第九章 方差分析及回归分析

§1 单因素的方差分析
§2 双因素的方差分析
§3 一元线性回归
§4 多元线性回归

§3 一元线性回归

多个变量之间的关系：

$$\begin{cases} \text{确定性关系——函数关系} \\ \text{非确定性关系——相关关系} \end{cases}$$

回归分析（regression analysis）：研究变量之间相关关系的数学方法。即用一个或一组变量（自变量）去估计或表达另一个变量（因变量）

$$\begin{cases} \text{自变量 } x \text{ ——普通变量} \\ \text{因变量 } Y \text{ ——线性回归模型} \end{cases}$$

回归分析的任务：

(1) 建立相关关系的变量之间数学关系式

(2) 检验所建立的关系式是否有效

(3) 利用有效的关系式进行预测和控制

$$\text{回归分析}\begin{cases}\text{线性回归分析}\begin{cases}\text{一元线性回归}\\ \text{多元线性回归}\end{cases}\\ \text{非线性回归分析}\end{cases}$$

一元线性回归

普通变量x取定一组不完全相同的值$x_1, x_2, \ldots, x_n$，$Y_1, Y_2, \ldots, Y_n$分别是对应的独立观察结果，称

$$(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$$

是一个样本，对应的样本值记为：

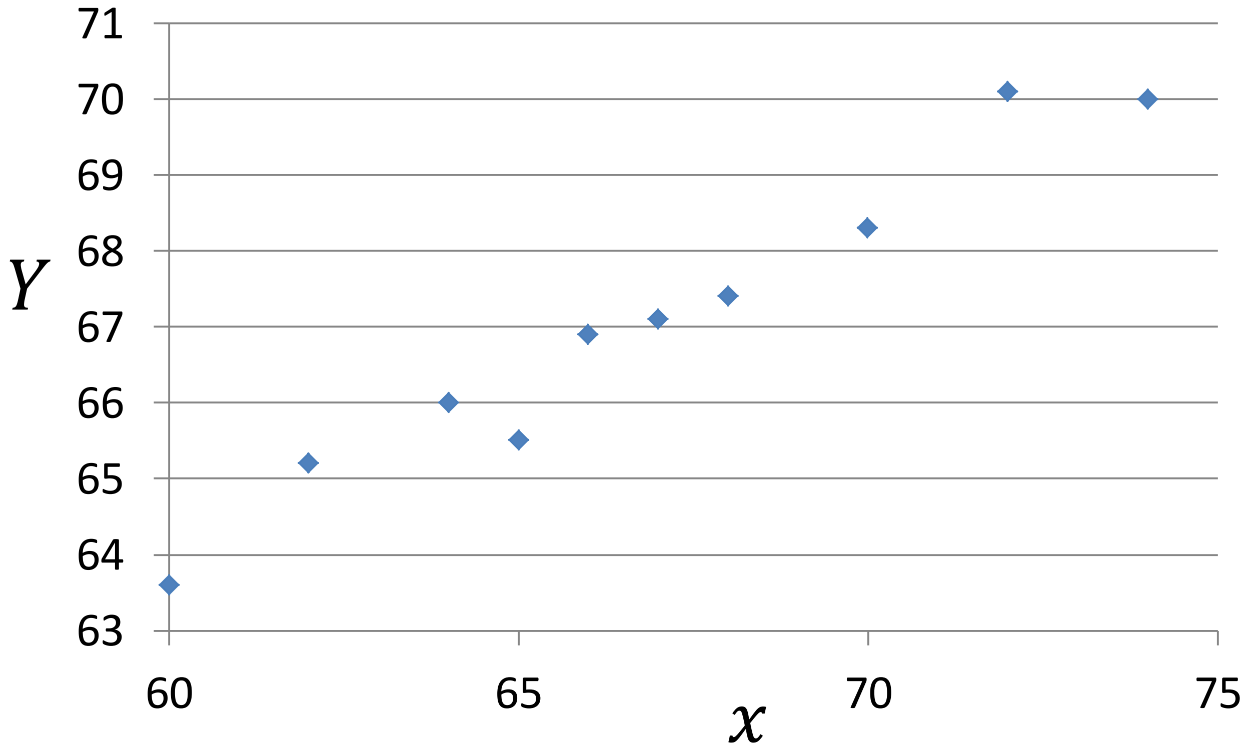
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

$\mu(x)$：Y关于x的回归函数，简称回归。

利用样本估计$\mu(x)$的问题称为求Y关于x的回归问题

例1 Francis Galton和K. Pearson收集了大量父亲身高和儿子身高的资料，其中的10对数据如下：

父亲身高(x)	60	62	64	65	66	67	68	70	72	74
儿子身高(Y)	63.6	65.2	66	65.5	66.9	67.1	67.4	68.3	70.1	70



一元线性函数：$\mu(x) = a + bx$

假设对于x的每个值，$Y \sim N(a + bx, \sigma^2)$

其中，a, b, σ^2是不依赖于x的未知参数。

记$\epsilon = Y - (a + bx)$，则有：

$$Y = \underbrace{a + bx}_{\text{线性函数}} + \underbrace{\epsilon}_{\text{随机误差}}, \qquad \epsilon \sim N(0, \sigma^2)$$

上式称为一元线性回归模型。b称为回归系数。

a,b的估计

通过n次独立试验得到样本$(x_1,Y_1),\dots,(x_n,Y_n)$，$Y_i\sim N(a+bx_i,\sigma^2),i=1,\dots,n$。

$Y_1,\dots,Y_n$相互独立，其联合概率密度(样本的似然函数)为：

$$L=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y_i-a-bx_i)^2}{2\sigma^2}}$$

$$=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{\sum_{i=1}^{n}(y_i-a-bx_i)^2}{2\sigma^2}}$$

