身高受那些因素影响？

- 决定身高的因素是什么？父母遗传、生活环境、体育锻炼，还是以上各因素的共同作用
- 2004 年 12 月，中国人民大学国民经济管理系 02 级的两位学生，对人大在校生进行了问卷调查。问卷采取随机发放、当面提问当场收回
- 调查的样本量为 98 人，男性 55 人，女性 43 人。调查内容包括被调查者的身高(单位：cm)、性别、其父母身高、是否经常参加体育锻炼、家庭所在地是在南方还是在北方等等。部分数据如教材中的表所示(1 代表男性，0 代表女性)
- 父亲身高、母亲身高、性别是不是影响子女身高的主要因素呢？如果是，子女身高与这些因素之间能否建立一个线性关系方程，并根据这一方程对身高做出预测？
- 这就是本章将要讨论的多元线性回归问题

第 10 章　多元线性回归

10.1　多元线性回归模型

10.1.1　回归模型与回归方程

10.1.2　参数的最小二乘估计

10.1.1　回归模型与回归方程

多元回归模型

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 x1 ， x2 ，…， xk 和误差项 ε 的方程，称为多元回归模型
3. 涉及 k 个自变量的多元线性回归模型可表示为

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

- b_0 ， b_1，b_2 ，…， b_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. 正态性。误差项 ε 是一个服从正态分布的随机变量，且期望值为 0，即 $\varepsilon \sim N(0,\sigma^2)$
2. 方差齐性。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. 独立性。对于自变量 x_1，x_2，…，x_k 的一组特定值，它所对应的ε与任意一组其他值所对应的不相关

多元线性回归方程 (multiple linear 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

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

- b_1，b_2，…，b_k 称为偏回归系数
- b_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

据此来分析教材体系

估计的多元线性回归的方程(estimated multiple linear regression equation)

1.用样本统计量估计回归方程中的参数时得到的方程

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 $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ 的估计值

$\hat{y}$ 是 y 的估计值.

10.1.2 参数的最小二乘估计

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

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$$

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

2. 求解各回归参数的标准方程如下

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

例题分析

【例 10-1】一家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提高，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，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。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

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	回归统计						
4	Multiple R	0.8931					
5	R Square	0.7976					
6	Adjusted R Square	0.7571					
7	标准误差	1.7788					
8	观测值	25					
9							
10	方差分析						
11		df	SS	MS	F	Significance F	
12	回归	4	249.3712	62.3428	19.7040	1.0354E-06	
13	残差	20	63.2792	3.1640			
14	总计	24	312.6504				
15							
16		Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	-1.0216	0.7824	-1.3058	0.2064	-2.6536	0.6104
18	X Variable 1	0.0400	0.0104	3.8375	0.0010	0.0183	0.0618
19	X Variable 2	0.1480	0.0788	1.8787	0.0749	-0.0163	0.3124
20	X Variable 3	0.0145	0.0830	0.1750	0.8629	-0.1587	0.1877
21	X Variable 4	-0.0292	0.0151	-1.9368	0.0670	-0.0606	0.0022

模型汇总

模型	R	R 方	调整 R 方	标准 估计的误差
1	.893[a]	.798	.757	1.7788

a. 预测变量: (常量), 固定资产投资, 累计应收贷款, 贷款项目个数, 贷款余额。

Anova[b]

模型		平方和	df	均方	F	Sig.
1	回归	249.371	4	62.343	19.704	.000[a]
	残差	63.279	20	3.164		
	总计	312.650	24			

a. 预测变量: (常量), 固定资产投资, 累计应收贷款, 贷款项目个数, 贷款余额。
b. 因变量: 不良贷款

系数[a]

模型		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准 误差	试用版			容差	VIF
1	(常量)	-1.022	.782		-1.306	.206		
	贷款余额	.040	.010	.891	3.837	.001	.188	5.331
	累计应收贷款	.148	.079	.260	1.879	.075	.529	1.890
	贷款项目个数	.015	.083	.034	.175	.863	.261	3.835
	固定资产投资	-.029	.015	-.325	-1.937	.067	.360	2.781

a. 因变量: 不良贷款

表 10-4 中还给出了**准化回归系数**（standardized regression coefficient），它是将因变量和所有自变量都标准化后进行回归得到的回归系数。计算标准化回归系数时，首先将因变量和各个自变量进行标准化[1]处理，然后根据标准化后的值进行回归，得到的方程称为**标准化回归方程**（standardized regression equation），该方程中的回归系数就是标准化回归系数，用 β 表示。

