5 相关与回归

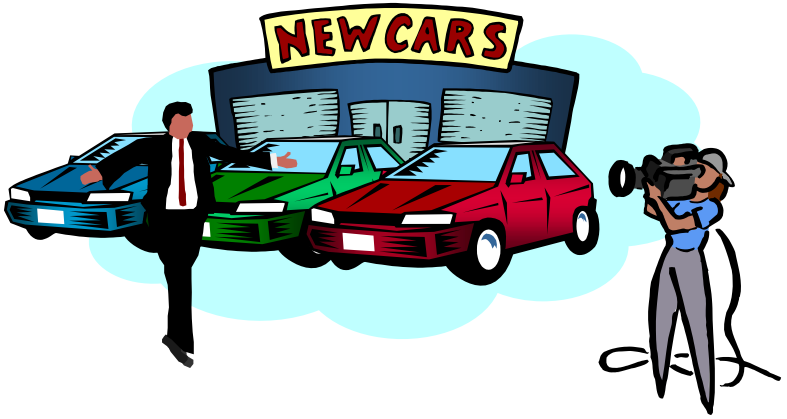
5.1 引例

5.2 相关分析

5.3 简单线性回归分析

5.4 多元线性回归分析

5.5 其他形式回归分析



5.1 引例

例5.1.1：Armand比萨餐馆

- 意大利食品连锁店——阿曼德比萨餐馆的成功之处在于它靠近大学校园。经过实地考察后，张经理推测，附近学校学生人数对餐馆的季度营业额会产生显著影响。为验证这一结论，张经理请小李收集了10家餐馆季度营业额和附近学校学生人数的数据如右表所示。现请在5%显著性水平下利用这些数据验证附近学生人数与季度销售额的相关关系。

表5.1.1 Armand数据

餐馆序号 i	学生人数（千人） x_i	季度销售额（千美元） y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

5.1 引例（续）

例5.1.2：Cravens的销售情况

- Cravens公司在一些销售区域销售产品，并且为每一个销售区域分别指定了独家经销商。近期，该公司为研究各销售区域的销售情况，随机收集了25个销售区域的销售数据，数据集的变量包括：

（1）Sales：区域经销商的实际总销售收入；

（2）Time：经销商从事销售时间（单位：月）；

（3）Poten：所在销售区域的潜在总销售额；

（4）AdvExp：销售区域的广告费用；

（5）Share：过去4年加权平均的市场份额；

（6）Change：过去4年间市场份额的变化；

（7）Accounts：顾客可以赊购的商店数目；

（8）Work：工作量；

（9）Rating：经销商的综合排序。

5.1 引例（续）

具体数据如下表所示，根据这些数据，对9个变量进行两两相关分析；

表5.1.2 Cravens各经销商的销售数据（部分）

Sales	Time	Poten	AdvExp	Share	Change	Accounts	Work	Rating
3669.88	43.1	74065.1	4582.9	2.51	0.34	74.86	15.05	4.9
3473.95	108.13	58117.3	5539.8	5.51	0.15	107.32	19.97	5.1
2295.1	13.82	21118.5	2950.4	10.91	-0.72	96.75	17.34	2.9
4675.56	186.18	68521.3	2243.1	8.27	0.17	195.12	13.4	3.4
6125.96	161.79	57805.1	7747.1	9.15	0.5	180.44	17.64	4.6
2134.94	8.94	37806.9	402.4	5.51	0.15	104.88	16.22	4.5
5031.66	365.04	50935.3	3140.6	8.54	0.55	256.1	18.8	4.6
3367.45	220.32	35602.1	2086.2	7.07	-0.49	126.83	19.86	2.3

5.2 相关分析

变量之间的关系

函数关系

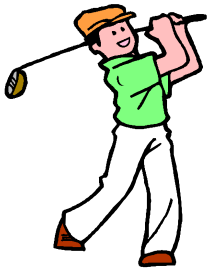
- **定义**：完全确定的对应关系。
- **实例**：圆面积与半径之间关系。

相关关系

- **定义**：变量之间具有较强的依赖关系，但是它们之间不能完全唯一地相互确定。或称统计相关、线性相关。
- **实例**：商品消费量与居民收入之间的关系；收入水平与受教育程度之间的关系。

