线性回归

本章主要内容

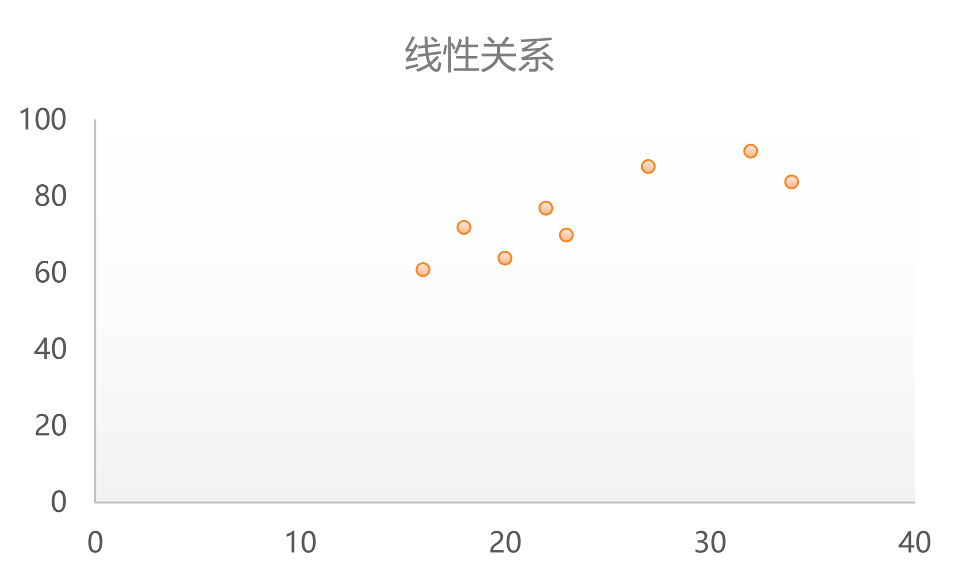
- 一元线性回归
- 多元线性回归
- 多重共线性问题
- 用SPSS做回归分析

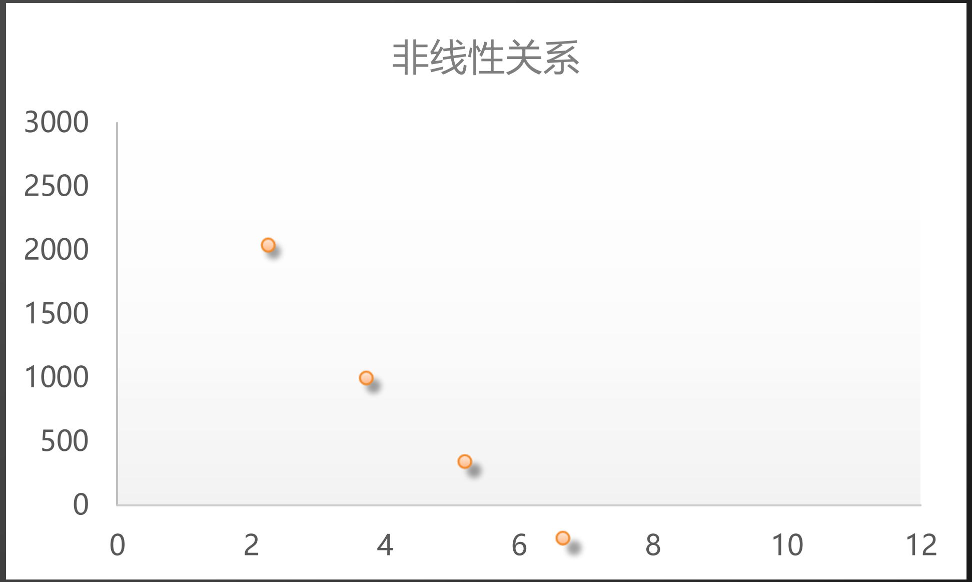
本章主要内容

- 一元线性回归
- 多元线性回归
- 多重共线性问题
- 用SPSS做回归分析

线性关系

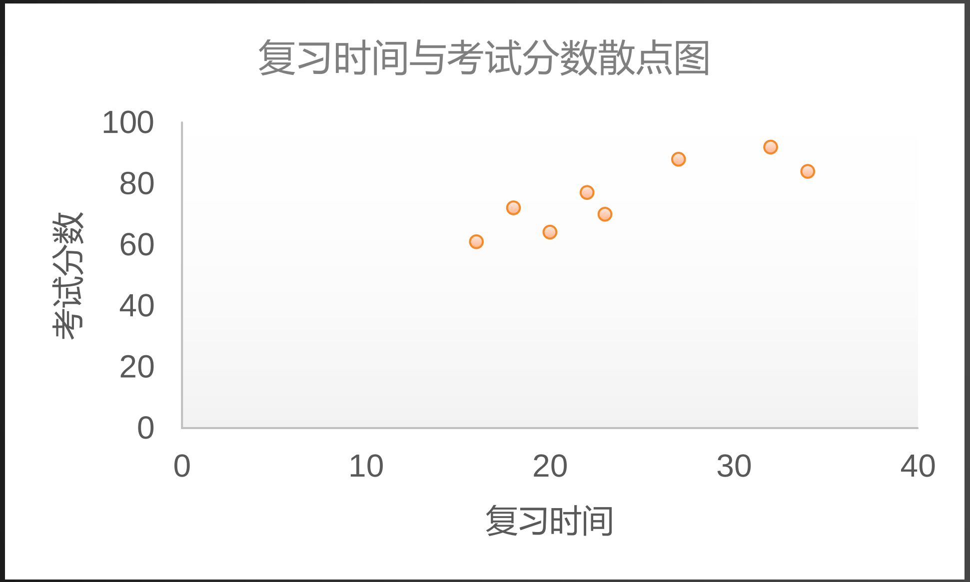
- 线性回归，顾名思义，就是线性+回归，线性指线性关系，回归是指建立拟合曲线。
- 线性回归就是为具有线性关系的两个变量建立一条拟合曲线（直线）。
- 那什么是线性关系呢？看下图。