用最大似然估计，即求下面函数的最小值：

$$Q(a,b)=\sum_{i=1}^{n}(y_i-a-bx_i)^2$$

为此，令Q对于参数a,b的偏导数等于0，得：

$$\begin{cases}\dfrac{\partial Q}{\partial a}=-2\displaystyle\sum_{i=1}^{n}(y_i-a-bx_i)=0\\ \dfrac{\partial Q}{\partial b}=-2\displaystyle\sum_{i=1}^{n}(y_i-a-bx_i)x_i=0\end{cases}$$

a, b的最大似然估计值为：

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

其中，

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

经验回归函数：$\hat{\mu}(x) = \hat{a} + \hat{b}x$

经验回归方程：$\hat{y} = \hat{a} + \hat{b}x$

其图形称为回归直线。

为了方便计算，引入下述记号：

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

这样a, b的估计值可以写成：

$$\begin{cases} \hat{a} = \bar{y} - \hat{b}\bar{x} \\ \hat{b} = \dfrac{S_{xy}}{S_{xx}} \end{cases}$$

例1的解：

由所给的数据计算得：$\bar{x} = 66.8, \bar{y} = 67.01,$

求得a, b的估计值为：

$$\begin{cases} \hat{a} = 35.977 \\ \hat{b} = 0.4646 \end{cases}$$

故儿子身高关于父亲身高的经验回归方程为：

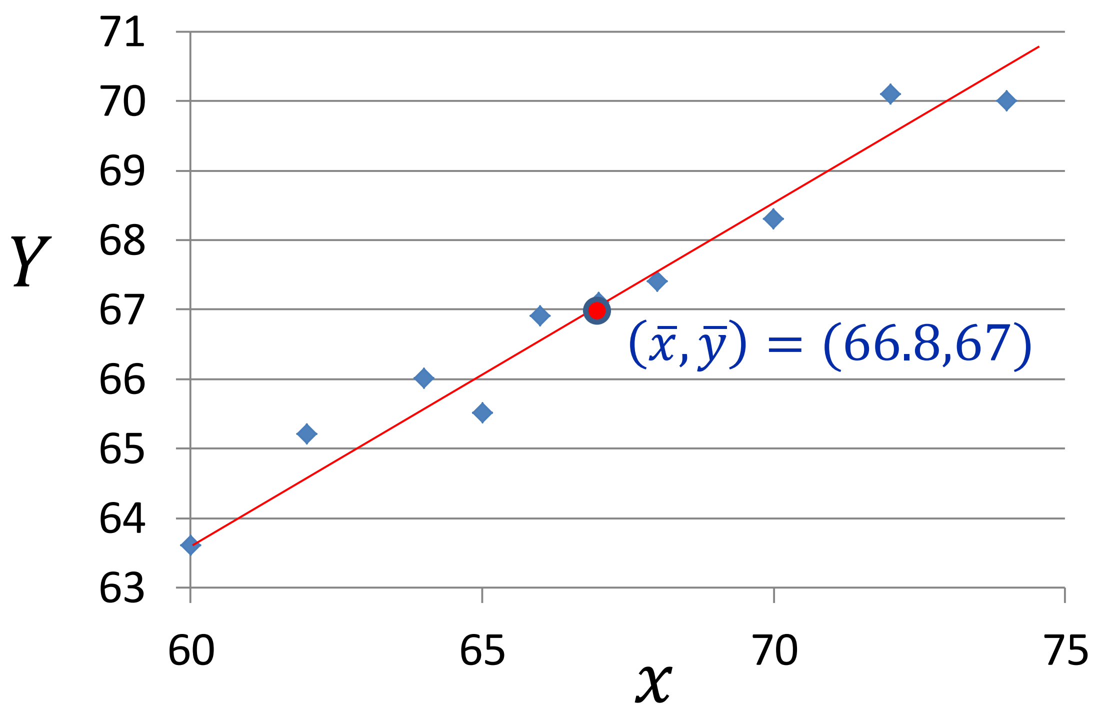
$$\hat{y} = 0.4646x + 35.977$$

回归方程可写为：

$$\hat{y} = \bar{y} + \hat{b}(x - \bar{x})$$

$$\boxed{\begin{aligned} \hat{y} &= \hat{a} + \hat{b}x \\ \hat{a} &= \bar{y} - \hat{b}\bar{x} \end{aligned}}$$

即，对于样本值 $(x_1, y_1), \ldots, (x_n, y_n)$，回归直线通过散点图的几何中心$(\bar{x},\ \bar{y})$。



备注：

- “回归”是由英国著名生物学家兼统计学家Francis Galton（1822～1911）在研究人类遗传问题时提出来的。1855年， Galton发表《遗传的身高向平均数方向的回归》
- 回归效应：当父母身高走向极端，子女的身高不会象父母身高那样极端化，其身高要比父母们的身高更接近平均身高，即有“回归”到平均数去的趋势
- 现代统计学中，回归广泛地指根据一种变量预测另一种变量或多种变量之间关系的描述方法

最小二乘估计（least squares estimation, LSE）：

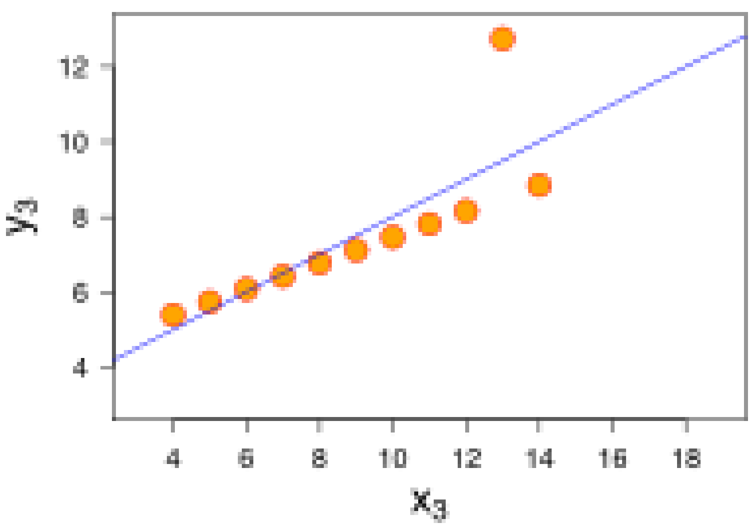
如果Y不是正态变量，可以直接最小化Y的观察值y_i和回归函数$a+bx_i$的偏差的平方和来求参数a,b，即最小化：