标准化回归系数 β_i 的含义是：在其他自变量取值不变的条件下，自变量 x_i（这里是指原始数据）每变动一个标准差，因变量平均变动 β_i 个标准差。显然，某个自变量变动一个标准差改变的因变量的标准差绝对值越大，说明该自变量对因变量的影响就越大，相对于其他自变量而言，它对因变量的预测也就越重要。

例如，表 10—4 中给出的各标准化回归系数为：$\beta_1 = 0.891$，$\beta_2 = 0.260$，$\beta_3 = 0.034$，$\beta_4 = -0.325$。x_1 的标准化回归系数 $\beta_1 = 0.891$表示：在其他自变量不变的条件下，贷款余额每改变一个标准差，不良贷款平均改变 0.891 个标准差。其它系数的含义类似。按标准化回归系数的绝对值大小排序为：$\beta_1 > \beta_4 > \beta_2 > \beta_3$。可见在 4 个自变量中，贷款余额（$x_1$）是预测不良贷款的最重要变量，而贷款项目个数（$x_3$）则是最不重要的变量。

10.2 拟合优度和显著性检验

10.2.1 模型的拟合优度

多重决定系数(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

调整的多重决定系数(adjusted multiple coefficient of determination)

1. 用样本量 n 和自变量的个数 k 去调整 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

多重相关系数(multiple correlation coefficient)

1. 多重决定系数的平方根 R
2. 反映因变量 y 与 k 个自变量之间的相关程度
3. 实际上 R 度量的是因变量的观测值 y 与由多元回归方程得到的预测值之间的关系强度，即多重相关系数 R 等于因变量的观测值与估计值之间的简单相关系数即

$R = \sqrt{R^2} = r_{y\hat{y}}$ (一元相关系数 r 也是如此，即 $r_{xy} = r_{y\hat{y}}$ 。读者自己去验证)

估计标准误差 S_e

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

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

10.2.2 显著性检验

导入：中华传统文化成语故事：滥竽充数

出处：

《韩非子·内储说上》:“齐宣王使人吹竽，必三百人。南郭处士请为王吹竽，宣王说

之，廪食以数百人。宣王死，愍王立。好一一听之，处士逃。”

寓意：

滥竽充数的故事告诉人们：（1）弄虚作假是经不住时间的考验，终究会露出马脚的，一个人如果像不会吹竽的南郭先生那样，没有真本事，只靠装样子糊弄人，在别人还不了解真相的时候，能够蒙混一阵子，但是总有真相大白的一天；（2）讽刺那些虚荣，对人不加以鉴别而重用的掌权者。

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：合奏效果（不找南郭先生）

线性关系检验可借助滥竽充数这个成语故事理解：

（1）线性关系显著相当于合奏乐队只要有一个乐师会吹奏，其他人即便不会吹奏，只要装模做样就行，也能骗过齐宣王。

（2）线性关系不显著相当于合奏乐队没有一个乐师会吹奏（这种情形概率非常低）

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

回归系数检验：独奏效果（找出南郭先生）

回归系数检验亦可借助滥竽充数这个成语故事理解：

（1）回归系数检验显著相当于齐愍王让乐师一个一个独奏，会吹奏的通过。

（2）回归系数检验不显著相当于不会吹奏的在一个一个独奏的过程中，自然暴露出来。

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

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

其中

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

10.3 多重共线性及其处理

10.3.1 多重共线性及其识别

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

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

- 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 $\mathbf{1-R_i^2}$。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。显然，VIF 越大多重共线性就越严重。一般要求 VIF 小于 5，也可放宽到小于 10。如果大于 10 则认为存在严重的多重共线性。

相关矩阵及其检验(SPSS)

