第5章 线性回归模型

第5.1节 线性模型理论

第5.2节 一元回归与相关分析

第5.3节 多元回归分析

东北大学数学系

第5.1节 线性模型理论

5.1.1 线性模型的定义

定义5.1.1 y 是可观察的随机变量，x_1，…，x_m 是可观察的分类或数值变量，β_0，…，β_k 是未知参数，ε 是不可观察随机误差($\varepsilon \sim N(0, \sigma^2)$)。

$$y = \beta_0 + \sum_{i=1}^{k} f_i(x_1, \cdots, x_m)\beta_i + \varepsilon$$

称为是线性模型。

Remark

① 线性模型中“线性”是针对未知参数 β 而言，
许多表面上的非线性模型本质也是线性的:

$y = \alpha e^{\beta x} \times \varepsilon$， $y = \alpha x^{\beta} \times \varepsilon$， $\ln \varepsilon \sim N(0, \sigma^2)$ ；

而有些模型是实质上的非线性模型：

$$y = \alpha e^{\beta x} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \ ;$$

以及 *Logistic* 模型：

$$y = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} + \varepsilon$$

② 一些统计学家喜欢把线性模型表示成：

$$E\,y = \beta_0 + x_1\beta_1 + \ldots + x_k\beta_k$$

含义是：线性模型就是一个随机变量的数学期望具有未知参数线性结构的统计模型。

在这种意义下 x_1，…，x_k 很自然就被称为“自变量”，y 也就被称作“因变量”。

自变量与因变量关系是一种统计上的关系，即因变量的均值是自变量的函数，而决不能认为因变量是自变量的函数。

③ 把模型 $E\,y = \beta_0 + x_1\beta_1 + \ldots + x_k\beta_k$ 改写成：

$$y = \beta_0 + x_1\beta_1 + \ldots + x_k\beta_k + \varepsilon\,,\ E\,\varepsilon = 0$$

为方便处理，进一步假定 $\varepsilon \sim N(0, \sigma^2)$。

要处理这 $k+1$ 个未知参数 β_0，…，β_k，需要至少做 n 次独立试验 $(n > k+1)$；这些试验都在不同自变量取值下进行，其它条件都保持不变，最后写成矩阵的表达式，就是：

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon\,,\ E\,\varepsilon = 0$$

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon\ ,\ E\,\varepsilon = 0$$

这里$\mathbf{Y} = (y_1, \ldots, y_n)^{\mathrm{T}}$表示可观察的因变量；

$\beta = (\beta_0, \ldots, \beta_k)^{\mathrm{T}}$表示待估计或检验的未知参数；

$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$是随机误差，一般假定$\varepsilon_i \sim N(0, \sigma^2)$。

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \cdot & \cdot & \cdots & \cdot \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

表示自变量，

$n > k+1$

5.1.2 线性模型参数的估计

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

1. 未知参数 β 的估计

采用最小二乘的标准，即寻找 $\hat{\beta}$，使得：

$$\|\mathbf{Y}-\mathbf{X}\hat{\beta}\|^2 = \inf \|\mathbf{Y}-\mathbf{X}\beta\|^2 \text{ 对所有 } \beta \in \mathbf{R}^{k+1}$$

这样得到的 β 的估计称为是最小二乘估计(*LSE*)

LSE 的求解思路：<u>平方和分解</u>

$$\|Y-X\beta\|^2 = \|Y-X\hat{\beta} + X(\hat{\beta}-\beta)\|^2$$

$$= \|Y-X\hat{\beta}\|^2 + \|X(\hat{\beta}-\beta)\|^2 + 2(\hat{\beta}-\beta)^T X^T(Y-X\hat{\beta})$$