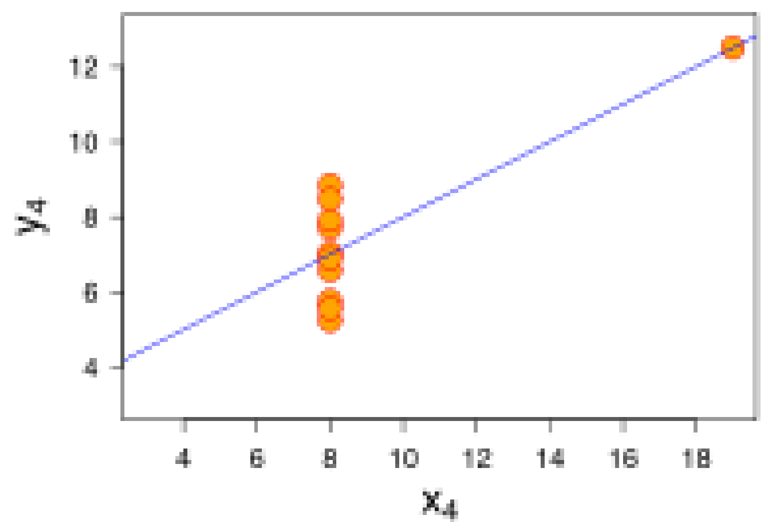
$$Q(a,b)=\sum_{i=1}^{n}(y_i-a-bx_i)^2$$

- 非常常用的求经验公式的方法！
- 当Y是正态变量时，与最大似然估计的解相同

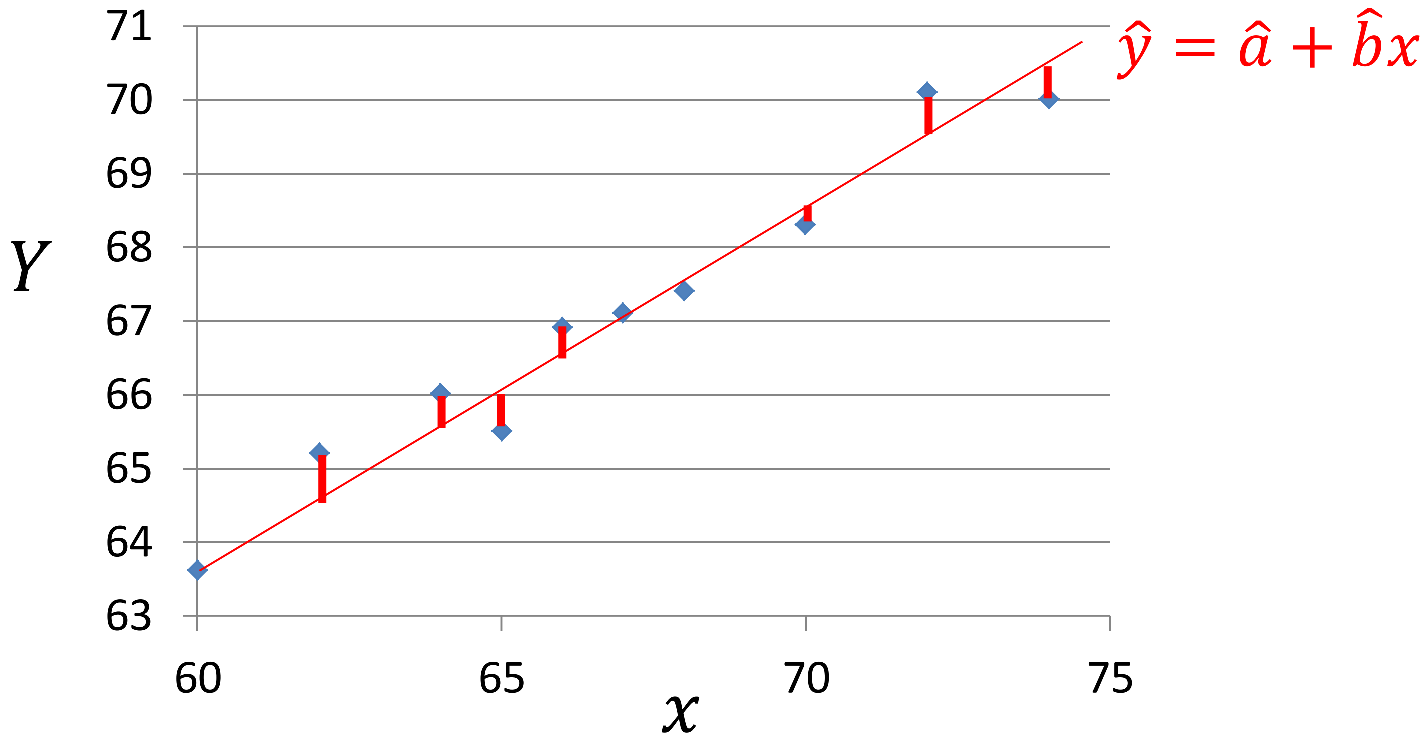
注意：

噪声点对结果的影响





- σ^2的估计



由于$\hat{y}_i = \hat{a} + \hat{b}x_i$ ，得x_i处的残差：$y_i - \hat{y}_i$

残差平方和： $Q_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \hat{a} - \hat{b}x_i\right)^2$

$$Q_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}\left(y_i-\bar{y}-\hat{b}(x_i-\bar{x})\right)^2$$

$$=\sum_{i=1}^{n}(y_i-\bar{y})^2+\left(\hat{b}\right)^2\sum_{i=1}^{n}(x_i-\bar{x})^2-2\hat{b}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$$