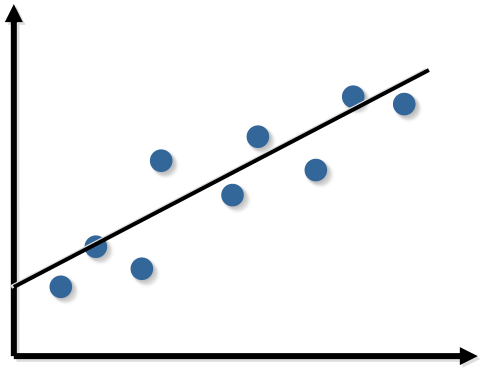
5.2 相关分析（续）

相关关系的类型

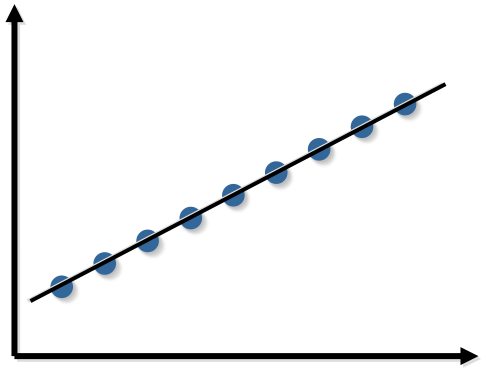


正相关

- 定义：变量之间的变动方向是相同的。
- 正相关图示：



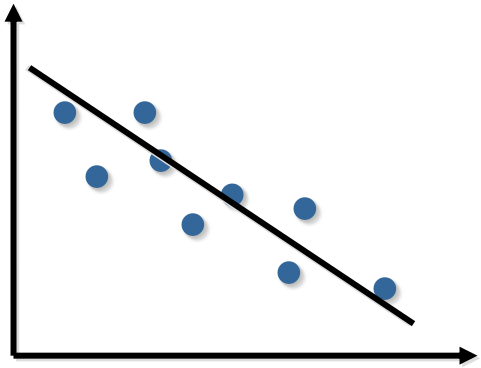
正相关



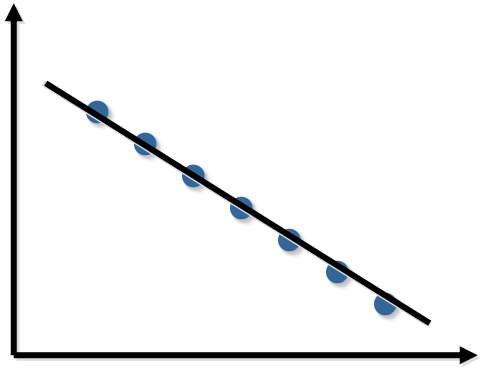
完全正相关

5.2 相关分析（续）

- 负相关
  - 定义：变量之间的变动方向是相反的。
  - 负相关图示：

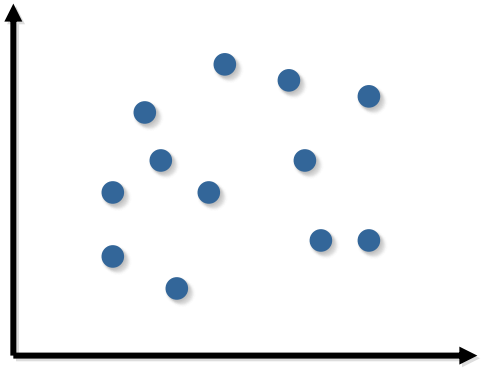


负相关



完全负相关

- 不相关
  - 定义：变量之间的变动没有关系。
  - 不相关图示：



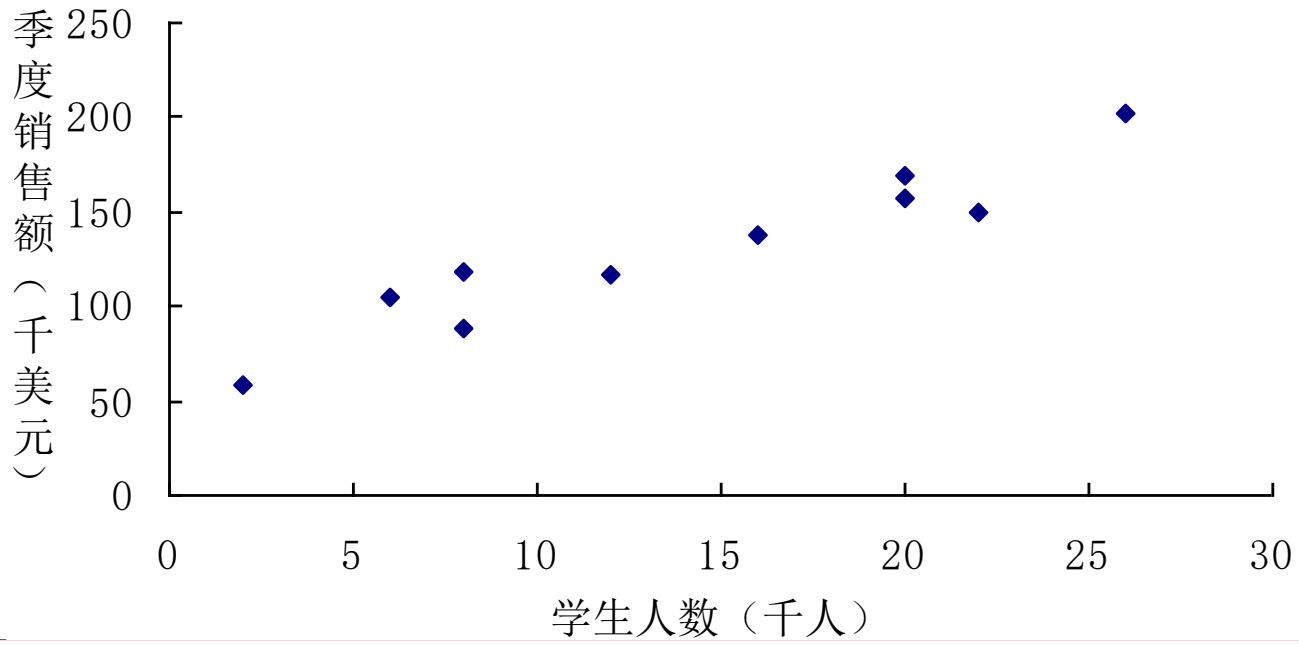
不相关

5.2 相关分析（续）

相关关系的测定方法

散点图

- 以点的方式将变量取值描绘在二维平面图上以表示变量之间的相关关系。
- 示例：Armand季度营业额与学生人数的散点图



5.2 相关分析（续）

- **软件实现**：绘制季度营业额与学生人数的散点图

Excel：选中两个变量的数据，点击“插入”-“散点图”

MATLAB：

```
[Data,~, All]=xlsread('Armand.xls'); %读取Excel
Population=Data(:,2); Sale=Data(:,3);
figure('color','w'); hold all;
plot(Population, Sale,'o','Linewidth',1,'MarkerFaceColor','b');
xlabel('学生人数); ylabel('季度营业额');
```

Python：

```
import pandas as pd
Armand = pd.read_excel('Armand.xls
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif']='Simhei’
plt.scatter(Armand['Population'], Armand['Sales'])
plt.xlabel(‘学生人数‘); plt.ylabel('季度营业额')
```

5.2 相关分析（续）

➢ 相关系数：Pearson定义

- **总体相关系数**：根据总体数据计算的，即

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

其中，σ_x、σ_y、σ_{xy} 分别为变量x的总体标准差、变量y的总体标准差和x与y的总体协方差。

总体相关系数是刻画变量真实关系的最佳指标，它常无法计算，需利用样本相关系数进行估计。

- **样本相关系数**：根据样本数据计算的，即

$$r = \frac{s_{xy}}{s_x s_y}$$

其中，s_x、s_y、s_{xy} 分别为变量x的样本标准差、变量y的样本标准差和x与y的样本协方差。

5.2 相关分析（续）

- 简易计算：

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n(n-1)}\left(n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i \sum_{i=1}^{n}y_i\right)$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n(n-1)}\left(n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right)}$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \sqrt{\frac{1}{n(n-1)}\left(n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2\right)}$$

$$r = \frac{s_{xy}}{s_x s_y}$$

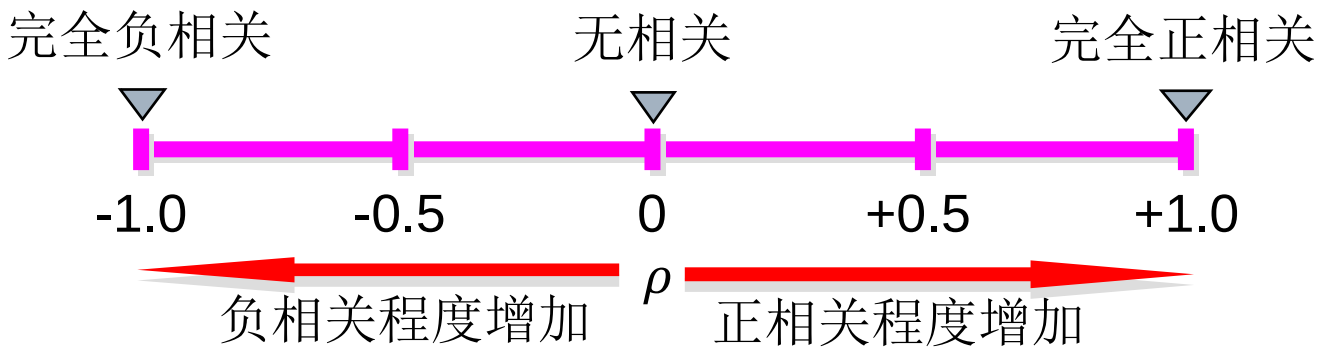
5.2 相关分析（续）

- **相关系数的取值范围及其意义**：

（1）$0 < \rho \leq 1$，正相关，当$\rho = 1$时，完全正相关；

（2）$\rho = 0$，不相关；

（3）$-1 \leq \rho < 0$，相关，当$\rho = -1$时，完全负相关。



- **问题讨论**：变量之间没有相关关系是否意味着变量之间没有任何关系？

5.2 相关分析（续）

- 计算示例：**Armand**季度营业额与学生人数的样本相关系数

$$r_{xy}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}$$

$$=\frac{10\times 21040-140\times 1300}{\sqrt{10\times 2528-140\times 140}\sqrt{10\times 184730-1300\times 1300}}$$

$$=0.9501$$

计算附表：

	x_i	y_i	x_iy_i	x_i^2	y_i^2	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
	2	58	116	4	3364	144	5184	864
	6	105	630	36	11025	64	625	200
	8	88	704	64	7744	36	1764	252
	8	118	944	64	13924	36	144	72
	12	117	1404	144	13689	4	169	26
	16	137	2192	256	18769	4	49	14
	20	157	3140	400	24649	36	729	162
	20	169	3380	400	28561	36	1521	234
	22	149	3278	484	22201	64	361	152
	26	202	5252	676	40804	144	5184	864
Σ	140	1300	21040	2528	184730	568	15730	2840

5.2 相关分析（续）

➢ 相关系数的显著性检验

- **检验假设：** $H_0:\rho=0$，$H_1:\rho\neq 0$
- **检验统计量：**

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

- **查找临界值并比较和判断：** 若 $|t_0|<t_{\alpha/2}$，则不能拒绝零假设，因此，相关关系可能是不显著的，否则，存在显著相关。
- 也可分别利用左侧和右侧检验来检验负相关和正相关关系。

5.2 相关分析（续）

➢ 相关分析的示例

● 示例**1**：**Armand**季度营业额与学生人数的相关系数显著性检验：

检验过程如下：

（1）构建检验假设：$H_0:\rho \le 0, H_1:\rho > 0$，

（2）计算检验统计量：

$$t_0 = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.9501}{\sqrt{(1-0.9501^2)/(10-2)}} = 8.6146$$

（3）给定5%的显著性水平，查表得到如下临界值：

$$t_{0.05}(8) = 1.860$$

由于$t_0 > t_{0.05}(8)$，因此，两者具有显著的正相关关系。

5.2 相关分析（续）

- **软件实现**：季度营业额与学生人数的相关系数

Excel：点击“数据”-“相关系数”-设置参数即可计算

MATLAB：

```
[Data,~, All]=xlsread('Armand.xls'); %读取Excel
Population=Data(:,2); Sale=Data(:,3);
[rho,pval] = corr(Population,Sale)
[r,p] = corrcoef(Population,Sale)
[RHO,PVAL] =
corrplot([Sale,Population],'varNames',{'Sale','Population'})
```