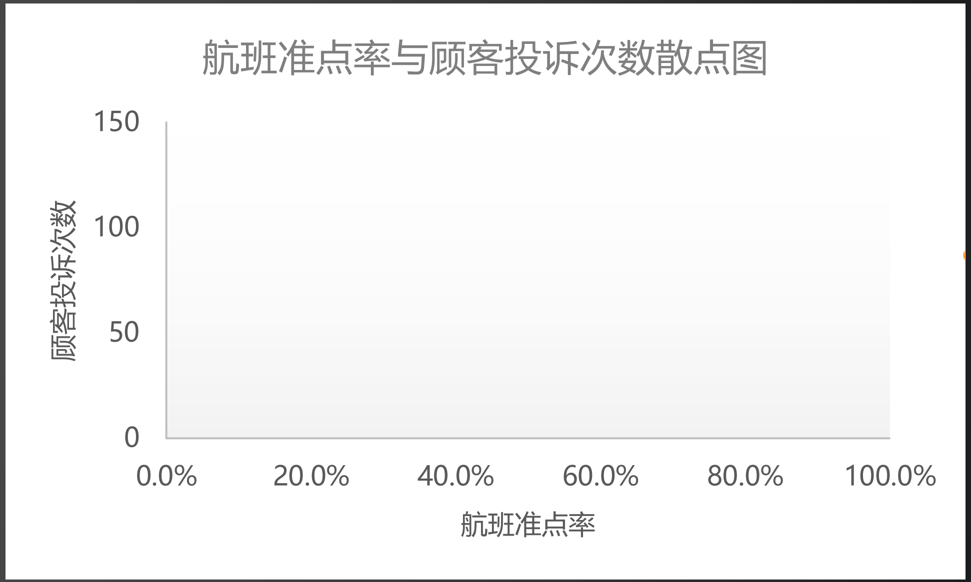




线性关系

- 线性关系又分为正相关和负相关。





复习时间与考试成绩有线性关系，呈现出正相关。

航班准点率与顾客投诉次数有线性关系，呈现出负相关。

相关系数

散点图只能够大致反映出两个变量之间的线性关系，如果需要准确度量两个变量之间的线性关系的强弱，就需要一个统计量，即相关系数。

相关系数是根据样本计算出来的度量两个变量之间线性关系强度的统计量，计算公式为：

$$
\begin{aligned}
\rho_{XY} &= \frac{E\{[X-E(X)][Y-E(Y)]\}}{\sqrt{D(X)}\sqrt{D(Y)}} \\
&= \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-[E(X)]^2}\sqrt{E(Y^2)-[E(Y)]^2}} \\
&= \frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-(\sum x)^2}\sqrt{n\sum y^2-(\sum y)^2}}
\end{aligned}
$$

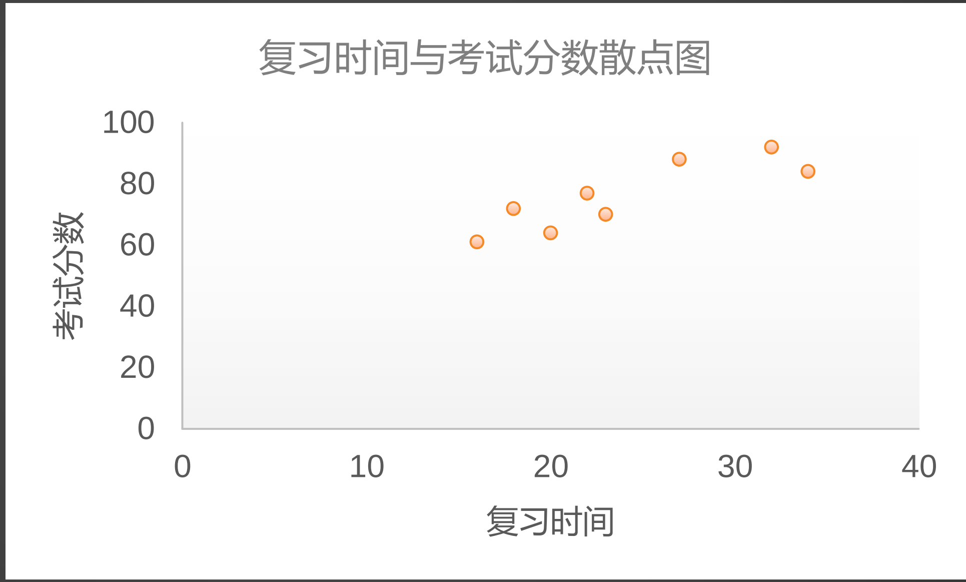
实际应用中，一般通过Excel公式计算相关系数，公式为：

- CORREL
- PEARSON

相关系数

- 例如，想知道学生的复习时间（小时）和考试分数（分）之间是否有关系，抽取了由8名学生组成的一个样本，数据如下。

复习时间	考试分数
20	64
16	61
34	84
23	70
27	88
32	92
18	72
22	77



- 通过Excel公式计算出相关系数为：0.86

相关系数矩阵

- 如果同时求多个变量之间的相关系数，得到的就是相关系数矩阵。
- 例如，有以下关于银行不良贷款等数据，通过Excel可以得到相关系数矩阵。

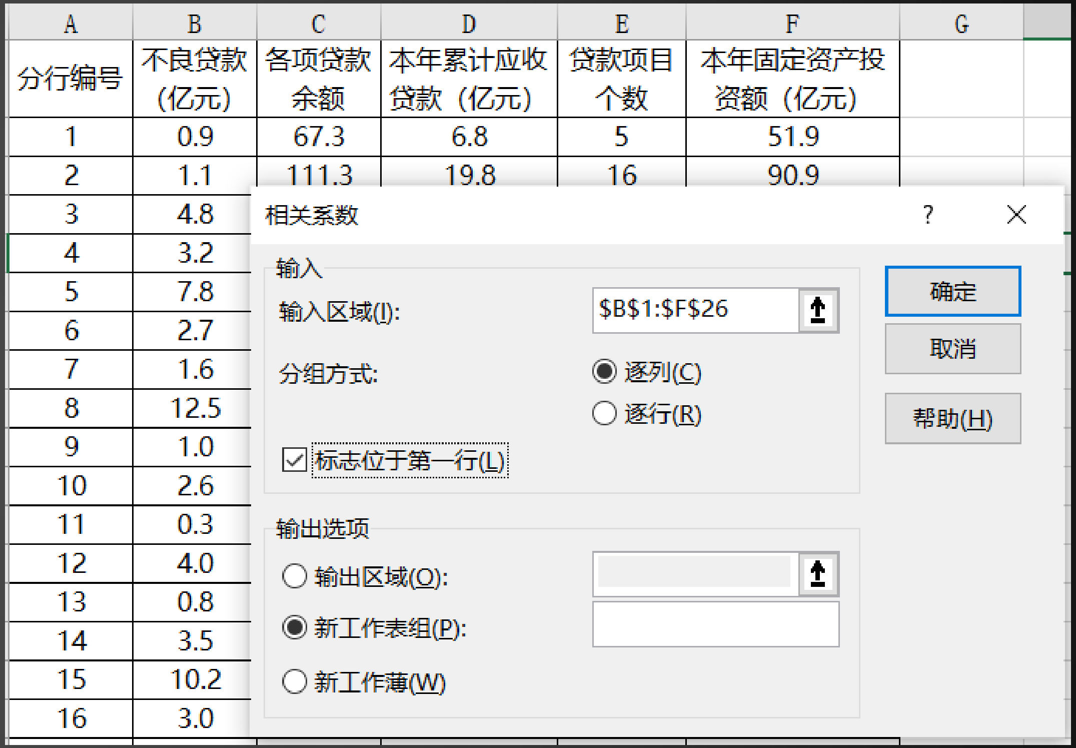
分行编号	不良贷款（亿元）	各项贷款余额	本年累计应收贷款（亿元）	贷款项目个数	本年固定资产投资额（亿元）
1	0.9	67.3	6.8	5	51.9
2	1.1	111.3	19.8	16	90.9
3	4.8	173.0	7.7	17	73.7
4	3.2	80.8	7.2	10	14.5
5	7.8	199.7	16.5	19	63.2
6	2.7	16.2	2.2	1	2.2
7	1.6	107.4	10.7	17	20.2
8	12.5	185.4	27.1	18	43.8
9	1.0	96.1	1.7	10	55.9
10	2.6	72.8	9.1	14	64.3