因此要使得对一切 $\beta \in R^{k+1}$ 都有

$\|Y-X\beta\|^2 \geq \|Y-X\hat{\beta}\|^2$，充分必要条件是：

$(\hat{\beta}-\beta)^T X^T(Y-X\hat{\beta}) = 0$ 对一切 $\beta \in R^{k+1}$ 都成立

由于 β 是 R^{k+1} 中任意一个向量，

所以 $X^T(Y-X\hat{\beta})$ 必须是一个 $k+1$ 维零向量，即：

$$(X^TX)\hat{\beta} = X^TY \quad \text{(正规方程)}$$

如果X 是满秩矩阵即 $rk(X) = k+1$ 时，正规方程的解：

$$\hat{\beta} = (X^TX)^{-1}X^TY \equiv S^{-1}X^TY$$

就称为是线性模型 $Y = X\beta + \varepsilon$

中参数向量 β 的<u>最小二乘估计</u> (*LSE*)

$X\hat{\beta}$ 称为是<u>经验回归函数</u>，

$Y = X\hat{\beta}$是<u>经验回归方程</u>。

2. 误差方差σ^2的估计

把线性模型 $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ 改写成如下形式：

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad 1 \le i \le n$$

定义“残差” (*Residual*)

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}, \quad 1 \le i \le n$$

作为随机误差ε_i的“估计”，则残差平方和：

$$Q_e = e_1^2 + e_2^2 + \ldots + e_n^2$$

$$= \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 = \mathbf{Y}^T(\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^T)\mathbf{Y}$$

可以作为 σ^2 的估计(注意需要修正！)

定理5.1.1　线性模型的最小二乘估计

(1) 对于模型 $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$，$\beta$ 的*LSE* 是

$$\hat{\beta} = \mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

(2) σ^2的*LSE*是

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\mathbf{Y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{Y}$$

思考

既然n 个观察数据 y_1，…，y_n 的方差都是σ^2，为什么不使用这组数据的样本方差,而是要用残差平方和修正以后去估计σ^2？

最小二乘估计的矩阵代数含义

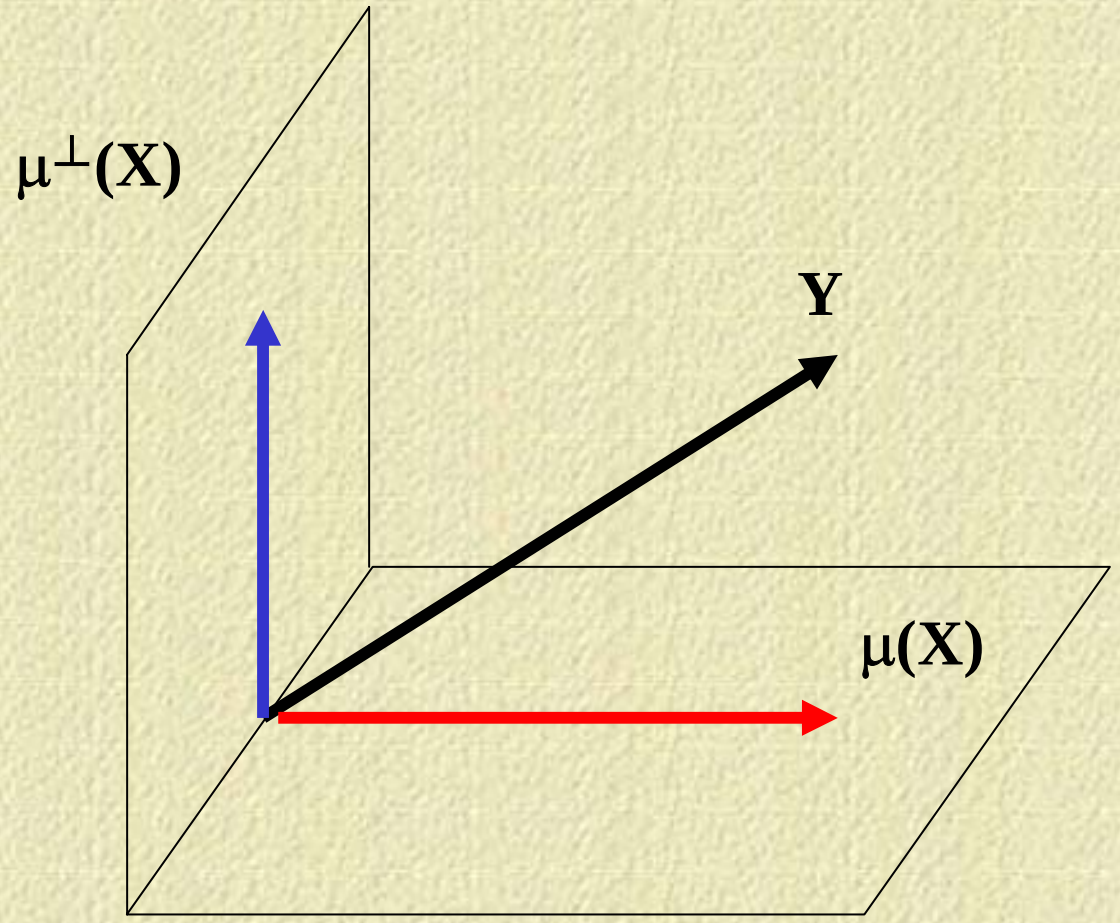
考虑矩阵 $\mathbf{X}$ 的$k+1$ 个n 维列向量生成的$\mathbf{R}^n$ 中的线性子空间 $\mu(\mathbf{X})$，不难证明 $\mu(\mathbf{X}) = \mu(\mathbf{X}\mathbf{X}^{\mathrm{T}})$。