相关性

		贷款余额	累计应收贷款	贷款项目个数	固定资产投资
贷款余额	Pearson 相关性	1	.679**	.848**	.780**
	显著性（双侧）		.000	.000	.000
	N	25	25	25	25
累计应收贷款	Pearson 相关性	.679**	1	.586**	.472*
	显著性（双侧）	.000		.002	.017
	N	25	25	25	25
贷款项目个数	Pearson 相关性	.848**	.586**	1	.747**
	显著性（双侧）	.000	.002		.000
	N	25	25	25	25
固定资产投资	Pearson 相关性	.780**	.472*	.747**	1
	显著性（双侧）	.000	.017	.000	
	N	25	25	25	25

**. 在 .01 水平（双侧）上显著相关。
*. 在 0.05 水平（双侧）上显著相关。

多重共线性的处理

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

奥克姆剃刀(Occam's Razor)

1. 模型选择可遵循奥克姆剃刀的基本原理
   - 最好的科学模型往往最简单，且能解释所观察到的实事
2. 对于线性模型来说，奥克姆剃刀可表示成简约原则
   - 一个模型应包括拟合数据所必需的最少变量
3. 如果一个模型只包含数据拟合所必需的变量，这个模型就称为简约模型(parsimonious model)
   - 实际中的许多多元回归模型都是对简约模型的扩展

10.3.2　变量选择与逐步回归

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(P 值最小的)，并将其首先引入模型
3. 分别拟合引入模型外的 *k*-1 个自变量的二元线性回归模型
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(这些模型中在每一个都有 *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(这些模型中每一个都有 *k*-2 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
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

参数的最小二乘估计(逐步回归)

【例 10-4】根据例 10-1 的数据，用逐步回归方法建立不良贷款 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，并求出不良贷款的置信区间和预测区间

逐步回归 (例题分析—SPSS 输出结果)

变量的进入和移出标准

输入／移去的变量[a]

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

两个模型的主要统计量

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), 贷款余额。
b. 预测变量: (常量), 贷款余额, 固定资产投资。

两个模型的方差分析表

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), 贷款余额。
b. 预测变量: (常量), 贷款余额, 固定资产投资。
c. 因变量: 不良贷款

两个模型的参数估计和检验

系数[a]

模型		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准 误差	试用版			容差	VIF
1	(常量)	-.830	.723		-1.147	.263		
	贷款余额	.038	.005	.844	7.534	.000	1.000	1.000
2	(常量)	-.443	.697		-.636	.531		
	贷款余额	.050	.007	1.120	6.732	.000	.392	2.551
	固定资产投资	-.032	.015	-.355	-2.133	.044	.392	2.551

a. 因变量: 不良贷款

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

多重共线性、自相关与异方差现象处理中的次第问题（主次矛盾）

多重共线性的影响：

（1）完全共线性下参数估计量不存在

（2）近似共线性会导致参数估计量的方差随着解释变量共线性的增加而增加；参数估计量的置信区间随着解释变量共线性的增加而变大；解释变量存在严重共线性时，t 检验失效，预测精度降低；参数估计量的经济含义不合理，回归模型缺乏稳定性。