说明：这里只显示了前10条记录。

相关系数矩阵

用Excel求相关系数矩阵步骤如下：

1. 菜单【数据】-【数据分析】，选择【相关系数】
2. 在弹出的“相关系数”对话框中，选择：
   - 【输入区域】：选择数据区域
   - 勾选【标志位于第一行】
   - 输出选项：默认“新工作表组”



	不良贷款（亿元）	各项贷款余额	本年累计应收贷款（亿元）	贷款项目个数	本年固定资产投资额（亿元）
不良贷款（亿元）	1				
各项贷款余额	0.843571364	1			
本年累计应收贷款（亿元）	0.731505008	0.678771764	1		
贷款项目个数	0.700281491	0.848416404	0.58583149	1	
本年固定资产投资额（亿元）	0.51851809	0.779702158	0.47243096	0.746645845	1

线性回归

说完了线性之后，再说说回归，回归就是为两个具有线性关系的变量找到一条拟合直线。

例如，前面提到的复习时间与考试分数，散点图如下。

利用Excel可以直接获得这个拟合直线，还可以得到回归方程及判定系数R^2。

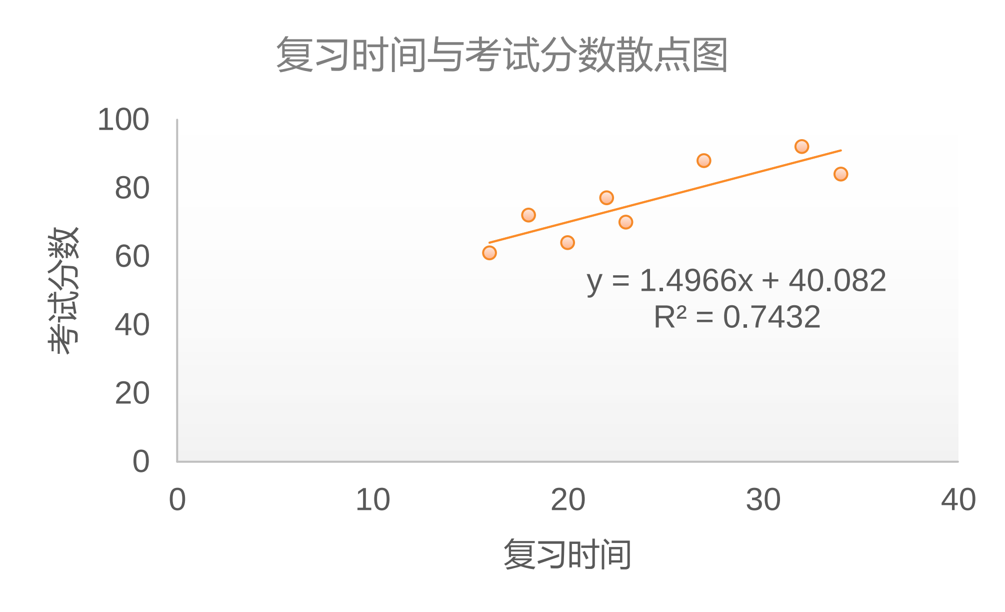
这里将复习时间作为自变量x，数值型；

考试分数作为因变量y，数值型。

x与y之间为线性关系，回归方程为：

$y = 1.4966x + 40.082$

判定系数：$R^2 = 0.7432$，表示拟合效果。



线性回归

线性回归的本质是研究数值型自变量和数值型因变量之间的线性关系。

- 自变量：用$x_1, x_2, \cdots, x_n$，一个或者多个，均为数值型。
- 因变量：y，一个，为数值型。

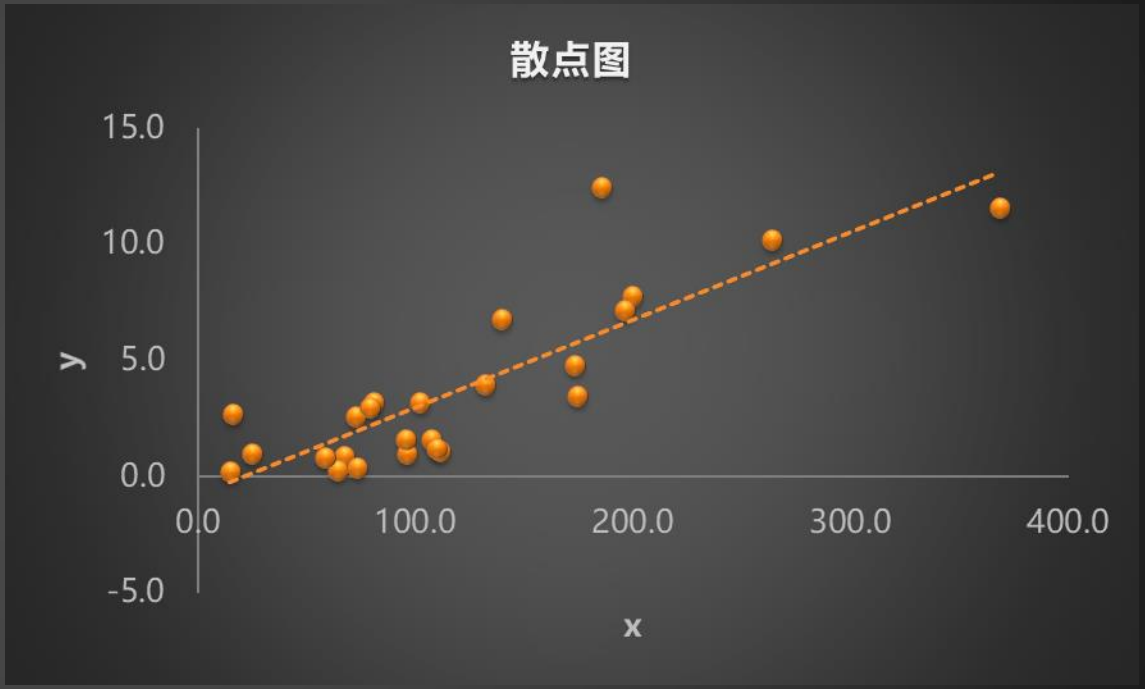
根据自变量X的个数，线性回归分为：

- 一元线性回归：一个自变量
- 多元线性回归：多个自变量

一元线性回归

- 特点：只有一个自变量x
- 回归方程：y= $wx + b$
- 一次函数：y=kx+b

- 任务目标：估计参数w, b



一元线性回归示意图

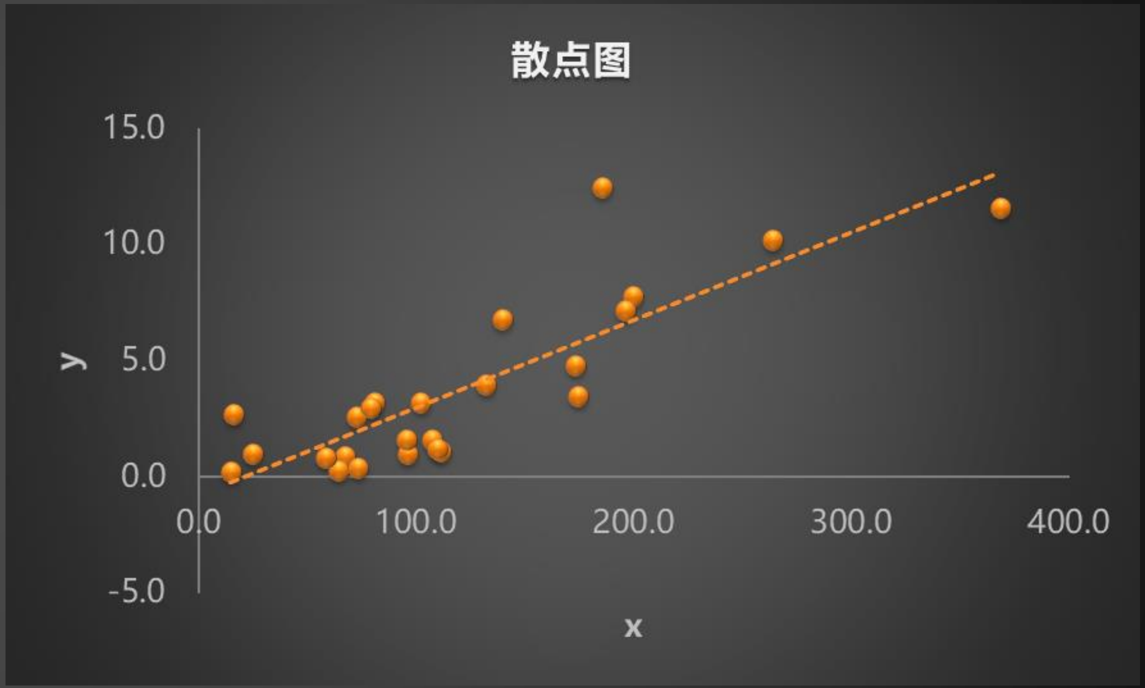
最小二乘法

假定有m个样本，$(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$

最小二乘法的基本思想是通过最小化这些点到直线的总误差来估计参数w和b。

根据最小二乘法，使以下这个式子最小。

$$\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{m}(y_i - wx_i - b)^2$$



这个式子采用的是误差的平方和，加上平方是为了避免正负相抵。

最小二乘法

目标函数：

$$l(w,b)=\sum_{i=1}^{m}(y_i-wx_i-b)^2$$

在给定样本数据后，l是w和b的函数。根据微积分的极值定理，对l求相应于w和b的偏导数，并令其等于0,即

$$\begin{cases}\frac{\partial l}{\partial w}=-2\sum_{i=1}^{m}x_i(y_i-wx_i-b)=0\\ \frac{\partial l}{\partial b}=-2\sum_{i=1}^{m}(y_i-wx_i-b)=0\end{cases}$$