$$=S_{yy}+\left(\hat{b}\right)^2 S_{xx}-2\hat{b}S_{xy}$$

由 $\hat{b}=\frac{S_{xy}}{S_{xx}}$，得

$$Q_e=S_{yy}-\hat{b}S_{xy}$$

可以证明(见本章后附录)，统计量

$$\frac{Q_e}{\sigma^2}\sim\chi^2(n-2)$$

所以，

$$E\left[\frac{Q_e}{\sigma^2}\right] = n-2$$

即

$$E\left[\frac{Q_e}{n-2}\right] = \sigma^2$$

得σ^2的无偏估计量：

$$\hat{\sigma}^2 = \frac{Q_e}{n-2} = \frac{1}{n-2}(S_{YY} - \hat{b}S_{xY})$$

- 线性假设的显著性检验

$\mu(x)$是否具有形式$a+bx$?

假设检验：

$$H_0: b = 0, \qquad H_1: b \neq 0$$

可以证明：$\hat{b} \sim N(b, \sigma^2/S_{xx})$, 则

$$\frac{\hat{b}-b}{\hat{\sigma}}\sqrt{S_{xx}} \sim t(n-2)$$

H_0的拒绝域为：

$$|t| = \frac{|\hat{b}|}{\hat{\sigma}}\sqrt{S_{xx}} \geq t_{\frac{\alpha}{2}}(n-2)$$

当拒绝H_0时，认为回归效果是显著的。

回归效果不显著的原因可能有：

(1) 影响Y取值的除了x及随机误差外还有其他不可忽略的因素；

(2) $E[Y]$与 x的关系不是线性的；

(3) Y与 x不存在关系。

- **系数b的置信区间**

回归效果显著时，由

$$\frac{\hat{b}-b}{\hat{\sigma}}\sqrt{S_{xx}} \sim t(n-2)$$

可以得到b的置信水平为$1-\alpha$的置信区间：

$$\left(\hat{b} \pm t_{\frac{\alpha}{2}}(n-2) \times \frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right)$$

- 回归函数值的点估计和置信区间

x_0是自变量x的某一指定值，则$\mu(x_0) = a + bx_0$的点估计即为：

$$\hat{\mu}(x_0) = \hat{a} + \hat{b}x_0$$

考虑统计量$\hat{Y}_0=\hat{a}+\hat{b}x_0$，由于

$$\frac{\hat{Y}_0-(a+bx_0)}{\sigma\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}}\sim N(0,1),\qquad \frac{(n-2)\hat{\sigma}^2}{\sigma^2}=\frac{Q_e}{\sigma^2}\sim\chi^2(n-2)$$

于是，

$$\frac{\hat{Y}_0-(a+bx_0)}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}}\sim t(n-2)$$

所以，$\mu(x_0)=a+bx_0$的置信水平为$1-\alpha$的置信区间为：

$$\left(\hat{a}+\hat{b}x_0\pm t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1/n+(x_0-\bar{x})^2/S_{xx}}\right)$$

置信区间的长度是x_0的函数，它随$|x_0-\bar{x}|$的增加而增加。当$x_0=\bar{x}$时为最短。

- Y的观察值的点预测和预测区间

Y_0是$x = x_0$时Y的观察值，已知：

$$Y_0 = a + bx_0 + \epsilon_0, \epsilon_0 \sim N(0, \sigma^2)$$

Y_0的点预测：

$$\hat{Y}_0 = \hat{\mu}(x_0) = \hat{a} + \hat{b}x_0$$

可以证明，

$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

所以，Y_0的置信水平为$1-\alpha$的预测区间为：

$$\left(\hat{a}+\hat{b}x_0 \pm t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}\right)$$

- 预测区间的长度是x_0的函数，它随$|x_0-\bar{x}|$的增加而增加。当$x_0=\bar{x}$时为最短。
- 在相同的置信水平下，回归函数值$\mu(x_0)$的置信区间要比Y_0的预测区间短，这是因为$Y_0=a+bx_0+\epsilon_0$比$\mu(x_0)=a+bx_0$多了一项ϵ_0的缘故。

- 可化为一元线性回归的例子

通过适当的变量变换，复杂的回归问题可以化为一元线性回归来处理。

例如：

(1) $Y = ae^{\beta x} \cdot \epsilon,\ \ln \epsilon \sim N(0, \sigma^2)$

(2) $Y = ax^{\beta} \cdot \epsilon,\qquad \ln \epsilon \sim N(0, \sigma^2)$

(3) $Y = a + \beta h(x) + \epsilon,\ \epsilon \sim N(0, \sigma^2)$

$$\rightarrow Y' = a + bx' + \epsilon', \epsilon \sim N(0, \sigma^2)$$

§4 多元线性回归

随机变量Y往往与多个普通变量$x_1, x_2, \dots$，x_p $(p > 1)$有关。对于自变量$x_1, x_2, \dots$，x_p的一组确定的值，Y的数学期望$\mu(x_1, x_2, \dots$，$x_p)$即Y关于x的回归函数。

假设$\mu(\cdot)$为线性函数，即：

$$Y = b_0 + b_1 x_1 + \cdots + b_p x_p + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

其中，$b_0, b_1, \dots, b_p, \sigma^2$ 是$x_1, x_2, \dots$，x_p无关的未知参数。

- 参数的估计

设$(x_{11},\dots,x_{1p},y_1),\dots,(x_{n1},\dots,x_{np},y_n)$是$Y$和$x$的一个样本。

用最大似然估计法估计参数，即求下面函数的最小值：

$$Q(\boldsymbol{b})=\sum_{i=1}^{n}\left(y_i-b_0-b_1x_{i1}-\cdots-b_px_{ip}\right)^2$$