Python：

```
import pandas as pd
Armand = pd.read_excel('Armand.xls
Armand = Armand[['Population','Sales']]
r=Armand.corr()
```

5.2 相关分析（续）

- 示例2：Cravens9个变量的两两相关分析

	Sales	Time	Poten	AdvExp	Share	Change	Accounts	Work
Time	0.6229							
Poten	0.5978	0.454						
AdvExp	0.5962	0.2492	0.1741					
Share	0.4835	0.1061	-0.2107	0.2645				
Change	0.4892	0.2515	0.26829	0.3765	0.0855			
Accounts	0.754	0.7578	0.47864	0.2	0.403	0.3274		
Work	-0.117	-0.1794	-0.2588	-0.2722	0.3493	-0.288	-0.1988	
Rating	0.4019	0.1012	0.3587	0.4115	-0.024	0.5493	0.22861	-0.2769

从结果看出，Sales与Work的相关系数较小，Time与Accounts、Change与Rating的相关系数均较大，相关系数大于0.7的变量一般不应同时引入模型中。

5.2 相关分析（续）

- 软件实现：Cravens9个变量的两两相关分析

Excel：点击“数据”-“相关系数”-设置参数即可计算

MATLAB：

```
[~,~, All]=xlsread('Cravens.xls'); %读取Excel
Data=cell2mat(All(2:end,:)); %获取分析数据
VariableName=All(1,:); %获取变量名称
Option={'type','Pearson','testR','on','tail','both','varNames',VariableName}; %设定检验选项
CorrelationAnalysis=CorrelationAnalysis(Data,VariableName,Option) %相关性检验
```

Python：

```
import pandas as pd; import seaborn as sns
Cravens = pd.read_excel('Cravens.xls’)
corr=Cravens .corr() #计算相关系数
sns.heatmap(corr, xticklabels=corr.columns, yticklabels=corr.columns) #计算相关系数矩阵图
```

5.3 简单回归分析

回归分析的一般问题

➢ 回归分析的概念

- 研究一个（或某些）变量对另一个（或一些）变量是否产生影响以及会产生多大的影响程度的一种统计分析方法。比如，每增加一元的广告支出将会对商品销售额产生什么样的影响？
- **与相关分析的关系：**相关分析是回归分析的前提和基础；回归分析是相关分析的深化和扩展；两者的变量地位和性质不同、分析侧重点不同。

➢ 回归分析的类型

- **简单回归分析：**研究一个变量对另一个变量的影响关系，又称一元回归分析。
- **多元回归分析：**研究多个变量对另一个变量的影响关系。

5.3 简单回归分析（续）

- **线性回归分析**：变量之间的影响关系是线性的；
- **非线性回归分析**：变量之间的影响关系是非线性的；

➢ 回归分析的基本步骤

- 第1步，通过定性分析和相关分析方法判定变量之间的关系；
- 第2步，构建回归分析模型（一元或多元、线性或非线性等）；
- 第3步，对回归分析模型及其参数进行估计和显著性检验；
- 第4步，利用获得的回归分析模型进行估计、预测和决策。

5.3 简单回归分析（续）

简单线性回归模型的基本形式

➢ 简单线性总体回归模型

- **目标**：研究两个变量之间的线性影响关系。
- **形式**：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

其中，x_i、y_i分别称为自变量（解释变量）和因变量（被解释变量）取值（i=1,2,…,N）；β_0、β_1分别称为常数项和自变量的回归系数；ε_i称为随机扰动项（随机误差项）。

- **因变量变化的构成**：从总体回归模型可知，因变量y的变化由下列两部分组成：

一是，**由自变量x变化影响引起的变化**（即$\beta_1 x_i$的变化部分），该变化是主要的、线性的；

二是，由除自变量x以外的**其他因素影响引起的变化**（即ε_i），该变化是次要的、随机的。

5.3 简单回归分析（续）

- **模型意义**：总体回归模型反映了在总体中x对y的影响关系，这种关系是客观的、真实的、不确定的（近似线性），但通常是无法知道的。

➢ 总体回归方程（直线）

- **方程形式**：

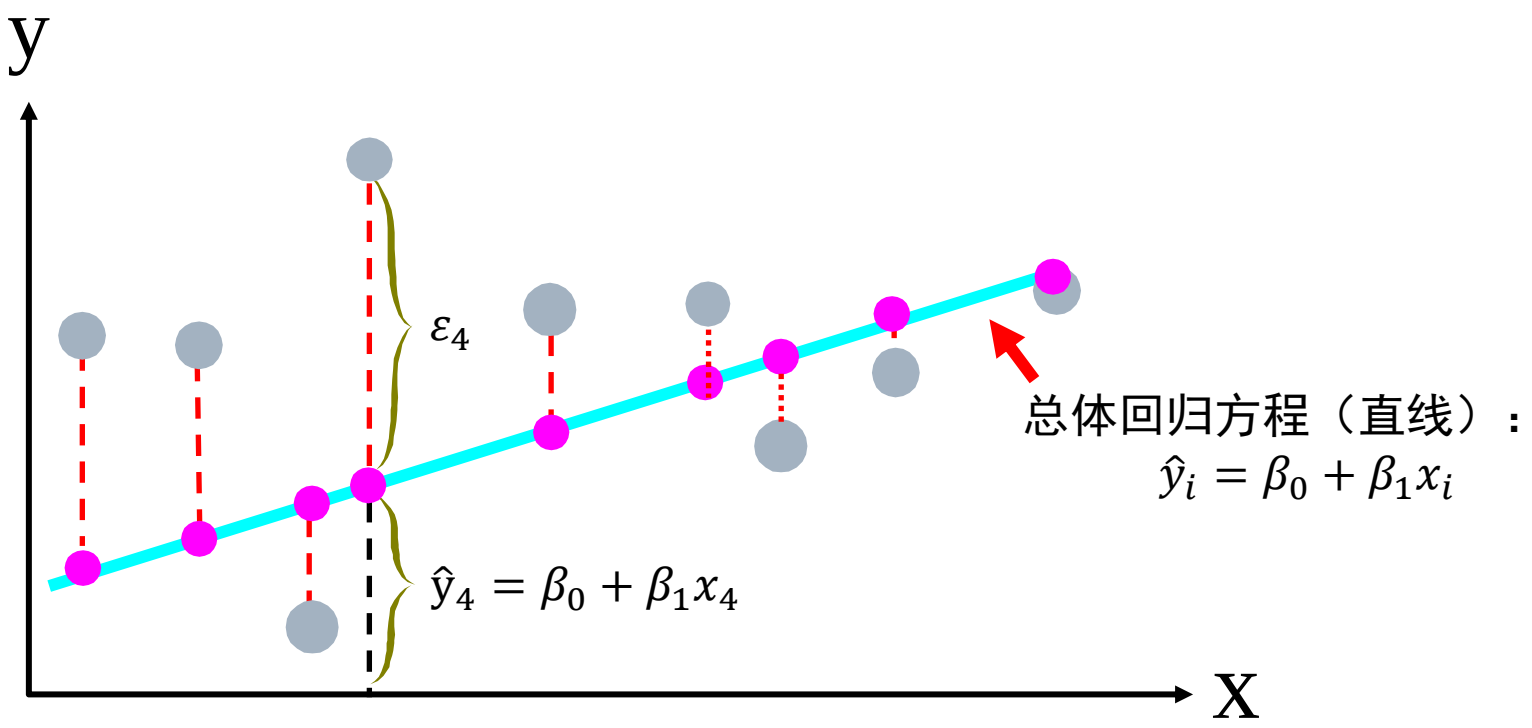
$$E(y_i) = \beta_0 + \beta_1 x_i$$

其中，$E(y_i)$为因变量的期望，β_0、β_1分别为方程（直线）的截距和斜率。

- **方程意义**：总体回归方程反映了自变量对因变量的期望的影响关系，这种关系是客观的、真实的、确定的（完全线性），但通常也无法知道。

5.3 简单回归分析（续）

- 图示：



5.3 简单回归分析（续）

- **最优的总体回归直线**：要使回归方程作出的对因变量的所有估计值 $\hat{y}_i$（i=1, 2, …, N）最接近于因变量的所有样本观察值 y_i（i=1, 2, …, N），则必须使随机误差均方和（Mean Sum of the Squared Errors，MSE）最小，其定义为：

$$MSE = \mathrm{E}\{\varepsilon_i^2\} = E\{(y_i - \hat{y}_i)^2\}$$

5.3 简单回归分析（续）

➢ 样本回归模型

- **模型形式：**

$$y_i = b_0 + b_1 x_i + e_i$$

其中，x_i、y_i分别为自变量和因变量样本观察值（i=1, 2, …, n）；b_0、b_1分别为样本回归模型的常数项和和回归系数，e_i为样本回归模型的残差项。