解上述方程组得

$$\begin{cases}w=\frac{m\sum_{i=1}^{m}x_iy_i-\sum_{i=1}^{m}x_i\sum_{i=1}^{m}y_i}{m\sum_{i=1}^{m}x_i^2-(\sum_{i=1}^{m}x_i)^2}\\ b=\frac{1}{m}\sum_{i=1}^{m}(y_i-wx_i)\end{cases}$$

大家只要了解原理，一般借助于计算机来计算！！！

评估回归效果

将由回归方程计算出来的y值记为$\hat{y}$，对于每一个实际观测值y_i，其误差的大小用$y_i - \hat{y}_i$来表示，所有观测值的总误差可以通过每个观测值的误差的平方和来表示，即

$$\sum (y_i - \hat{y}_i)^2$$

以上式子可以称为误差平方和，或者残差平方和。
接着对误差做一下变形。

$$y_i - \hat{y}_i = y_i - \overline{y} - (\hat{y}_i - \overline{y})$$

移项得

$$y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})$$

上式本质上是误差分解，不妨将其称为**误差分解式。**

接着，将误差分解式两边平方，并求和，得

$$\sum(y_i-\overline{y})^2=\sum[(y_i-\hat{y}_i)+(\hat{y}_i-\overline{y})]^2$$

将右边平方展开，得

$$\sum(y_i-\overline{y})^2=\sum(y_i-\hat{y}_i)^2+\sum(\hat{y}_i-\overline{y})^2+2\sum(y_i-\hat{y}_i)(\hat{y}_i-\overline{y})$$

上式中，可以证明$\sum(y_i-\hat{y}_i)(\hat{y}_i-\overline{y})=0$

说明：证明时，需要将$\hat{y}_i=wx+b$代入上式中，前面已经求得参数w和b的值。

于是，有

$$\sum(y_i-\overline{y})^2=\sum(y_i-\hat{y}_i)^2+\sum(\hat{y}_i-\overline{y})^2$$

评估回归效果

$$\sum(y_i-\overline{y})^2=\sum(y_i-\hat{y}_i)^2+\sum(\hat{y}_i-\overline{y})^2$$

在上式中，左边称为**总平方和**，记为SST，右边第一项称为**误差平方和**，记为SSE，右边第二项称为**回归平方和**，记为SSR。三个平方和的关系为：

$$总平方和(SST)=残差平方和(SSE)+回归平方和(SSR)$$

我们的目标是希望**误差平方和**尽可能小。在总平方和一定的情况下，回归平方和越大，误差平方和就越小，所以可以借助于回归平方和占总平方和的比例来评估回归方程的好坏，我们将这个比例称之为**判定系数**，记为R^2，其表达式为：

$$R^2=\frac{SSR}{SST}=\frac{\sum(\hat{y}_i-\overline{y})^2}{\sum(y_i-\overline{y})^2}=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\overline{y})^2}$$