为此，令Q对于参数$b_0, b_1, \dots, b_p$的偏导数等于0，得：

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\displaystyle\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip}\right) = 0 \\ \dfrac{\partial Q}{\partial b_1} = -2\displaystyle\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip}\right)x_{i1} = 0 \\ \vdots \\ \dfrac{\partial Q}{\partial b_p} = -2\displaystyle\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip}\right)x_{ip} = 0 \end{cases}$$

引入矩阵表示：

$$\boldsymbol{X}=\begin{bmatrix}1 & x_{11} & \cdots & x_{1p}\\ 1 & x_{21} & \cdots & x_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & \cdots & x_{np}\end{bmatrix},\boldsymbol{Y}=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix},\boldsymbol{B}=\begin{bmatrix}b_0\\ b_1\\ \vdots\\ b_p\end{bmatrix}$$

所以，有正规方程组：

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}=\boldsymbol{X}^T\boldsymbol{Y}$$

$b_0,b_1,\ldots,b_p$的最大似然估计：

$$\widehat{\boldsymbol{B}}=\begin{bmatrix}\hat{b}_0\\ \hat{b}_1\\ \vdots\\ \hat{b}_p\end{bmatrix}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

取：

$$\hat{b}_0 + \hat{b}_1 x_1 + \cdots + \hat{b}_p x_p \stackrel{\text{def}}{=} \hat{y}$$

$\mu(x_1, x_2, \ldots, x_p)$的估计为：

$$\hat{\mu}(x) = \hat{b}_0 + \hat{b}_1 x_1 + \cdots + \hat{b}_p x_p$$

此式称为p元经验回归函数。

p元经验回归方程：

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \cdots + \hat{b}_p x_p$$

例 4 某种产品每件平均价格Y(元)与批量x(件)之间的关系如下表：

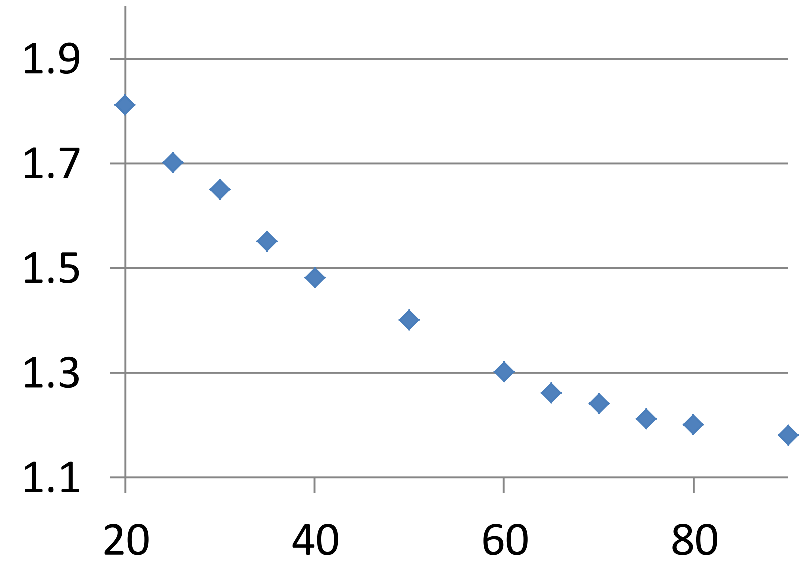
x	20	25	30	35	40	50	60	65	70	75	80	90
y	1.81	1.70	1.65	1.55	1.48	1.40	1.30	1.26	1.24	1.21	1.20	1.18

散点图如下，我们选取模型：

$$Y = b_0 + b_1 x + b_2 x^2 + \epsilon,$$

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

来拟合，求回归方程？



解：令$x_1 = x, x_2 = x^2$,模型可以转化为二元线性回归模型：

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

由样本值：

$$\boldsymbol{X} = \begin{bmatrix} 1 & 20 & 400 \\ 1 & 25 & 625 \\ \vdots & \vdots & \vdots \\ 1 & 90 & 8100 \end{bmatrix}, \boldsymbol{Y} = \begin{bmatrix} 1.81 \\ 1.70 \\ \vdots \\ 1.18 \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

得参数$\boldsymbol{B}$的估计为：

$$\widehat{\boldsymbol{B}} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} = \begin{bmatrix} 2.19826629 \\ -0.02252236 \\ 0.00012507 \end{bmatrix}$$

故回归方程为：

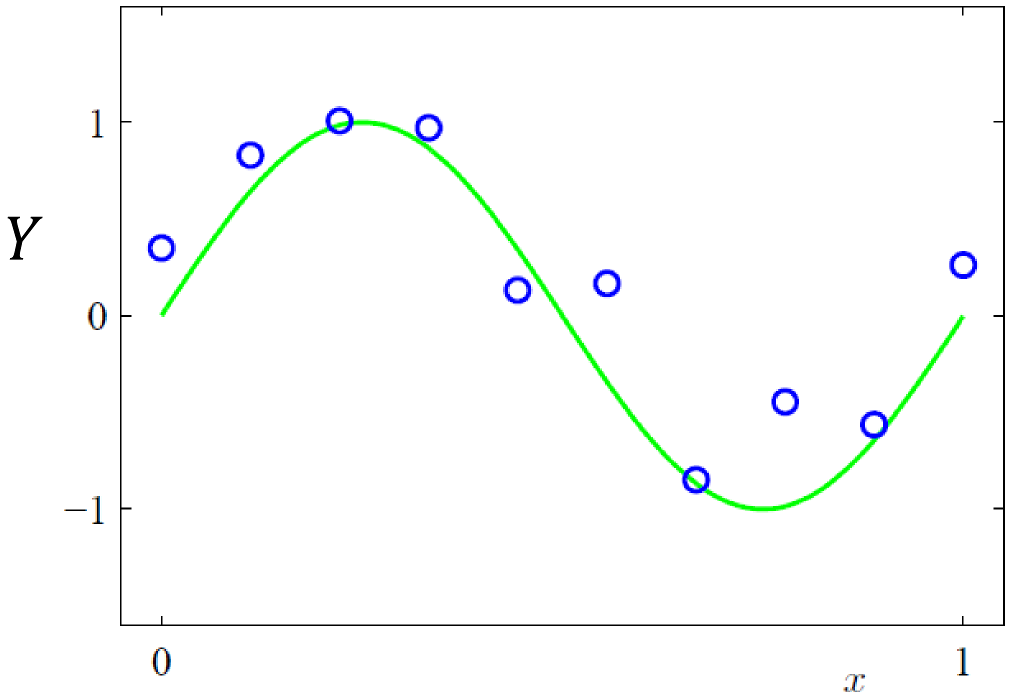
$$\hat{y} = 2.19826629 - 0.02252236 x_1 + 0.00012507 x_2$$