自相关性的影响：

（1）参数估计值不具有最优性（仍具有无偏性，但不再具有最小方差性）

（2）低估随机误差项的方差

（3）模型的统计检验失效

（4）区间估计和预测区间的精度降低

异方差性的影响：

（1）不影响模型参数最小二乘估计值的无偏性

（2）模型参数最小二乘估计量不是一个有效的估计量

（3）t 检验来判断解释变量影响的显著性将失去意义

（4）直接影响回归模型的估计、检验和应用

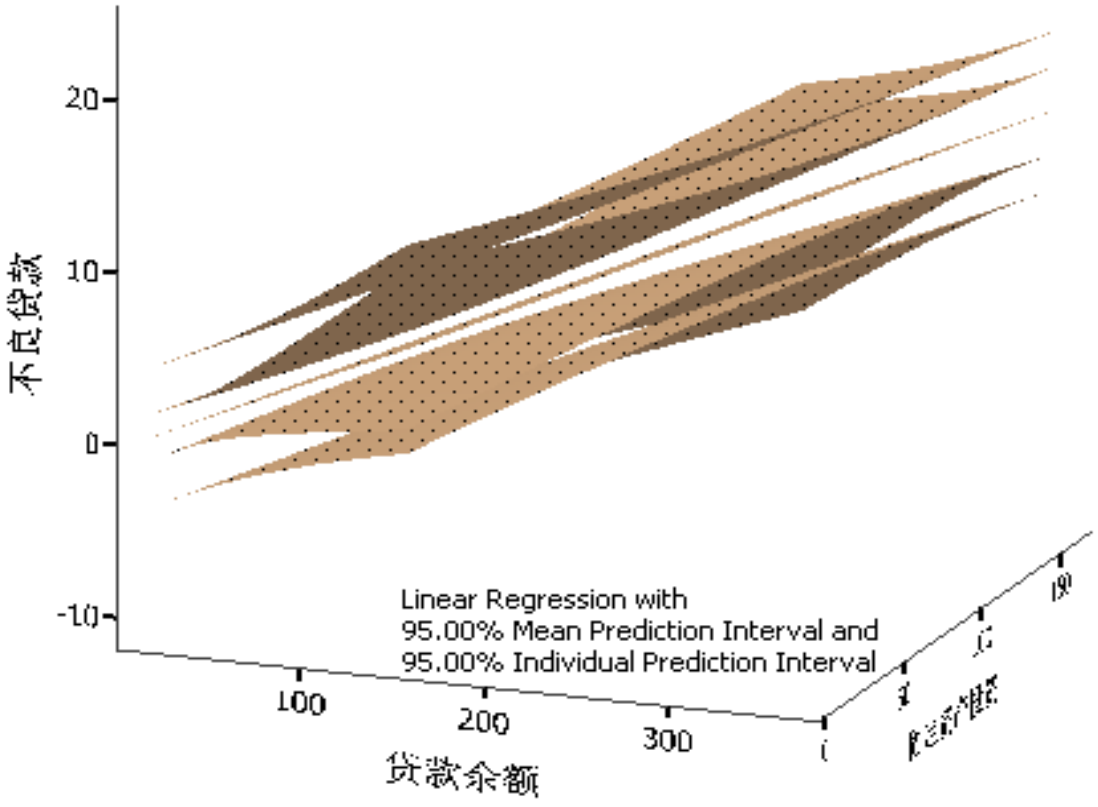
10.4 利用回归方程进行预测

置信区间和预测区间(例题分析)

	分行编号	不良贷款	PRE_1	RES_1	LMCI_1	UMCI_1	LICI_1	UICI_1
1	1	.9	1.288	-.388	.275	2.301	-2.665	5.242
2	2	1.1	2.259	-1.159	.918	3.599	-1.792	6.309
3	3	4.8	5.913	-1.113	4.975	6.850	1.978	9.848
4	4	3.2	3.161	.039	1.918	4.404	-.858	7.180
5	5	7.8	7.592	.208	6.225	8.958	3.533	11.650
6	6	2.7	.302	2.398	-1.061	1.665	-3.756	4.359
7	7	1.6	4.318	-2.718	3.015	5.620	.280	8.355
8	8	12.5	7.491	5.009	5.888	9.093	3.347	11.635
9	9	1.0	2.610	-1.610	1.787	3.433	-1.299	6.519
10	10	2.6	1.169	1.431	.011	2.328	-2.824	5.163
11	11	.3	1.426	-1.126	.471	2.380	-2.513	5.365
12	12	4.0	3.763	.237	2.891	4.636	-.157	7.684
13	13	.8	1.779	-.979	.737	2.820	-2.182	5.740
14	14	3.5	4.609	-1.109	3.134	6.083	.512	8.705
15	15	10.2	8.139	2.061	6.269	10.008	3.884	12.393
16	16	3.0	2.594	.406	1.641	3.547	-1.345	6.533
17	17	.2	-1.042	1.242	-2.524	.441	-5.141	3.058
18	18	.4	2.449	-2.049	1.439	3.459	-1.504	6.402
19	19	1.0	.372	.628	-.857	1.602	-3.642	4.387
20	20	6.8	4.521	2.279	3.732	5.311	.619	8.424
21	21	11.6	12.860	-1.260	10.321	15.398	8.272	17.448
22	22	1.6	2.954	-1.354	2.139	3.769	-.954	6.861
23	23	1.2	2.907	-1.707	2.035	3.778	-1.013	6.827
24	24	7.2	8.165	-.965	6.314	10.016	3.919	12.412
25	25	3.2	1.603	1.597	-.002	3.208	-2.542	5.748