可以看出，判定系数是一个介于0到1之间的数，判定系数越接近于1，说明回归方程拟合效果越好。

案例：分析不良贷款形成的原因

一家大型商业银行在多个地区有25家分行，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。下面是该银行所属的25家分行2002年的有关业务数据。

分行编号	不良贷款（亿元）	各项贷款余额	本年累计应收贷款（亿元）	贷款项目个数	本年固定资产投资额（亿元）
1	0.9	67.3	6.8	5	51.9
2	1.1	111.3	19.8	16	90.9
3	4.8	173.0	7.7	17	73.7
4	3.2	80.8	7.2	10	14.5
5	7.8	199.7	16.5	19	63.2
6	2.7	16.2	2.2	1	2.2
7	1.6	107.4	10.7	17	20.2
8	12.5	185.4	27.1	18	43.8
9	1.0	96.1	1.7	10	55.9
10	2.6	72.8	9.1	14	64.3

案例：分析不良贷款形成的原因

任务目标：

将不良贷款作为因变量y，从其余变量中选择一个作为自变量x，然后做一元线性回归

实现步骤：

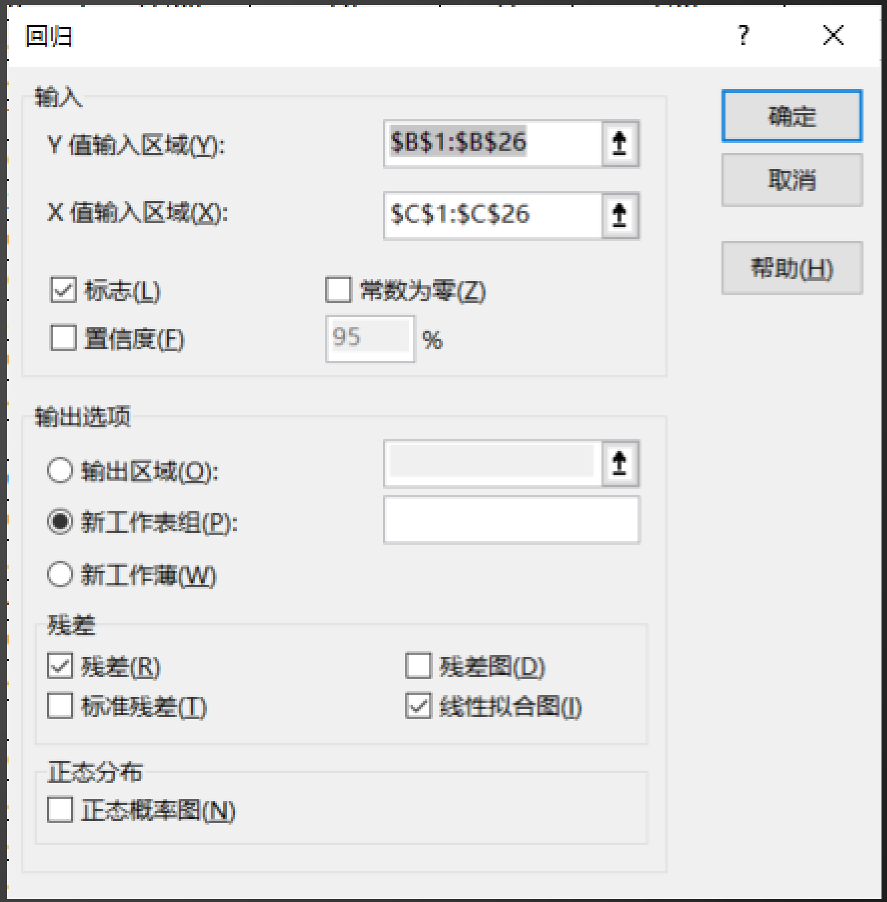
1. 求出变量间的相关系数矩阵，利用相关系数矩阵确定选用哪个变量作为自变量x
2. 根据确定的自变量x和因变量y，做一元线性回归
3. 利用判定系数R^2评估整体回归效果
4. 回归分析的显著性检验：线性关系检验、回归系数的检验（p值，越小越好）

一元线性回归：Excel

主要步骤：

1. 菜单【数据】-【数据分析】，选择【回归】
2. 在弹出的【回归】对话框中，选择Y和X
3. 勾选【残差】、【线性拟合图】

单击【确定】按钮后，Excel自动给出分析结果。



一元线性回归：Excel

SUMMARY OUTPUT					
回归统计					
Multiple R	0.843571364				
R Square	0.711612647	判定系数			
Adjusted R Square	0.699074066				
标准误差	1.979947533	越小越好			
观测值	25				
方差分析					
	df	SS	MS	F	Significance F
回归分析	1	222.4859787	222.4859787	56.75384406	1.18349E-07
残差	23	90.16442134	3.920192232		
总计	24	312.6504			
	Coefficients	标准误差	t Stat	P-value	Lower 95%
Intercept	-0.829520617	0.723043295	-1.147262719	0.263067597	-2.325249632
各项贷款余额	0.037894707	0.00503015	7.533514722	1.18349E-07	0.027489049

线性关系的显著性：说明自变量和因变量之间存在着显著的线性关系

回归方程：$y = -0.83 + 0.038x$

回归系数的显著性：说明自变量x对因变量的影响时是显著的

本章主要内容

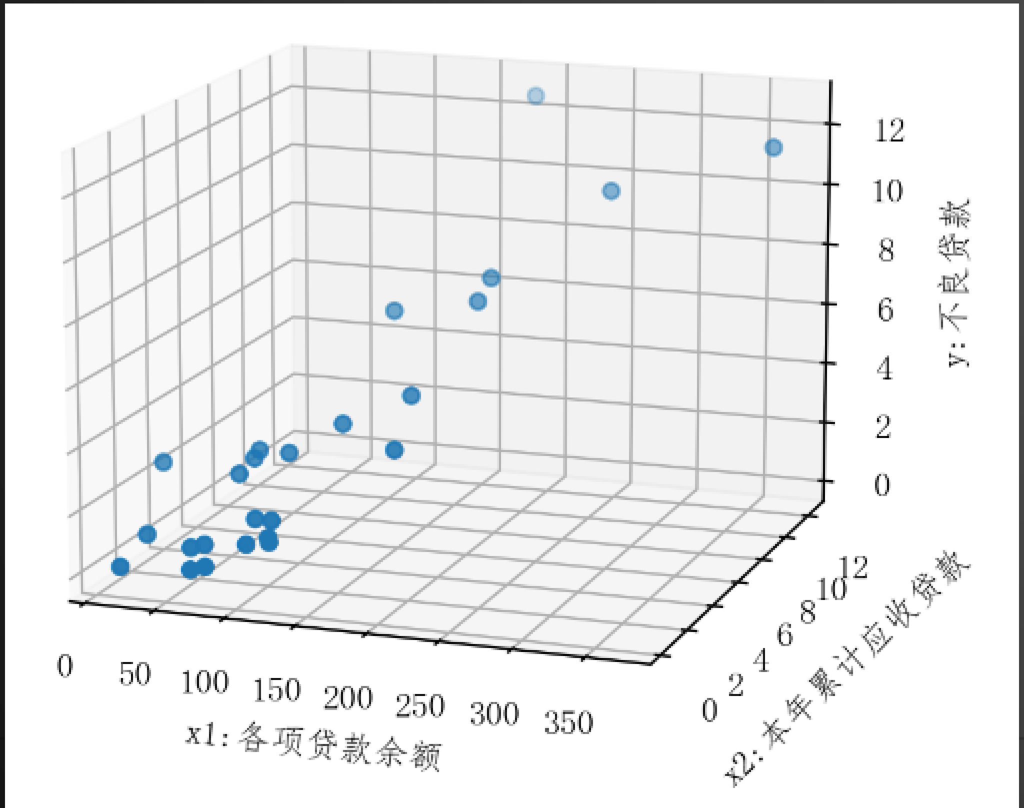
- 一元线性回归
- 多元线性回归
- 多重共线性问题
- 用SPSS做回归分析