- 可化为多元线性回归的例子

$$Y = b_0 + b_1\phi_1(x) + \cdots + b_M\phi_M(x) + \epsilon$$
$$= \sum_{i=0}^{M} b_i\phi_i(x) + \epsilon = \boldsymbol{b}^T\boldsymbol{\phi}(x) + \epsilon$$

- $\boldsymbol{b}^T = [b_0, b_1, \dots, b_M]$
- $\boldsymbol{\phi}^T = [\phi_0, \phi_1, \dots, \phi_M]$
- 基函数可以是非线性的，例如：
  - ✓ $\boldsymbol{\phi}^T = [1, x, x^2, \dots, x^M]$
  - ✓ $\phi_i = e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}$

例 3 非线性回归问题



$\sin(2\pi x)$

数据产生方式：$Y = \sin(2\pi x) + \epsilon,\ n = 10$

试估计样本的回归函数？

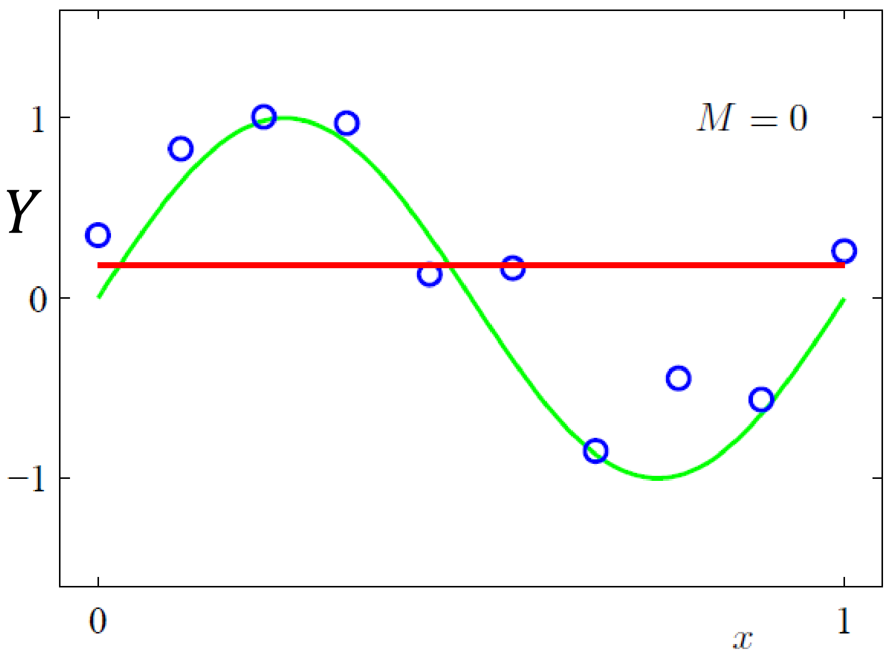
解：

建立线性回归模型：

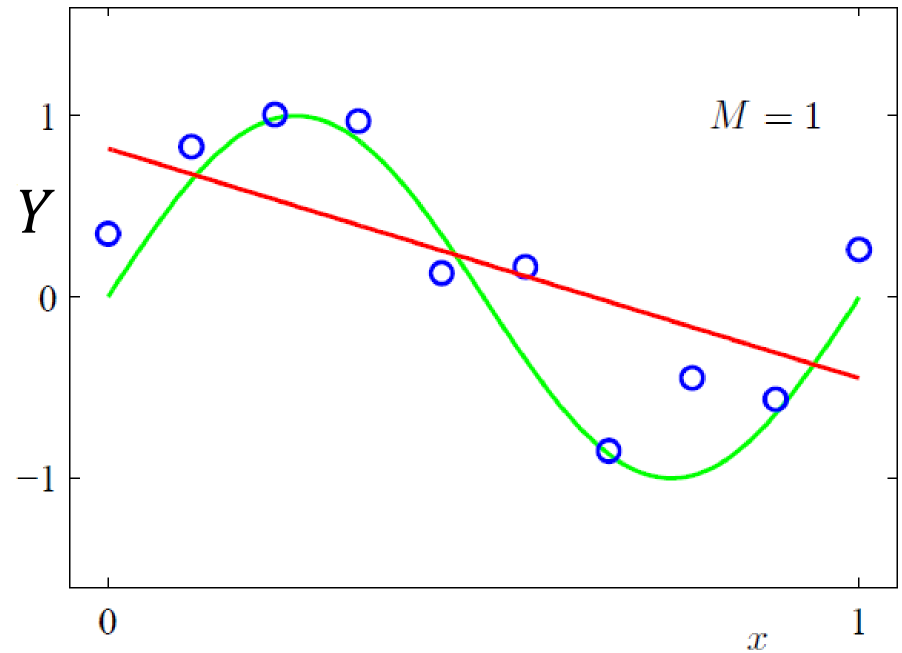
$$\mu(x) = b_0 + b_1 x + \cdots + b_M x^M(x) = \sum_{i=0}^{M} b_i x^i$$