置信区间和预测区间(例题分析)

不良贷款的置信面和预测面



10.5 哑变量回归

10.5.1 在模型中引进哑变量

10.5.2 含有一个哑变量的回归

哑变量(dummy variable)

1. 也称虚拟变量。用数字代码表示的定性自变量
2. 哑变量可有不同的水平
   - 只有两个水平的哑变量
     - 比如，性别(男，女)
   - 有两个以上水平的哑变量
     - 贷款企业的类型(家电，医药，其他)
3. 哑变量的取值为 0，1

$$x=\begin{cases}1 & \text{男}\\ 0 & \text{女}\end{cases}$$

在回归中引进哑变量

1. 回归模型中使用哑变量时，称为哑变量回归
2. 当定性变量只有两个水平时，可在回归中引入一个哑变量
   - 比如，性别(男，女)
3. 一般而言，如果定性自变量有 *k* 个水平，需要在回归中模型中引进 *k*-1 个哑变量

$$x_1=\begin{cases}1 & \text{水平}1\\ 0 & \text{其他水平}\end{cases},x_2=\begin{cases}1 & \text{水平}2\\ 0 & \text{其他水平}\end{cases},\cdots,x_{k-1}=\begin{cases}1 & \text{水平}k-1\\ 0 & \text{其他水平}\end{cases}$$

例题分析

【例 10-6】为研究考试成绩与性别之间的关系，从某大学商学院随机抽取男女学生各 8 名，得到他们的市场营销学课程的考试成绩如下表

	A	B	C
1	考试成绩 *y*	性别	*x*
2	75	男	0
3	96	女	1
4	68	男	0
5	51	男	0
6	78	女	1
7	81	女	1
8	72	男	0
9	69	男	0
10	88	女	1
11	93	男	0
12	62	男	0
13	76	女	1
14	45	男	0
15	75	女	1
16	65	女	1
17	95	女	1

10.5.2 含有一个哑变量的回归（例题分析）

	A	B	C
1	考试成绩 y	性别	x
2	75	男	0
3	96	女	1
4	68	男	0
5	51	男	0
6	78	女	1
7	81	女	1
8	72	男	0
9	69	男	0
10	88	女	1
11	93	男	0
12	62	男	0
13	76	女	1
14	45	男	0
15	75	女	1
16	65	女	1
17	95	女	1

- 引进哑变量时，回归方程表示为 $E(y)=\beta_0+\beta_1 x$
  - 男(x=0)：$E(y)=\beta_0$—男学生考试成绩的期望值
  - 女(x=1)：$E(y)=\beta_0+\beta_1$—女学生考试成绩的期望值
- 注意：当指定哑变量 0，1 时
  - β0 总是代表与哑变量值 0 所对应的那个分类变量水平的平均值
  - β_1 总是代表与哑变量值 1 所对应的那个分类变量水平的平均响应与哑变量值 0 所对应的那个分类变量水平的平均值的差值，即

平均值的差值 $=(\beta_0+\beta_1)-\beta_0=\beta_1$

考试成绩与性别的回归

	A	B	C	D	E
1					
2		Coefficients	标准误差	t Stat	P-value
3	Intercept	66.875	4.558	14.67	6.81E-10
4	X Variable 1	14.875	6.445	2.308	3.68E-02