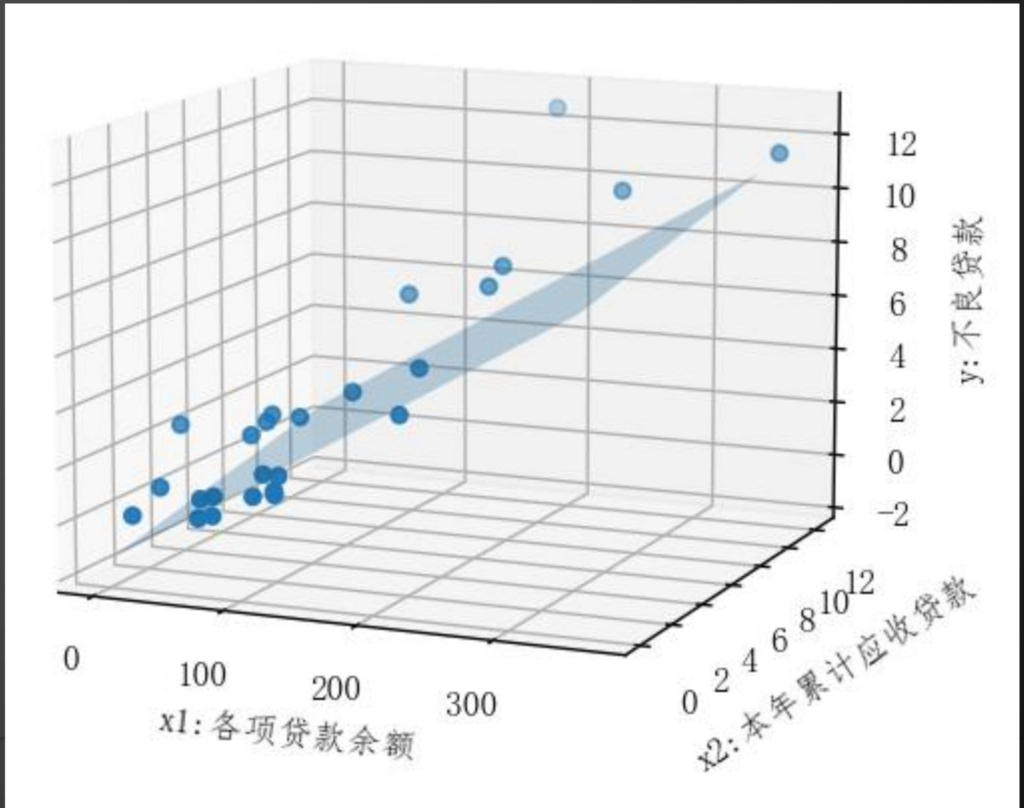
多元线性回归

- 特点：涉及多个自变量，$x_1, x_2, \cdots, x_n$
- 回归方程：$y = w_1 x_1 + w_2 x_2$+...+ $w_n x_n$+b
- 任务目标：估计参数$w_1, w_2, \cdots, w_n, \mathrm{b}$

二元线性回归示意图

- 假设有两个自变量，x_1, x_2，加上一个因变量y，可以绘制一个三维散点图。
- 此时，回归方程：$y = w_1x_1 + w_2x_2 + b$，是一个平面，用这个平面去拟合这些样本点。





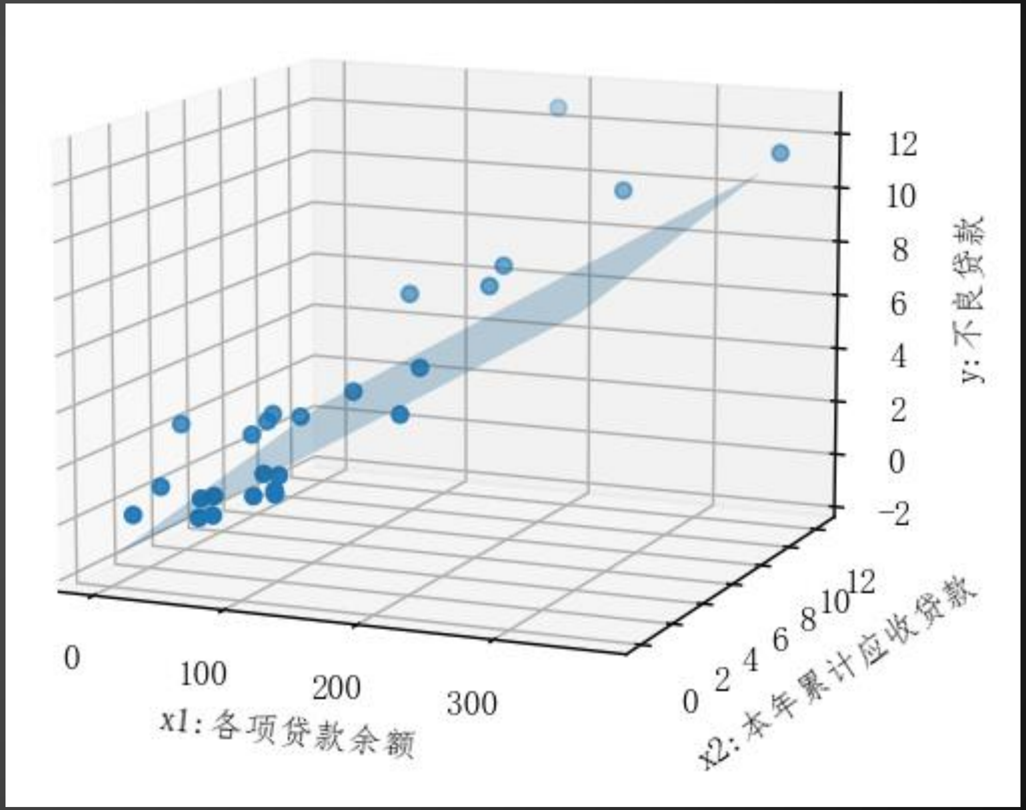
二元线性回归的最小二乘法

假定有m个样本，$\left(x_1^{(1)}, x_2^{(1)}, y_1\right), \left(x_1^{(2)}, x_2^{(2)}, y_2\right), \cdots, \left(x_1^{(m)}, x_2^{(m)}, y_m\right)$