- **模型意义：**样本回归模型反映了在样本中x对y的影响关系，因这种影响随样本的不同而不同，因而它是随机的，也即常数项和回归系数是随机变量，而总体回归模型中的常数项和回归系数是确定变量。

5.3 简单回归分析（续）

➢ 样本回归方程（直线）

- 方程形式：

$$\hat{y}_i = b_0 + b_1 x_i$$

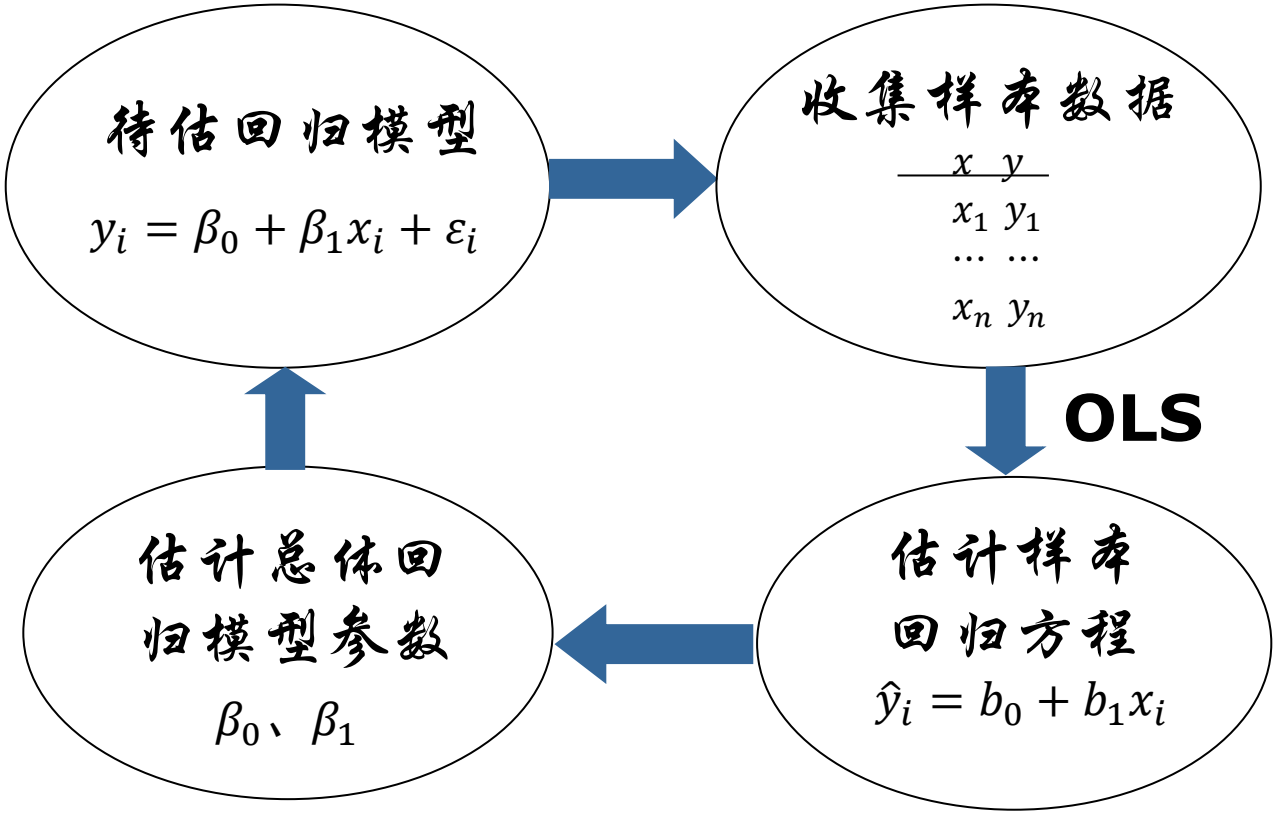
其中，$\hat{y}_i$ 为因变量的第i个估计值，b_0、b_1 为方程（直线）的截距和斜率。

因此，回归分析的目的就是估计出样本回归方程的参数 b_0、b_1，并将其作为总体回归方程的参数 β_0、β_1 的估计。

5.3 简单回归分析（续）

简单线性回归模型的参数估计

参数估计过程



回归模型参数估计过程

5.3 简单回归分析（续）

- **OLS**：要使样本回归方程作出的对因变量的所有估计 $\hat{y}_i$（i=1, 2, …, n）最接近于因变量的所有样本观察值 y_i（i=1, 2, …, n），则必须使残差平方和（Sum of the Squared Errors，SSE）：

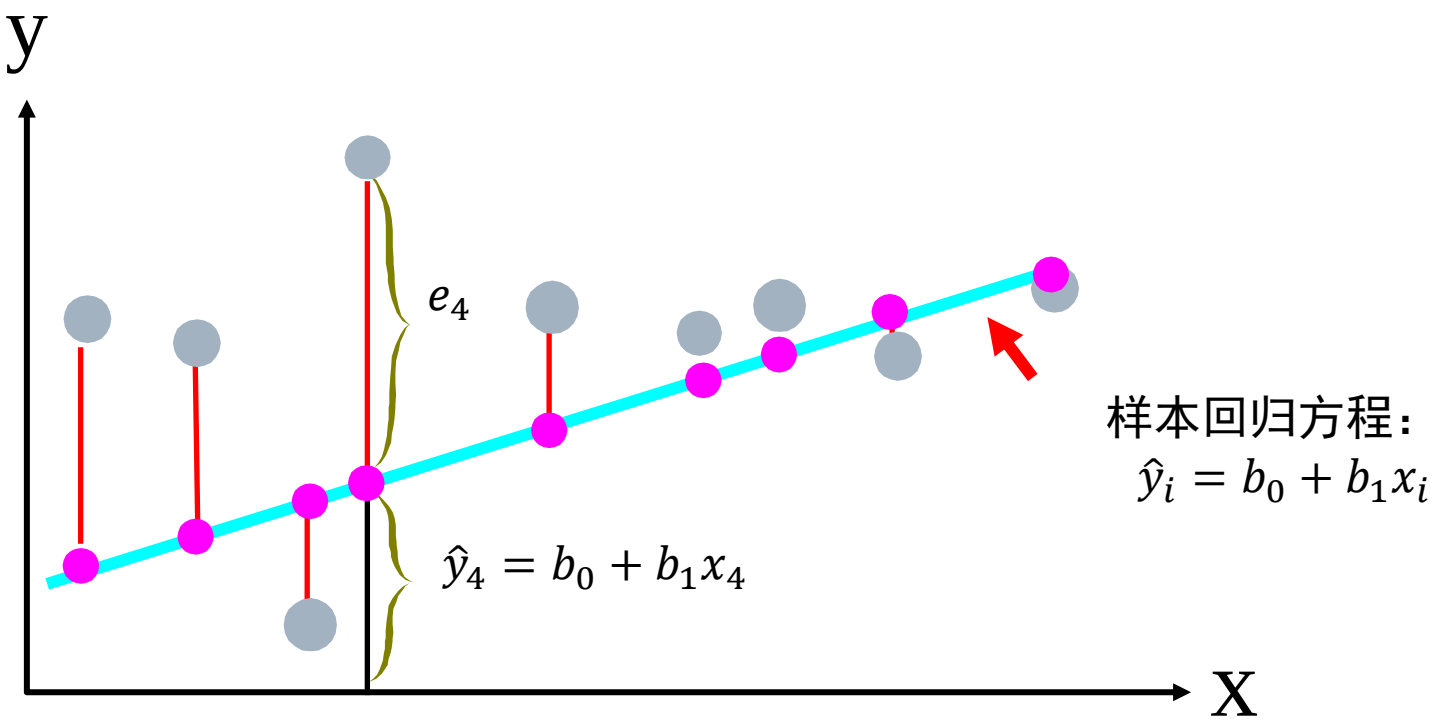
$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