主体间效应的检验

因变量:考试成绩

源	III 型平方和	df	均方	F	Sig.
校正模型	885.063[a]	1	885.063	5.326	.037
截距	88357.563	1	88357.563	531.731	.000
性别	885.063	1	885.063	5.326	.037
误差	2326.375	14	166.170		
总计	91569.000	16			
校正的总计	3211.438	15			

a. R 方 = .276（调整 R 方 = .224）

男=1，女=0。

女学生考试成绩的期望值=81.75 分；男学生比女学生平均低 14.875 分

例题分析

【例 10-8】为研究工资水平与工作年限和性别之间的关系，在某行业中随机抽取 10 名职工，所得数据如下表

	A	B	C	D
1	月工资收入（元） y	工作年限（年） x_1	性别 x_2	x_2
2	2900	2	男	1
3	3000	6	女	0
4	4800	8	男	1
5	1800	3	女	0
6	2900	2	男	1
7	4900	7	男	1
8	4200	9	女	0
9	4800	8	女	0
10	4400	4	男	1
11	4500	6	男	1

Excel 输出的结果

	A	B	C	D	E	F	G
1	回归统计						
2	Multiple R	0.7309					
3	R Square	0.5342					
4	Adjusted R Square	0.4759					
5	标准误差	781.0223					
6	观测值	10					
7							
8	方差分析						
9		df	SS	MS	F	Significance F	
10	回归	1	5596033.06	5596033.06	9.1739	0.0163	
11	残差	8	4879966.94	609995.87			
12	总计	9	10476000				
13							
14		Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
15	Intercept	2147.2727	604.9773	3.5493	0.0075	752.1917	3542.3538
16	X Variable 1	304.1322	100.4120	3.0288	0.0163	72.5815	535.6829

主体间效应的检验

因变量:月工资收入

源	III 型平方和	df	均方	F	Sig.
校正模型	9000922.601	2	4500461.300	21.357	.001
截距	3481514.702	1	3481514.702	16.522	.005
工作年限	8088255.934	1	8088255.934	38.383	.000
性别	3404889.543	1	3404889.543	16.158	.005
误差	1475077.399	7	210725.343		
总计	1.564E8	10			
校正的总计	10476000.00	9			

a. R 方 = .859（调整 R 方 = .819）

参数估计

因变量:月工资收入

参数	B	标准 误差	t	Sig.	95% 置信区间	
					下限	上限
截距	930.495	466.974	1.993	.087	-173.723	2034.714
工作年限	387.616	62.565	6.195	.000	239.673	535.559
[性别=男]	1262.693	314.127	4.020	.005	519.902	2005.485
[性别=女]	0[a]	.	.	.	.	.

a. 此参数为冗余参数，将被设为零。

用工作年限和性别预测的月工资水平及其残差

	月工资收入	工作年限	性别	PRE_1	RES_1
1	2900	2	男	2968.42	-68.42
2	3000	6	女	3256.19	-256.19
3	4800	8	男	5294.12	-494.12
4	1800	3	女	2093.34	-293.34
5	2900	2	男	2968.42	-68.42
6	4900	7	男	4906.50	-6.50
7	4200	9	女	4419.04	-219.04
8	4800	8	女	4031.42	768.58
9	4400	4	男	3743.65	656.35
10	4500	6	男	4518.89	-18.89

- 引进哑变量时，回归方程写为

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

  - 女(x_2=0)：$E(y|\text{女性}) = \beta_0 + \beta_1 x_1$
  - 男(x_2=1)：$E(y|\text{男性}) = (\beta_0 + \beta_2) + \beta_1 x_1$
- β_0 的含义表示：女性职工的期望月工资收入
- $(\beta_0 + \beta_2)$的含义表示：男性职工的期望月工资收入
- β_1 含义表示：工作年限每增加 1 年，男性或女性工资的平均增加值
- β_2 含义表示：男性职工的期望月工资收入与女性职工的期望月工资收入之间的差值 $(\beta_0 + \beta_2) - \beta_0 = \beta_2$