由于$\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}$ 是一个对称、幂等的n 阶方阵，即它是一个正投影阵，恰好是$\mu(\mathbf{X})$的投影矩阵；

而 $\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}$ 是 $\mu(\mathbf{X})$ 的正交子空间 $\mu^{\perp}(\mathbf{X})$ 的投影矩阵，因此

$\mathbf{X}\hat{\beta}$ 是 $\mathbf{Y}$ 到子空间 $\mu(\mathbf{X})$ 中的投影，

σ^2的LSE只和 $\mathbf{Y}$ 在 $\mu^{\perp}(\mathbf{X})$ 的投影向量有关。


$\mu^{\perp}(X)$
Y
$\mu(X)$

3. 最小二乘估计的无偏性质

引理5.1.2　随机向量的期望与方差公式

(1) 如果Y 是*n* 维随机向量，A 是*n* 阶对称矩阵

则 $E(\mathbf{Y}^{\mathrm{T}}\mathbf{A}\mathbf{Y}) = (E\mathbf{Y})^{\mathrm{T}}\mathbf{A}(E\mathbf{Y}) + \mathrm{tr}\{\mathbf{A}[\mathrm{Var}(\mathbf{Y})]\}$；

(2) 如果Y 是*n* 维随机向量，B 是$m\times n$ 阶矩阵

则 $\mathrm{Var}(\mathbf{B}\mathbf{Y}) = \mathbf{B}[\mathrm{Var}(\mathbf{Y})]\mathbf{B}^{\mathrm{T}}$

Remark

Var (Y)是Y 的协方差矩阵 ($Cov(y_i, y_j)$)$_{n\times n}$ ；

迹 (*Trace*)具有如下性质：

$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$，$\mathrm{tr}(\mathbf{A}-\mathbf{B}) = \mathrm{tr}(\mathbf{A}) - \mathrm{tr}(\mathbf{B})$

注意到线性模型的形式 $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$，因此

$$E\mathbf{Y} = \mathbf{X}\beta，\ \mathrm{Var}(\mathbf{Y}) = \sigma^2\mathbf{I}_n$$

说明 β 的*LSE* $\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ 是无偏估计。

根据引理5.1.2，残差平方和的数学期望是：

$$E(Q_e) = E[\mathbf{Y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{Y}]$$

$$= \beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{X}\beta + \mathrm{tr}[(\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}})\sigma^2\mathbf{I}_n]$$

$$= 0 + \sigma^2\,\mathrm{tr}(\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}) = \sigma^2[n - \mathrm{tr}(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}})]$$

$$= \sigma^2[n - \mathrm{tr}(\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X})] = (n - k - 1)\sigma^2 。$$

5.1.3 估计量的分布

对于线性模型 $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$，$\beta \in \mathbf{R}^{k+1}$，