达到最小。

因此，OLS以**残差平方和达到最小**为估计准则。

5.3 简单回归分析（续）

- **OLS图解：**



5.3 简单回归分析（续）

- **模型假定**：为便于模型的估计和检验，常需对总体回归模型的随机误差项做如下假定：

假定 1：随机误差项的期望为零，即为对于所有的 i，总有

$$E(\varepsilon_i) = 0$$

假定 2：随机误差项的方差为常数，即为对于所有的 i，总有

$$Var(\varepsilon_i) = \sigma^2$$

假定 3：随机误差项之间互不相关，即为对于所有的 i 和 j（i≠j），总有

$$cov(\varepsilon_i, \varepsilon_j) = 0$$

假定 4：随机误差项与自变量之间互不相关，即为对于所有的 i，总有

$$cov(\varepsilon_i, x_i) = 0$$

假定 5：随机误差项服从正态分布，即为对于所有的 i，总有

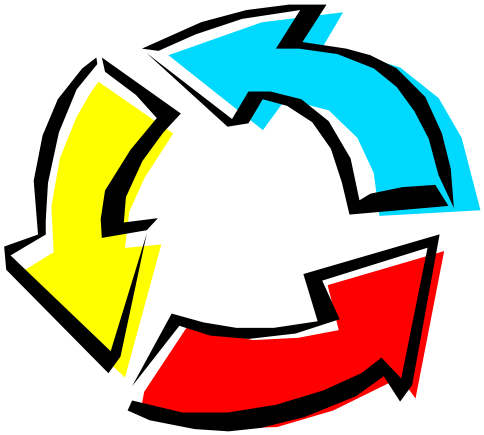
$$\varepsilon_i \sim N(0, \sigma^2)$$

5.3 简单回归分析（续）

➢ 参数OLS估计量

- **OLS估计量形式：**

$$\begin{cases} b_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ b_0 = \bar{y} - b_1\bar{x} \end{cases}$$



- **OLS估计量的均值和标准误：**

$$\begin{cases} E(b_0) = \beta_0 \\ s_{b_0} = \sigma\sqrt{\dfrac{1}{n} + \dfrac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \end{cases}、\quad \begin{cases} E(b_1) = \beta_1 \\ s_{b_1} = \sigma\sqrt{\dfrac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \end{cases}$$

5.3 简单回归分析（续）

- 随机误差项的标准差估计
  - 随机误差项的标准差 σ，因常无法知道，需要进行估计。
  - **无偏估计量形式：**

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

  - 该项常称为估计标准误差。
- 估计中的3个平方和
  - **总平方和：**

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

即为因变量样本实际观察值与其样本均值的平方和，反映样本观察值的总离散程度和总差异程度。

5.3 简单回归分析（续）

- **回归平方和：**

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

即为因变量的估计值与样本均值的离差平方和，反映样本回归方程(自变量) 可解释的那部分样本观察值的离散程度和差异程度；

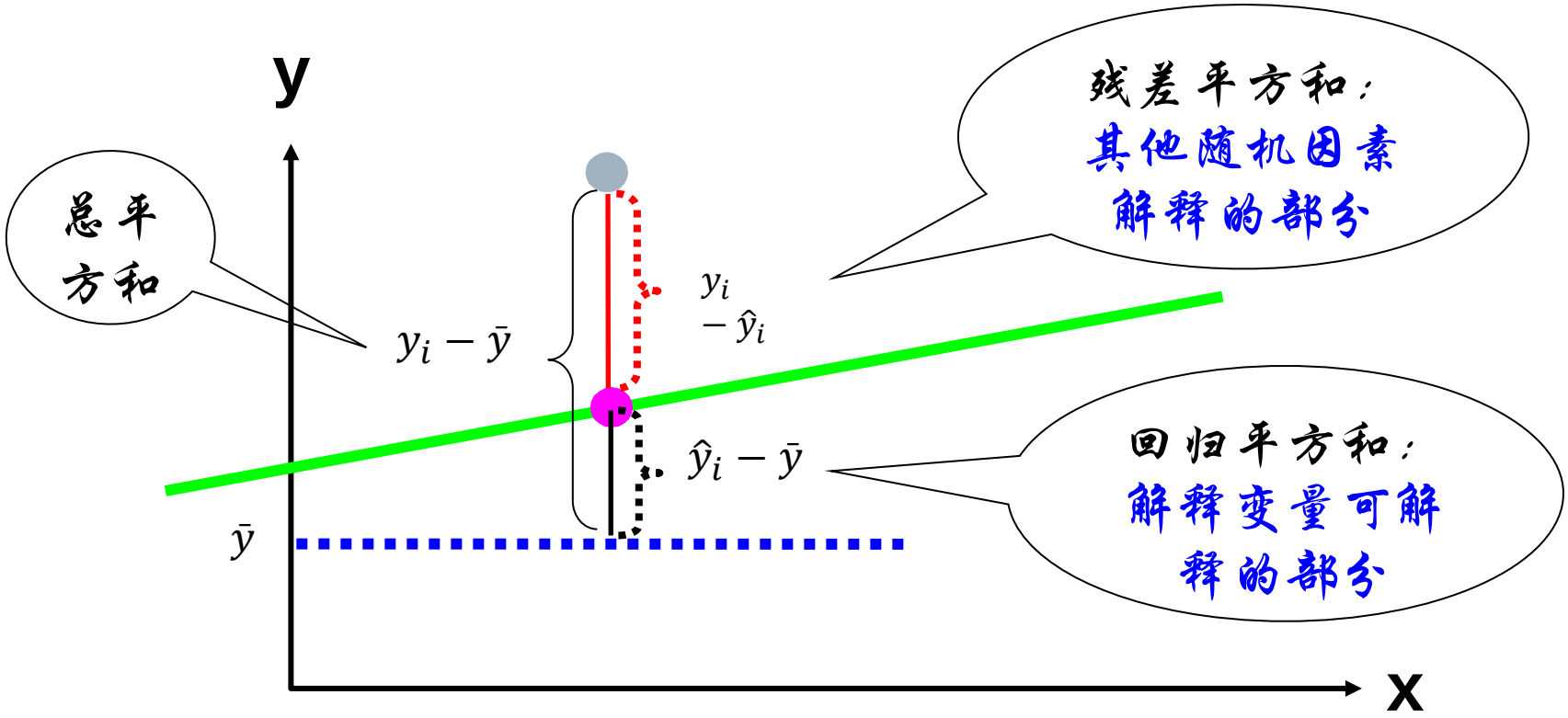
- **残差方和：**

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

因变量的样本实际观察值与估计值的离差平方和，反映样本回归方程（自变量）无法解释的而由除自变量以外的其他因素引起的样本观察值的离散程度和差异程度。

5.3 简单回归分析（续）

- 三者关系：$SST=SSR+SSE$。



5.3 简单回归分析（续）

简单线性回归模型参数的显著性检验：t检验

- 第1步，提出假设： $H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$
- 第2步，构造t检验统计量并计算其值 t_0：

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \sim t_{(n-2)}$$

- 第3步，给定显著水平，查临界值，与计算值比较并做出判断：

若 $|t_0| \geq t_{\alpha/2}(n-2)$，则拒绝$H_0$，认为X对Y具有显著影响；

若 $|t_0| < t_{\alpha/2}(n-2)$，则不能拒绝$H_0$，认为没有显著影响。

或者，计算p值，若 $p \leq \alpha$，则拒绝H0。

5.3 简单回归分析（续）

简单线性回归模型整体的显著性检验

模型整体的拟合优度评价：判定系数

- **拟合效果**：若SSE越小，SSR越大，则样本观察值聚集在样本回归直线周围就越紧密，样本回归直线对样本观察值的拟合效果就越好。
- **判定系数**：

$$R^2 = \frac{SSR}{SST}$$

取值范围为[0,1]，反映样本回归方程对样本观察值的拟合程度和自变量对因变量的解释能力，其值越大表明样本回归直线对样本观察值的拟合效果越好，所有自变量对因变量变化的解释能力越强。

- **与相关系数的关系**：

$$r_{xy} = (b_1\text{的符号})\sqrt{\text{判定系数}}$$

5.3 简单回归分析（续）

➢ 模型整体的显著性检验：F检验

- 第1步，提出假设：H_0：$\beta_1 = 0$，H_1：$\beta_1 \neq 0$
- 第2步，构建F检验统计量并计算其值 F_0：

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

- 第3步，给定显著性水平 α，查临界值 F_α，与检验统计量的计算值比较，并做出判断：

若 $F_0 > F_\alpha$，拒绝H_0，表明所估计的回归模型整体上是显著的。

- 提示：简单线性回归模型的t检验与F检验本质上是等价的。

5.3 简单回归分析（续）

- 回归模型的方差分析（Analysis of Variance，ANOVA）

	自由度（df）	平方和（SS）	均方和（MS）	F统计量	F检验p值
回归	p	SSR	MSR=SSR/p	$F=\frac{MSR}{MSE}$	p
残差	n-p-1	SSE	MSE=SSE/(n-p-1)		
总和	n-1	SST			