主体间效应的检验

因变量:月工资收入

源	III 型平方和	df	均方	F	Sig.
校正模型	9000922.601	2	4500461.300	21.357	.001
截距	3481514.702	1	3481514.702	16.522	.005
工作年限	8088255.934	1	8088255.934	38.383	.000
性别	3404889.543	1	3404889.543	16.158	.005
误差	1475077.399	7	210725.343		
总计	1.564E8	10			
校正的总计	10476000.00	9			

a. R 方 = .859（调整 R 方 = .819）

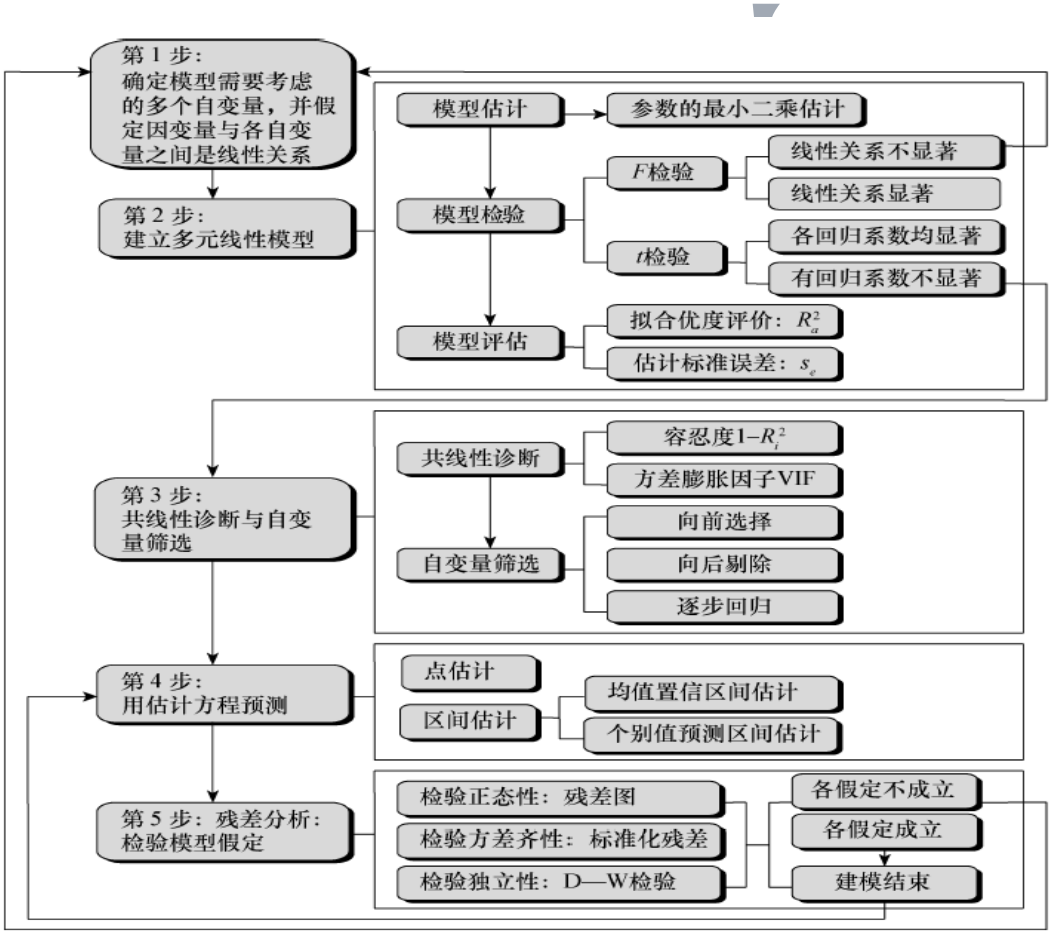
参数估计

因变量:月工资收入

参数	B	标准 误差	t	Sig.	95% 置信区间	
					下限	上限
截距	930.495	466.974	1.993	.087	-173.723	2034.714
工作年限	387.616	62.565	6.195	.000	239.673	535.559
[性别=男]	1262.693	314.127	4.020	.005	519.902	2005.485
[性别=女]	0[a]	.	.	.	.	.

a. 此参数为冗余参数，将被设为零。

本章图解



本章小结

- 多元线性回归模型、回归方程与估计的回归方程
- 回归方程的拟合优度与显著性检验
- 多重共线性问题及其处理
- 利用回归方程进行预测
- 哑变量回归
- 用 **Excel** 和 **SPSS** 进行回归分析
-

课后练习:

教材课后练习　**10-1、10-2、10-3、10-4、10-6、10-8、10-9**