$\mathbf{X}$ 是$n \times (k+1)$ 满秩矩阵，$\varepsilon \sim N(0, \sigma^2 \mathbf{I}_n)$

定理5.1.3

(1) β 的最小二乘估计服从 $k+1$ 维正态分布，

$$\hat{\beta} = \mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \sim N(\beta, \sigma^2 \mathbf{S}^{-1})\ ;$$

(2) σ^2的估计量服从卡方分布，即

$$\frac{n-k-1}{\sigma^2}\hat{\sigma}^2 = \frac{1}{\sigma^2}\mathbf{Y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{Y} \sim \chi^2(n-k-1)\ ;$$

(3) $\hat{\beta}$ 与$\hat{\sigma}^2$相互独立。

证明.

首先，根据多维正态分布的一个性质：

如果$\mathbf{Y}$ 服从 n 维正态分布 $\sim N(\mu, \Sigma)$，$\mathbf{A}$ 是任意一个 $m \times n$ 矩阵，则$\mathbf{AY} \sim N(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^{T})$

因为 $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$， $\varepsilon \sim N(0, \sigma^2\mathbf{I}_n)$，得到 $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$，现在已知 $\hat{\beta} = \mathbf{S}^{-1}\mathbf{X}^{T}\mathbf{Y}$，而 $\mathbf{S}$ 是一个对称矩阵，因此显然有β 的*LSE*

$$\hat{\beta} \sim N(\mathbf{S}^{-1}\mathbf{X}^{T}\mathbf{X}\beta, \mathbf{S}^{-1}\mathbf{X}^{T}(\sigma^2\mathbf{I}_n)\mathbf{X}\mathbf{S}^{-1})$$

$$\sim N(\beta, \sigma^2\mathbf{S}^{-1})\ ;$$

其次，注意到表达式：

$$\hat{\beta} = S^{-1}X^{T}Y = \beta + S^{-1}X^{T}\varepsilon ,$$

$$\hat{\sigma}^{2} = \frac{1}{n-k-1} Y^{T}(I_{n} - XS^{-1}X^{T})Y$$

$$= \frac{1}{n-k-1} \varepsilon^{T}(I_{n} - XS^{-1}X^{T}) \varepsilon$$

所以，

$$[\frac{X(\hat{\beta}-\beta)}{\sigma}]^{T}[\frac{X(\hat{\beta}-\beta)}{\sigma}] = (\frac{\varepsilon}{\sigma})^{T} XS^{-1}X^{T} (\frac{\varepsilon}{\sigma})$$，以及

$$(\frac{n-k-1}{\sigma^{2}})\hat{\sigma}^{2} = (\frac{\varepsilon}{\sigma})^{T}(I_{n} - XS^{-1}X^{T}) (\frac{\varepsilon}{\sigma})$$

都是 n 个标准正态的二次型，根据Cochren定理，定理5.1.3 成立。

练习5.1.1

对于线性模型：

$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$，　其中 $\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \cdot & \cdot & \cdots & \cdot \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$