5.3 简单回归分析（续）

利用样本回归方程对因变量进行估计和预测

因变量 y_p 及其均值 $E(y_p)$ 的点估计

- 只要将自变量某个取值代入样本回归方程，即可求出因变量的点估计值 $\hat{y}_p = b_0 + b_1 x_p$，同时 $\hat{y}_p$ 也是因变量均值的点估计 $E(y_p) = b_0 + b_1 x_p$ 。

5.3 简单回归分析（续）

Armand餐馆季度销售额对学生人数的回归分析

回归参数估计

- 根据计算附表，可得

$$\begin{cases} b_1 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \dfrac{2840}{568} = 5 \\ b_0 = \bar{y} - b_1\bar{x} = 130 - 5 \times 14 = 60 \end{cases}$$

- 所估计的简单线性回归方程为 $\hat{y}_i = 60 + 5x_i$，反映了学生人数每增加1千人，季度销售额将增加5千美元。

计算附表：

	x_i	y_i	x_iy_i	x_i^2	y_i^2	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
	2	58	116	4	3364	144	5184	864
	6	105	630	36	11025	64	625	200
	8	88	704	64	7744	36	1764	252
	8	118	944	64	13924	36	144	72
	12	117	1404	144	13689	4	169	26
	16	137	2192	256	18769	4	49	14
	20	157	3140	400	24649	36	729	162
	20	169	3380	400	28561	36	1521	234
	22	149	3278	484	22201	64	361	152
	26	202	5252	676	40804	144	5184	864
Σ	140	1300	21040	2528	184730	568	15730	2840

5.3 简单回归分析（续）

➢ 单个参数的显著性检验

- 估计标准误差s的估计：根据估计的回归方程，可估计出10个$\hat{y}_i$依次为：70, 90, 100, 100, 120, 40, 160, 160, 170, 190，由此，残差平方和的估计值为

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 1530$$

，因此，估计标准误差为

$$s = \sqrt{SSE/(n-2)} = \sqrt{1530/(n-2)} = 13.829。$$

- 回归系数标准差的估计：

$$s_{b_1} = s\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 13.829 \times \sqrt{1/568} = 0.58025$$

5.3 简单回归分析（续）

- 回归系数的t检验：

检验假设为：$H_0:\beta_1=0$，$H_1:\beta_1\neq 0$，检验统计量的计算值为：$t_0=(5-0)/0.58025=8.62$，给定1%的显著性水平，$t_{0.005}(10-2)=3.355$，从而，$|t_0|>t_{0.005}$，因此，拒绝H_0，认为回归系数显著不为零，从而表明学生人数对季度销售额具有显著的影响。

➢ 模型整体的显著性检验

- 判定系数：由于 $SST=\sum_{i=1}^{n}(y_i-\bar{y})^2=15730$，则 $SSR=SST-SSE=14200$，

因此，$R^2=SSR/SST=14200/15730=0.9027$，

判定系数取值较高，从而表明，学生人数对季度销售额的解释能力较强，样本回归直线对样本观察值的拟合效果较好。

5.3 简单回归分析（续）

- 模型的F检验：

检验假设为： $H_0:\beta_1=0$，$H_1:\beta_1\neq 0$ ，检验统计量的计算值为： $F_0=(14200/1)/(1530/8)=74.25$

给定1%显著性水平，$F_{0.01}(1,8)=11.26$，由于 $F_0>F_{0.01}$ ，所以拒绝H_0，因此，回归模型整体上是显著的。

➢ 季度销售额及其均值的预测

- 当学生人数为1万人的季度销售额及其均值的点估计值均为 $\hat{y}_i=60+5x_i=60+5\times 10=110$（千美元）。

5.3 简单回归分析（续）

- 回归分析的总体结论
  - Armand餐馆季度销售额对附近学校学生人数的简单线性回归方程为 $\hat{y}_i = 60 + 5x_i$。
  - 学生人数对季度销售额具有显著影响，回归模型整体上是显著的。
  - 学生人数为1万人的某店季度销售额及其均值的估计值均为110千美元。

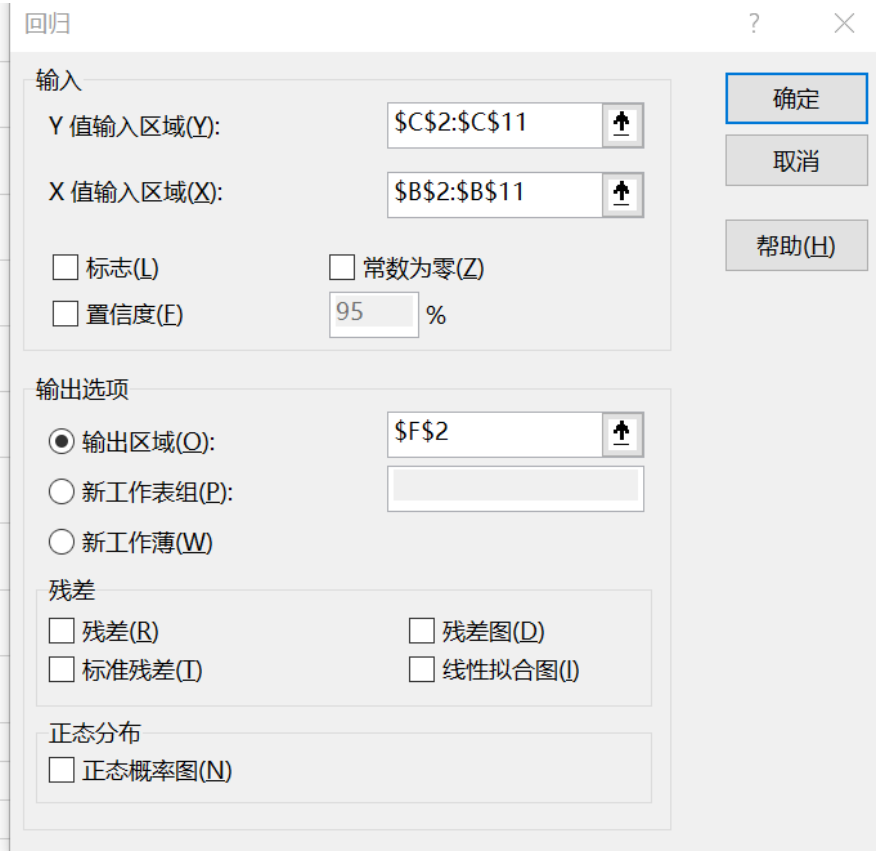
5.3 简单回归分析（续）

简单回归分析的软件实现

Excel操作步骤

- 点击“数据”-“回归”，在回归窗口进行相关设置即可运行回归结果

Restaurant	Population	Sales
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202



5.3 简单回归分析（续）

- Armand餐馆的Excel回归结果：

回归统计	
Multiple R	0.950123
R Square	0.9027336
Adjusted R Square	0.8905753
标准误差	13.829317
观测值	10

判定系数

三平方和　　F值及其P值

方差分析					
	df	SS	MS	F	Significance F
回归分析	1	14200	14200	74.24837	2.5489E-05
残差	8	1530	191.25		
总计	9	15730			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	60	9.226035	6.503336	0.000187	38.7247256	81.275274
X Variable 1	5	0.580265	8.616749	2.55E-05	3.66190596	6.338094

截距　　斜率　　T值及其P值

5.2 简单线性回归分析（续）

- Python代码：Armand季度营业额与学生人数的简单回归分析

```
import statsmodels.api as sm; import pandas as pd
Armand = pd.read_excel('Armand.xls')
x=Armand[`Population`]; y=Armand['Sales']
x = sm.add_constant(x); SLR=sm.OLS(y,x).fit();
print(SLR.summary())
```

```
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.903
Model:                            OLS   Adj. R-squared:                  0.891
Method:                 Least Squares   F-statistic:                     74.25
Date:                Thu, 09 Jun 2022   Prob (F-statistic):           2.55e-05
Time:                        21:36:19   Log-Likelihood:                -39.342
No. Observations:                  10   AIC:                             82.68
Df Residuals:                       8   BIC:                             83.29
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         60.0000      9.226      6.503      0.000      38.725      81.275
Population     5.0000      0.580      8.617      0.000       3.662       6.338
==============================================================================
```