- 回归函数$\mu(x)$对于x是非线性的，而对于参数$\boldsymbol{b}^T = [b_0, b_1, \ldots, b_M]$是线性的。

回归曲线：

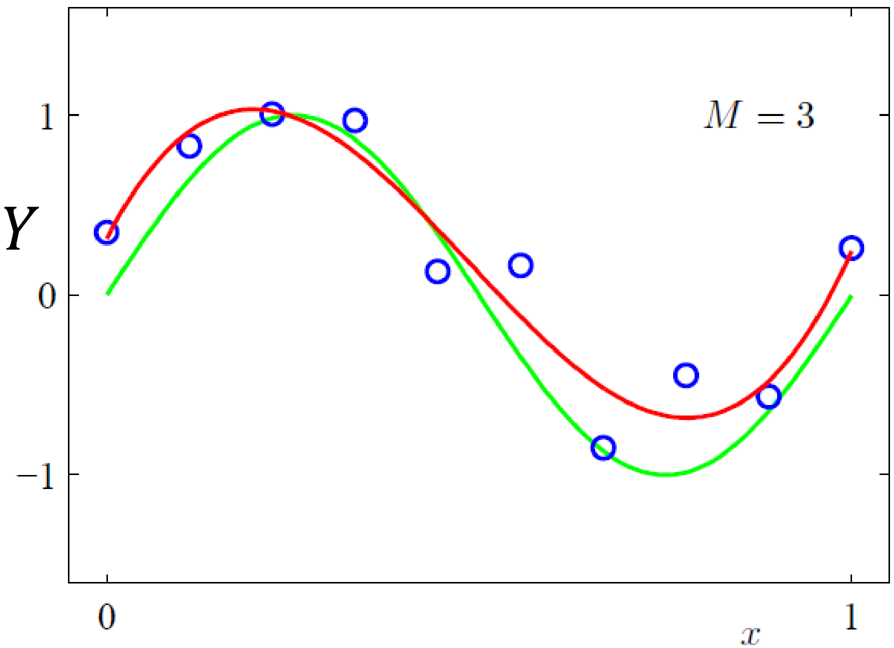


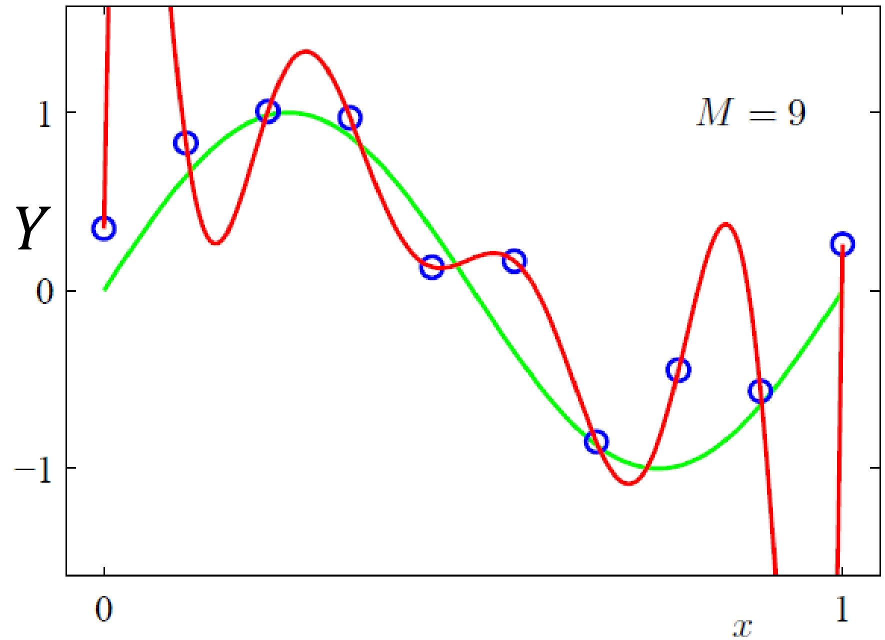
$$\mu(x) = b_0$$



$$\mu(x) = b_0 + b_1 x$$

回归曲线：



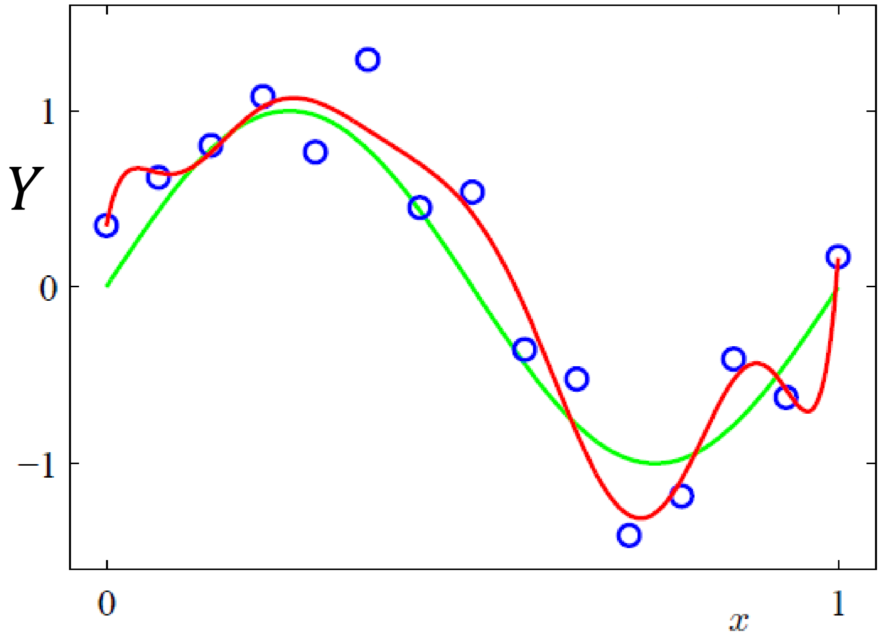


$$\mu(x)=b_0+b_1x+b_2x^2+b_3x^3$$

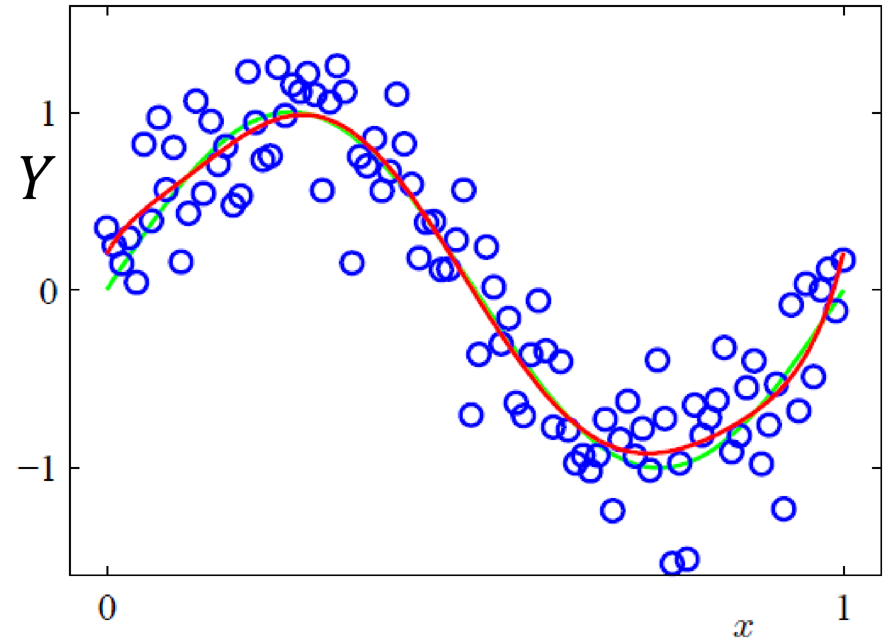
$$\mu(x)=b_0+b_1x+\cdots+b_9x^9$$

过拟合（overfitting）！

拟合结果与样本数量的关系：



$M=9, n=15$



$M=9, n=100$

小结——回归分析

1. 建立线性回归函数$\mu(x)=a+bx$，利用样本值$(x_1,y_1),\ldots,(x_n,y_n)$计算$a,b$的最大似然估计：
$$\hat{\mu}=\hat{a}+\hat{b}x$$
相应的回归方程为：$\hat{y}=\hat{a}+\hat{b}x$。

2. 误差ϵ的无偏估计：$\hat{\sigma}^2=\frac{Q_e}{n-2}=\frac{S_{yy}-\hat{b}S_{xy}}{n-2}$

3. 显著性检验：$H_0: b=0,\ H_1: b\neq 0$，拒绝域为：
$$|t|=\frac{|\hat{b}|}{\hat{\sigma}}\sqrt{S_{xx}}\geq t_{\frac{\alpha}{2}}(n-2)$$

4. 回归系数b的置信水平为$1-\alpha$的置信区间为：

$$\left(\hat{b} \pm t_{\frac{\alpha}{2}}(n-2) \times \frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right)$$

5. 回归函数$\mu(x)$在点x_0处的函数值$\mu(x_0)$的置信水平为$1-\alpha$的置信区间为：

$$\left(\hat{a} + \hat{b}x_0 \pm t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}\right)$$

6. Y在x_0处的观察值$Y_0 = a + bx_0 + \epsilon_0$的预测值为：$\hat{y}_0 = \hat{a} + \hat{b}x_0$，$Y_0$的置信水平为$1 - \alpha$的置信区间为：

$$\left(\hat{a} + \hat{b}x_0 \pm t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right)$$