最小二乘法的基本思想是通过最小化这些点到拟合平面的总误差来估计参数w_1, w_2, b

根据最小二乘法，使以下这个式子最小。

$$\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{m}\left(y_i - w_1 x_1^{(i)} - w_2 x_2^{(i)} - b\right)^2$$



二元线性回归的最小二乘法

假定有m个样本，$(x_1^{(1)}, x_2^{(1)}, y_1), (x_1^{(2)}, x_2^{(2)}, y_2), \cdots, (x_1^{(m)}, x_2^{(m)}, y_m)$。

目标函数：

$$l(w_1, w_2, b) = \sum_{i=1}^{m}(y_i - w_1 x_1^{(i)} - w_2 x_2^{(i)} - b)^2$$

注意：3个参数，m个样本。

根据微积分的极值定理，对l求相应于w_1, w_2和b的偏导数，并令其等于0，即

$$\begin{cases} \frac{\partial l}{\partial w_1} = -2\sum_{i=1}^{m} x_1^{(i)}(y_i - w_1 x_1^{(i)} - w_2 x_2^{(i)} - b) = 0 \\ \frac{\partial l}{\partial w_2} = -2\sum_{i=1}^{m} x_2^{(i)}(y_i - w_1 x_1^{(i)} - w_2 x_2^{(i)} - b) = 0 \\ \frac{\partial l}{\partial b} = -2\sum_{i=1}^{m}(y_i - w_1 x_1^{(i)} - w_2 x_2^{(i)} - b) = 0 \end{cases}$$

解上述方程组，就可以得到各参数w_1, w_2, b的值，多元回归分析也是类似的。

评估回归效果

与一元线性回归类似，在多元线性回归中，也是通过判定系数R^2来评估拟合效果，只是将其称为**多重判定系数。**

对于多元线性回归，也有这样一个平方和的关系式：

$$SST = SSR + SSE$$

其中，$SST = \sum(y_i - \bar{y})^2$为总平方和；

$SSR = \sum(\hat{y}_i - \bar{y})^2$为回归平方和；

$SSE = \sum(y_i - \hat{y}_i)^2$为残差平方和。

多重判定系数为：

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

多重判定系数也是一个介于0到1之间的数，判定系数越接近于1，说明回归方程拟合效果越好。

案例：分析不良贷款形成的原因

- 一家大型商业银行在多个地区有25家分行，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。下面是该银行所属的25家分行2002年的有关业务数据。

分行编号	不良贷款（亿元）	各项贷款余额	本年累计应收贷款（亿元）	贷款项目个数	本年固定资产投资额（亿元）
1	0.9	67.3	6.8	5	51.9
2	1.1	111.3	19.8	16	90.9
3	4.8	173.0	7.7	17	73.7
4	3.2	80.8	7.2	10	14.5
5	7.8	199.7	16.5	19	63.2
6	2.7	16.2	2.2	1	2.2
7	1.6	107.4	10.7	17	20.2
8	12.5	185.4	27.1	18	43.8
9	1.0	96.1	1.7	10	55.9
10	2.6	72.8	9.1	14	64.3

案例：分析不良贷款形成的原因

- 自变量有4个，包括x_1：各项贷款余额，x_2：本年累计应收贷款，x_3：贷款项目个数，x_4：本年固定资产投资额
- 因变量y：不良贷款

任务目标：根据选择的自变量和因变量做多元线性回归

实现步骤：

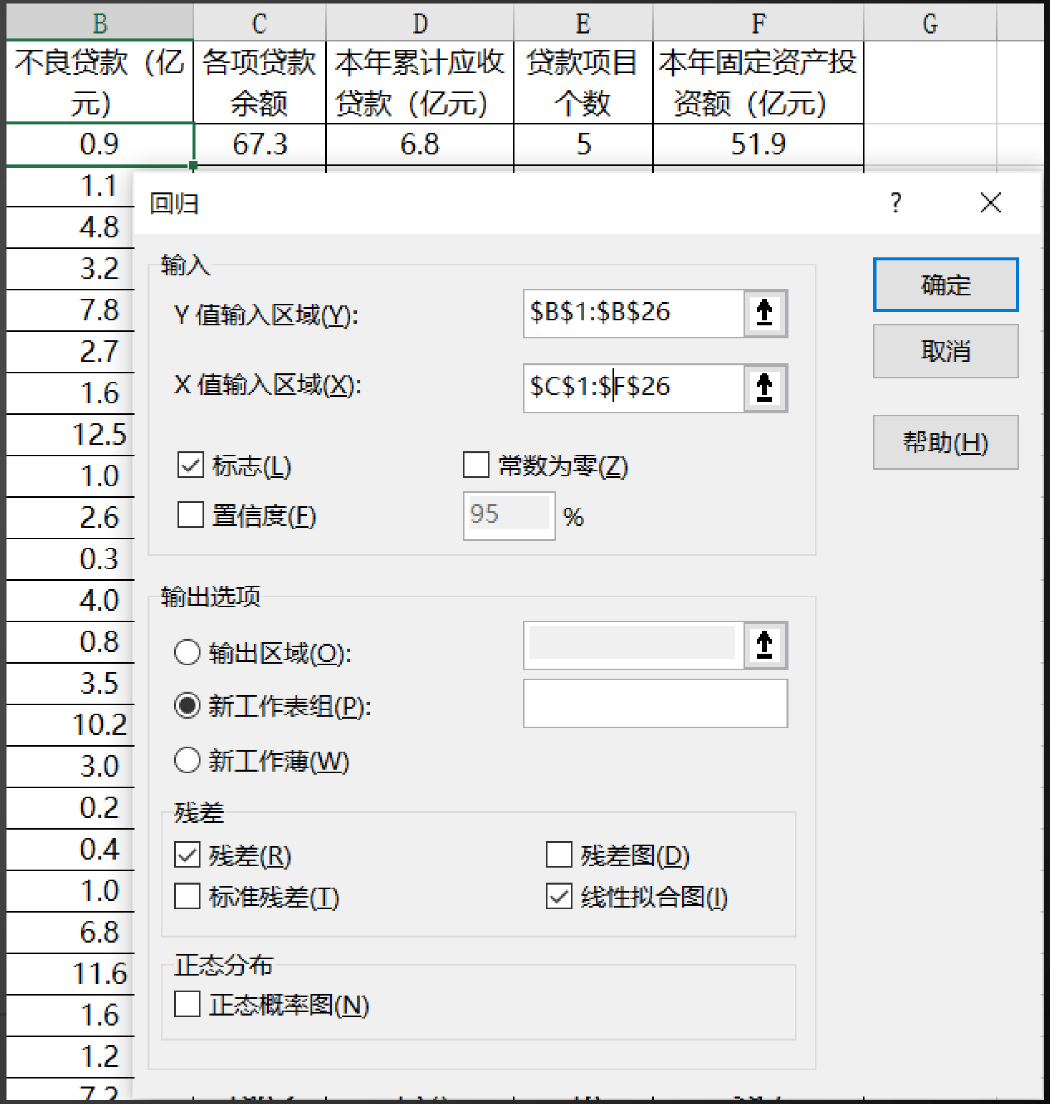
1. 根据确定的自变量x（多个自变量）和因变量y，做多元线性回归
2. 利用多重判定系数R^2评估整体回归效果
3. 回归分析的显著性检验：线性关系检验、回归系数的检验（p值，越小越好）