5.4 多元回归分析

多元线性回归模型的基本形式

总体回归模型

- 模型形式：

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

其中，$i = 1,2,\cdots,N$，β_0、β_i 分别为常数项和第i个解释变量的回归系数，ε_i 为随机误差项。

- 模型意义：将因变量的变化分为两部分：由p个解释变量变化引起的部分和除p个解释变量以外其他所有因素引起的部分。
- 模型的经典假定：（1）$E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2$；（2）$\varepsilon_i$ 与 ε_j $(i \neq j)$ 相互独立；（3）$\varepsilon_i \sim N(0, \sigma^2)$；（4）$x_i$与 x_j $(i \neq j)$ 相互独立，且 x_i与 ε_i相互独立。

5.4 多元回归分析（续）

- 总体回归方程（直线）

$$E(y_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

其中，$i = 1,2,\cdots,N$。

- 样本回归模型

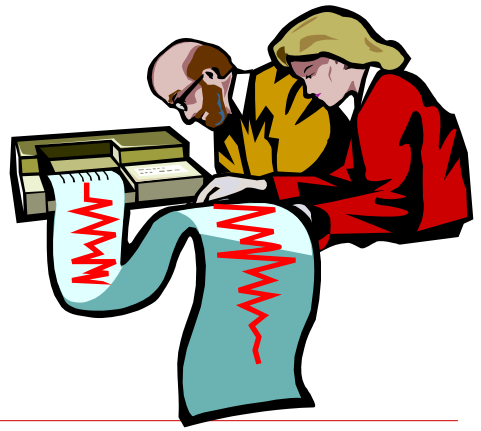
$$y_i = b_0 + b_1 x_{1i} + \cdots + b_p x_{pi} + e_i$$

其中，$i = 1,2,\cdots,n$。

- 样本回归方程（直线）

$$\hat{y}_i = b_0 + b_1 x_{1i} + \cdots + b_p x_{pi}$$

其中，$i = 1,2,\cdots,n$。



5.4 多元回归分析（续）

多元线性回归模型的参数估计

OLS估计量形式

- 记

$$Y=\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{pmatrix},\ X=\begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ & \cdots & \cdots & \cdots \\ 1 & x_{1N} & \cdots & x_{p\mathrm{N}} \end{pmatrix},\ B=\begin{pmatrix} b_0 \\ b_1 \\ \cdots \\ b_p \end{pmatrix}$$

则，参数向量的估计量为：

$$\boldsymbol{B}=(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}$$

5.4 多元回归分析（续）

➢ OLS估计量的均值和标准差

$$\begin{cases} E(b_k) = \beta_k \\ \sigma_{b_k} = \sigma\sqrt{\psi_{kk}} \end{cases}$$

其中，ψ_{kk}为 $(X^TX)^{-1}$ 的第k个对角线元素，σ 为随机误差项的标准差，常通过如下形式的估计标准误差s来估计：

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-p-1}}$$

5.4 多元回归分析（续）

多元线性回归模型的显著性检验

单个参数的显著性检验：t检验

- 第1步，提出假设：$H_0: \beta_i = 0$,$H_1: \beta_i \neq 0$
- 第2步，构造t检验统计量并计算其值 t_0：

$$t = \frac{b_i - \beta_i}{s_{b_i}} \sim t_{(n-p-1)}$$

其中，p为模型包含的自变量个数。

- 第3步，给定显著水平，查临界值，与计算值比较并做出判断：若 $|t_0| \geq t_{\alpha/2}(n-p-1)$，则拒绝$H_0$，认为 x_i对Y具有显著影响，反之，则不能拒绝H_0，认为没有显著影响。

或者，计算p值，若 $p \leq \alpha$，则拒绝H0。

5.4 多元回归分析（续）

➢ 拟合优度评价：判定系数

- 多元判定系数：

$$R^2 = \frac{SSR}{SST}$$

式中，各参数的含义意义同前。

- 修正多元判定系数：

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

式中，n为样本容量，p为自变量个数。

- **问题讨论**：随着自变量个数增加，多元判定系数和修正多元判定系数分别会如何变化？为什么？

5.4 多元回归分析（续）

➢ 模型显著性检验：F检验

- 第1步，提出假设：

  H_0：$\beta_1 = \beta_2 = \cdots = \beta_p = 0$，$H_1$：至少有一个$\beta_i \neq 0$

- 第2步，构建F检验统计量并计算其值 F_0：

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

- 第3步，给定显著性水平 α，查临界值 F_α，与检验统计量的计算值比较，并做出判断：若 $F_0 > F_\alpha$，则拒绝H_0，表明所估计的回归模型在整体上是显著的。

或者，计算p值，若 $p \leq \alpha$，则拒绝H0。

5.4 多元回归分析（续）

解释变量的选择

意义

- 在多元回归分析中，哪些变量作为解释变量至关重要，因此，需要通过一定的方法选择某些对被解释变量具有显著影响的变量。

确定方法

- **t检验法**：检验某解释变量的回归系数是否显著等于0，若它与0没有显著差异，则一般不将该变量引入回归模型中。
- **F检验法**：检验某些解释变量引入前后的模型是否存在显著差异，若存在显著差异，则一般可将这些解释变量引入回归模型中，F检验统计量形式为：

5.4 多元回归分析（续）

$$F = \frac{[SSE(x_1,\cdots,x_q) - SSE(x_1,\cdots,x_q,x_{q+1},\cdots,x_p)]/(p-q)}{SSE(x_1,\cdots,x_q,x_{q+1},\cdots,x_p)/(n-p-1)}$$
$$\sim F\,(p-q,n-p-1)$$

其中，$SSE(x_1,\cdots,x_q)$、$SSE(x_1,\cdots,x_q,x_{q+1},\cdots,x_p)$ 分别为包含解释变量 $x_1,\cdots,x_q$ 和 $x_1,\cdots,x_q,x_{q+1},\cdots,x_p$ 模型的残差平方和。给定 α 显著性水平，若 $F \geq F_\alpha$，则说明模型之间存在显著差异。

5.4 多元回归分析（续）

- **选择的具体方式：**

（1）**前向选择：**从模型中没有变量开始，按对因变量的贡献由大到小依次引进自变量（每次只增加一个自变量，被加入的变量不能删除），直到按某一标准没有变量被引进为止。

（2）**后向消元：**从模型中包含所有变量开始，按对因变量的贡献由小到大依次引进自变量（每次只删除一个自变量，被删除的变量不能加入），直到按某一标准没有变量被删除为止。

（3）**逐步回归：**是前向选择和后向消元的结合，可同时增加和删除变量。

（4）**最佳子集回归：**首先将模型划分成几大类，然后从每类中选择较佳回归，最后从所有最佳回归模型中选择最佳回归方程。

5.4 多元回归分析（续）

Cravens销售的多元回归分析

两两相关分析：Cravens数据

- 9个变量的两两相关系数如下表所示。从结果看出，Sales与Work的相关系数较小，Time与Accounts、Change与Rating的相关系数均较大，相关系数大于0.7的变量一般不应同时引入模型中。

	Sales	Time	Poten	AdvExp	Share	Change	Accounts	Work
Time	0.6229							
Poten	0.5978	0.454						
AdvExp	0.5962	0.2492	0.1741					
Share	0.4835	0.1061	-0.2107	0.2645				
Change	0.4892	0.2515	0.26829	0.3765	0.0855			
Accounts	0.754	0.7578	0.47864	0.2	0.403	0.3274		
Work	-0.117	-0.1794	-0.2588	-0.2722	0.3493	-0.288	-0.1988	
Rating	0.4019	0.1012	0.3587	0.4115	-0.024	0.5493	0.22861	-0.2769

5.4 多元回归分析（续）

- Sales对全部解释变量（8变量）的多元线性回归分析
  - 从下表可看出，判定系数和修正判定系数均较高，模型的整体性检验是显著的。

包含全部解释变量的多元回归分析结果1

回归统计					
Multiple R	0.960231				
R Square	0.922043				
Adjusted R Square	0.883064				
标准误差	449.0154				
观测值	25				
方差分析					
	df	SS	MS	F	Significance F
回归分析	8	38153712	4769214	23.65508	1.8149E-07
残差	16	3225837	201614.8		
总计	24	41379549			