小结——方差分析

1. 建立数学模型：$\begin{cases} X_{ij} = \mu + \delta_j + \epsilon_{ij} \\ \epsilon_{ij} \sim N(0, \sigma^2), \text{各}\epsilon_{ij}\text{独立} \\ \sum_{j=1}^{s} n_j \delta_j = 0 \\ i = 1, 2, \dots, n_j, j = 1,2, \dots, s \end{cases}$

2. 假设检验$\begin{cases} H_0\text{: } \delta_1 = \delta_2 = \cdots = \delta_s = 0 \\ H_1\text{: } \delta_1, \delta_2, \dots, \delta_s\text{不全为}0 \end{cases}$的拒绝域为：

$$F = \frac{S_A/(s-1)}{S_E/(n-s)} \geq F_\alpha(s-1, n-s)$$

方差来源	平方和	自由度	均方	F比
因素A	S_A	$s-1$	$\bar{S}_A = S_A/(s-1)$	$\bar{S}_A/\bar{S}_E$
误差	S_E	$n-s$	$\bar{S}_E = S_E/(n-s)$	
总和	S_T	$n-1$		

3. 参数估计：

σ^2	μ	μ_j	δ_j
$\widehat{\sigma^2} = \frac{S_E}{n-s}$	$\bar{X}$	$\bar{X}_{.j}$	$\widehat{\delta_j} = \bar{X}_{.j} - \bar{X}$

4. 当拒绝H_0时，总体$N(\mu_j, \sigma^2)$和$N(\mu_k, \sigma^2)$的均值差$\mu_j - \mu_k = \delta_j - \delta_k$的置信区间为：

$$\left(\bar{X}_{.j} - \bar{X}_{.k} \pm t_{\alpha/2}(n-s)\sqrt{\bar{S}_E\left(\frac{1}{n_j} + \frac{1}{n_k}\right)}\right)$$

方差分析与回归分析的比较：

	方差分析 （analysis of variance, ANOVA）	回归分析 （ regression analysis）
相同	• 采用线性模型 • 研究变量之间的关系	
不同	• 从观测变量的方差入手，研究一个或多个变量对特定变量的影响以及组间的差异	• 确定两种或两种以上变量间相互依赖的定量关系
	• 因素变量不一定是定量的	• 因素变量是定量的
	• 只需有选择的对某些试验水平进行试验	• 所有试验水平都进行相应的试验

作 业

- 概率论与数理统计

  pp. 267-269, #9, #11, #16