多元线性回归：Excel实现

Excel操作步骤：

1. 菜单【数据】-【数据分析】，选择【回归】
2. 在弹出的对话框中，选择Y和X（选择多个自变量）
3. 勾选【残差】、【线性拟合图】

单击【确定】按钮后，Excel自动给出分析结果



多元线性回归：Excel实现

A	B	C	D	E	F
SUMMARY OUTPUT					
回归统计					
Multiple R	0.893086776				
R Square	0.797603989	多重判定系数			
Adjusted R Square	0.757124787				
标准误差	1.778752284	越小越好			
观测值	25				
方差分析					
	df	SS	MS	F	Significance F
回归分析	4	249.3712062	62.34280156	19.7040442	1.03539E-06
残差	20	63.27919377	3.163959689		
总计	24	312.6504			
	Coefficients	标准误差	t Stat	P-value	Lower 95%
Intercept	-1.021639763	0.78237236	-1.305822925	0.206433969	-2.653639908
各项贷款余额	0.040039353	0.01043372	3.83749534	0.001028464	0.018274994
本年累计应收贷款（亿元）	0.148033891	0.078794333	1.878737798	0.074935421	-0.016328207
贷款项目个数	0.014529353	0.083033158	0.174982537	0.862852686	-0.158674781
本年固定资产投资额（亿元）	-0.029192866	0.015072973	-1.936768921	0.067030078	-0.060634537

线性关系的显著性：说明自变量和因变量之间存在着显著的线性关系

回归系数的显著性：只有一个自变量通过了显著性检验。

回归方程：$y = -1.02 + 0.04x_1 + 0.15x_2 + 0.01x_3 - 0.03x_4$

本章主要内容

- 一元线性回归
- 多元线性回归
- 多重共线性问题
- 用SPSS做回归分析

多重共线性问题

当回归模型中两个或两个以上的自变量彼此相关时，则称回归模型中存在多重共线性。

多重共线性会导致以下问题：

1. 可能会使回归的结果造成混乱，甚至会把分析引入歧途
2. 可能对参数估计值的正负号产生影响。

例如，前面提到的银行不良贷款问题中，多元回归分析的结果为：

	Coefficients	标准误差	t Stat	P-value
Intercept	-1.021639763	0.78237236	-1.305822925	0.206433969
各项贷款余额	0.040039353	0.01043372	3.83749534	0.001028464
本年累计应收贷款（亿元）	0.148033891	0.078794333	1.878737798	0.074935421
贷款项目个数	0.014529353	0.083033158	0.174982537	0.862852686
本年固定资产投资额（亿元）	-0.029192866	0.015072973	-1.936768921	0.067030078

问题：4个回归系数中，有3个回归系数未通过显著性检验。

多重共线性问题

例如，前面提到的银行不良贷款问题中，多元回归分析的结果为：

	Coefficients	标准误差	t Stat	P-value
Intercept	-1.021639763	0.78237236	-1.305822925	0.206433969
各项贷款余额	0.040039353	0.01043372	3.83749534	0.001028464
本年累计应收贷款（亿元）	0.148033891	0.078794333	1.878737798	0.074935421
贷款项目个数	0.014529353	0.083033158	0.174982537	0.862852686
本年固定资产投资额（亿元）	-0.029192866	0.015072973	-1.936768921	0.067030078

问题：本年固定资产投资额的系数为负值，说明随着本年固定资产投资额的增加，不良贷款是减少的。

实际上，通过对本年固定资产投资额（x）和不良贷款（y）做一元线性回归可知，随着本年固定资产投资额的增加，不良贷款是增加的，这个矛盾就是由多重共线性造成的。

说明：对本年固定资产投资额（x）和不良贷款（y）做一元线性回归，大家可以自行完成，这里不再演示。

多重共线性问题

多重共线性的判别：

1. 模型中各个自变量之间显著相关
2. 当模型的线性关系检验(F检验)显著时，但是几乎所有回归系数的t检验却不显著
3. 回归系数的正负号与预期的相反，例如，前面提到的不良贷款与本年固定资产投资额。
4. 容忍度与方差扩大因子（variance inflation factor，VIF）

- 一般来说，VIF越大，多重共线性越严重。
- 一般认为VIF大于10时，存在严重的多重共线性。

说明：用SPSS做回归分析时，可以给出VIF值。

多重共线性问题的处理

- 如果能在建立回归模型之前，对自变量进行一定的筛选，去掉那些不必要的自变量，不仅可以使模型变得容易，还能够使模型更有可操作性，也更容易解释。
- 解决多重共线性问题的核心：建模前，先选择变量，使进入模型的自变量尽可能不相关。
- 选择自变量的原则：将一个或一个以上的自变量引入回归模型中，看是否可以使残差平方和（SSE）显著减少。
- 回顾：根据最小二乘法建立回归模型，我们希望残差平方和尽可能小。

$$SSE=\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$$

多重共线性问题的处理

常用的选择变量的方法主要有：向前回归、向后回归、逐步回归、岭回归等。

- 向前选择：从零开始，不断增加自变量，使直至无法使SSR显著增加为止。
- 向后选择：从包含所有变量开始，不断减少自变量，直至无法使SSE显著减少为止。
- 逐步回归：将上述两种方法结合，进行自变量筛选。

- 由于Excel没有逐步回归的功能，所以需要借助于其他工具，如SPSS来完成逐步回归。

本章主要内容

- 一元线性回归
- 多元线性回归
- 多重共线性问题
- 用SPSS做回归分析

多元线性回归：SPSS实现

主要步骤：

1. 将数据粘贴到SPSS中，并调整各个字段的数据类型。
2. 选择菜单【分析】-【回归】-【线性】
3. 选择变量：X、Y，方法选择【步进】（逐步回归），勾选相关分析选项
4. 解释分析结果