5.4 多元回归分析（续）

包含全部解释变量的多元回归分析结果2

	Coefficients	标准误差	t Stat	P-value
Intercept	-1507.836	778.6084	-1.93658	0.070663
Time	2.010137	1.930514	1.041244	0.313241
Poten	0.037206	0.008202	4.53607	0.000337
AdvExp	0.150984	0.047107	3.205128	0.005518
Share	199.0402	67.02915	2.969457	0.009037
Change	290.8666	186.7769	1.557294	0.138958
Accounts	5.549745	4.775443	1.162142	0.26222
Work	19.79389	33.67517	0.587789	0.564878
Rating	8.189035	128.4985	0.063729	0.949976

- 从上表可看出，除了Poten、AdvExp、Share之外，其他解释变量的p值均大于0.05，说明该回归方程的解释变量选择不理想，需重新筛选。

5.4 多元回归分析（续）

➢ 最佳回归方程的确定

- 解释变量的筛选顺序：按照它们在包含全部解释变量的回归模型的p值从大到小（或与被解释变量的相关系数从小到大）的顺序，运用后向消元法在模型中依次剔除，直到剩余的所有解释变量的p值均小于0.5为止。
- 剔除Rating后的回归结果：

	Coefficients	标准误差	t Stat	P-value
Intercept	-1485.9	677.6281	-2.1928	0.042523
Time	1.975128	1.795677	1.099935	0.286693
Poten	0.037291	0.007851	4.749931	0.000185
AdvExp	0.151956	0.043245	3.513858	0.002663
Share	198.3252	64.11869	3.093096	0.006601
Change	295.8774	164.3831	1.799926	0.089644
Accounts	5.608949	4.544927	1.234112	0.23395
Work	19.89892	32.63471	0.609747	0.550093

5.4 多元回归分析（续）

- 剔除Work后的回归结果：

	Coefficients	标准误差	t Stat	P-value
Intercept	-1165.51	420.3538	-2.77269	0.012549
Time	2.26986	1.69895	1.336037	0.198184
Poten	0.038279	0.007547	5.072208	7.94E-05
AdvExp	0.140665	0.038392	3.663933	0.001776
Share	221.6221	50.58526	4.381158	0.00036
Change	285.1205	160.5564	1.775827	0.09267
Accounts	4.376557	3.999119	1.09438	0.288217

- 剔除Acounts后的回归结果：

	Coefficients	标准误差	t Stat	P-value
Intercept	-1113.87	419.8638	-2.65293	0.015702
Time	3.612255	1.181606	3.057072	0.006486
Poten	0.042088	0.006731	6.25297	5.27E-06
AdvExp	0.128854	0.037035	3.479266	0.002511
Share	256.969	39.13361	6.566451	2.75E-06
Change	324.5354	157.2766	2.063469	0.053001

5.4 多元回归分析（续）

- 剔除Change后的回归结果：

	Coefficients	标准误差	t Stat	P-value
Intercept	-1312.38	440.7254	-2.97778	0.007439
Time	3.816596	1.269733	3.005825	0.006984
Poten	0.044396	0.007158	6.202691	4.65E-06
AdvExp	0.152474	0.037982	4.014407	0.00068
Share	259.4843	42.18028	6.151793	5.2E-06

- 从上表可看出，四个变量Time、Poten、AdvExp、Share的p值均小于0.05，因此，这四个变量均可包含在回归模型中。最终的多元线性回归方程为：

$$\hat{Sales} = -1312.38 + 3.82Time + 0.0444Poten + 0.1525AdvExp + 259.48Share$$

5.5 其他形式的回归分析

对数线性回归模型

- 模型形式

$$Y_i = \alpha \prod_{k=1}^{p} X_{ki}^{\beta_k} e^{\varepsilon_i}$$

- 转换成线性回归模型：

$$\ln(Y_i) = \ln(\alpha) + \beta_1 \ln(X_{1i}) + \cdots + \beta_p \ln(X_{pi}) + \varepsilon_i$$

记 $y_i = \ln(Y_i)$，$\beta_0 = \ln(\alpha)$，$x_{1i} = \ln(X_{1i})$，…，$x_{pi} = \ln(X_{pi})$

则，对数线性回归模型可以转换成多元线性回归模型形式：

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

5.5 其他形式的回归分析（续）

半对数回归模型

- 模型形式：

$$Y_i = e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_{pi}x_{pi}+\varepsilon_i}$$

- 转换成多元线性回归模型：记 $y_i = \ln(Y_i)$，则半对数回归模型可转换成一般多元线性回归模型形式：

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{pi}x_{pi} + \varepsilon_i$$

线性概率模型

- 基本形式： $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{pi}x_{pi} + \varepsilon_i$
- 与一般多元线性回归模型的区别：因变量取值为离散值，若因变量只取0和1，则它的均值表示取0和1的概率。

5.5 其他形式的回归分析（续）

Probit/Logit模型

基本形式

- 若因变量取值为离散值，则将其直接与解释变量建立线性概率模型存在巨大局限性，因此，引入因变量的潜在变量，建立该潜在变量对解释变量的线性回归模型：

$$y_i^* = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{pi} x_{pi} + \varepsilon_i^*$$

其中，潜在变量与因变量的联系为：

$$y_i = \begin{cases} 0, y_i^* \le 0 \\ 1, y_i^* > 0 \end{cases}$$

5.5 其他形式的回归分析（续）

- **Probit模型**
  - 若 ε_i^* 服从标准正态分布函数，则其累积分布函数 $G(z)=\Phi(z)$，该模型称为Probit模型，而且

$$Prob(y_i=1|X)=\int_{-\infty}^{X^T\beta}\varphi(t)dt=\Phi(X^T\beta)$$

其中，$X^T=(1,x_{1i},x_{2i},\cdots,x_{pi})$，$\beta=(\beta_1,\beta_2,\cdots,\beta_p)^T$。

- **Logit模型**
  - 若 ε_i^* 服从逻辑分布函数，则其累积分布函数 $G(z)=e^z/(1+e^z)$，该模型称为Logit模型，而且

$$Prob(Y=1|X)=\frac{e^{X^T\beta}}{1+e^{X^T\beta}}$$

综述：回归分析形式

回归模型一般形式

一般形式

$$\boldsymbol{Y} = f(\boldsymbol{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}$$

其中，**Y**为因变量（或称被解释变量）（Dependent Variable），**X**为自变量（Independent Variable）（或称解释变量（Explanatory Variable），或称回归因子（Regressor）），**β**为待估参数（包括常数项和回归系数）（Parameter），ε为残差项（Residual Term），f(·)为**X**和**β**的任意函数。

- 因变量的变化包括两部分：一是由自变量变化引起的部分，其影响是主要的；二是由残差项变化引起的变化部分，其影响是次要的。
- **引起残差项变化的因素：**可能来源于被忽略的自变量、回归系数变动、变量测量误差、模型形式设定不合适以及数据生成过程的内在随机性等方面。

综述：回归分析形式（续）

- 多种形式
  - 简单线性回归（Sinple Linear Regression）
  - 多元线性回归（Multiple Linear Regression）
  - 多因变量正态回归（Multivariate Normal Regression）
  - 非线性回归（Nonlinear Regression）
  - 多项式回归（Polynomial Regression）
  - 分位数回归（Quantile Regression）
  - 广义线性回归（Generalized Linear Regression）（含Multinomial/Poisson/Logistic/Probit/Tobit等）
  - 岭回归（Ridge Regression）
  - 向量自回归（Vector Autoregression）

简要回顾：线性回归分析

模型名称		一元线性回归（SLR）	多元线性回归（MLR）	备注
模型形式				
自变量筛选				
OLS估计量				
拟合评价				
显著性检验	单个参数			
	模型整体			
预测	个体			
	均值			

相关与回归的主要问题回顾

- 1. 相关分析与回归分析的含义以及它们之间的联系与区别
- 2. 线性相关性的类型及其图示解释
- 3. 相关系数的定义、取值范围及其意义
- 4. 回归分析的基本步骤
- 5. 简单线性回归模型参数的OLS估计量及标准误形式
- 6. 简单线性回归分析的三个平方和形式及其意义
- 7. 判定系数的定义和作用
- 8. 简单线性回归模型参数的显著性检验步骤
- 9. 简单线性回归模型整体的显著性检验步骤