定义如下平方和：

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

$$\mathrm{RegSS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

证明必然成立分解关系　$\mathrm{TSS} = \mathrm{RegSS} + \mathrm{RSS}$

第5.2节 一元回归与相关分析

回归与相关分析是用于讨论数值变量之间关系的统计分析方法。

回归分析研究一个(或多个)自变量的变化如何影响因变量，

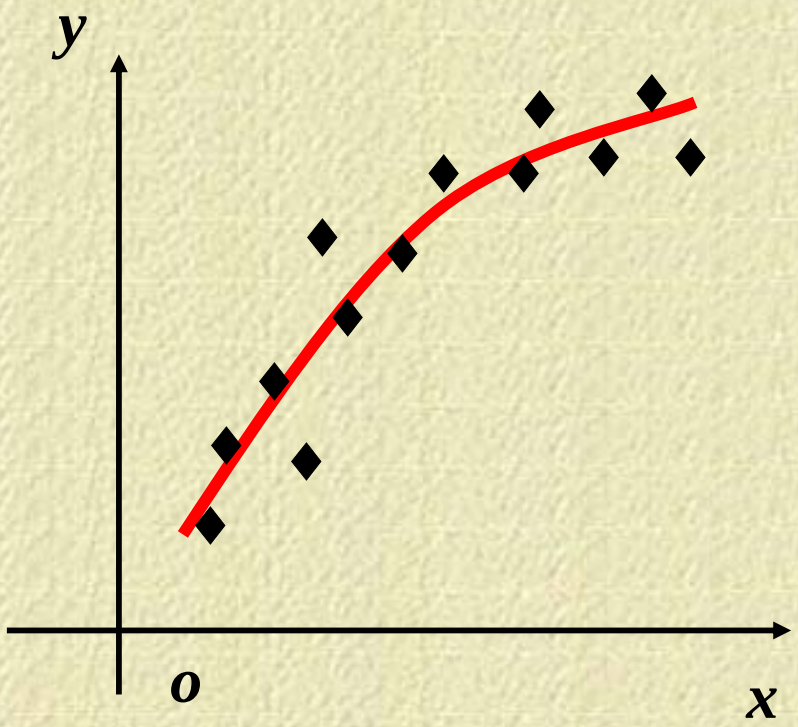
相关分析研究这两个数值变量的相关程度。

Regression

$y = 33.73 + 0.516\,x$　(单位：英寸)

“回归”的含义

直观上在一个总体中有两个特征(X, Y)，观察了n次得到平面上的n个点(x_1, y_1)，…，(x_n, y_n)。



如果一条曲线$y = f(x)$基本上通过这些点，或者这些点的大多数与这条曲线偏离很小，则称曲线是对观察值的拟合曲线，或者称为是y对于x的回归曲线。

在理论上，假定(X,Y)有联合分布，二阶矩存在，则当X取某个值x时Y有一个确定的条件分布$F(\cdot|x)$，这个分布的数学期望即条件期望$E(Y|x)$存在，$E(Y|x)$就称为Y对于x的回归(函数)

如果X是一维随机变量，则$E(Y|x)$就称为一元回归函数(主要是回归直线)；

当X是多维随机变量时就是多元回归(曲面)

Remark

采用条件期望$E(Y|x)$而不是其它的函数$y=g(x)$作为Y对于x的回归，原因是在均方误差的意义下条件期望是最优的。

如果 $E(Y|x)$ 就是 x 的线性函数，即：

$$E(Y|x) = \beta_0 + x_1\beta_1 + \ldots + x_k\beta_k ,$$

线性回归模型就定义成：

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i , \quad 1 \le i \le n$$

ε_i 独立同分布于 $N(0, \sigma^2)$

这时不再把 x 看成是随机变量 X 的观察值，而看成是一般的数量变量，因此线性回归模型也是一种线性模型：$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ ， $E\,\varepsilon = 0$

$y = \beta_0 + x_1\beta_1 + \ldots + x_k\beta_k$ 就称为是回归方程

5.2.1 一元线性回归模型

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i ,\ 1 \le i \le n$$

未知参数 β_0、β_1、σ^2 的估计以及估计量的性质根据定理5.1.1、定理5.1.3 决定。

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{L_{xy}}{L_{xx}}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}(L_{yy} - \hat{\beta}_1 L_{xy})$$

EXCEL 计算回归方程

自变量 X	x_1	x_2	…	x_n
因变量 Y	y_1	y_2	…	y_n

建立回归方程： y = 截距 + 斜率×x

截距 = ***INTERCEPT***(因变量数据, 自变量数据)

斜率 = ***SLOPE*** (因变量数据, 自变量数据)

$\hat{\sigma}$ = ***STEYX*** (因变量数据, 自变量数据)

或者直接使用函数***LINEST*** 得到回归方程

例5.2.1　食物中脂肪(克) 与所含热量(卡)的关系

随机选取16种食品，以脂肪含量作自变量 x ，因变量 y 是热量。

x	4	6	6	8	19	11	12	12	26
y	110	120	120	164	430	192	175	236	429
x	21	11	16	14	9	9	5		
y	318	249	281	160	147	210	120		

根据这组样本数据讨论热量与脂肪含量的关系。

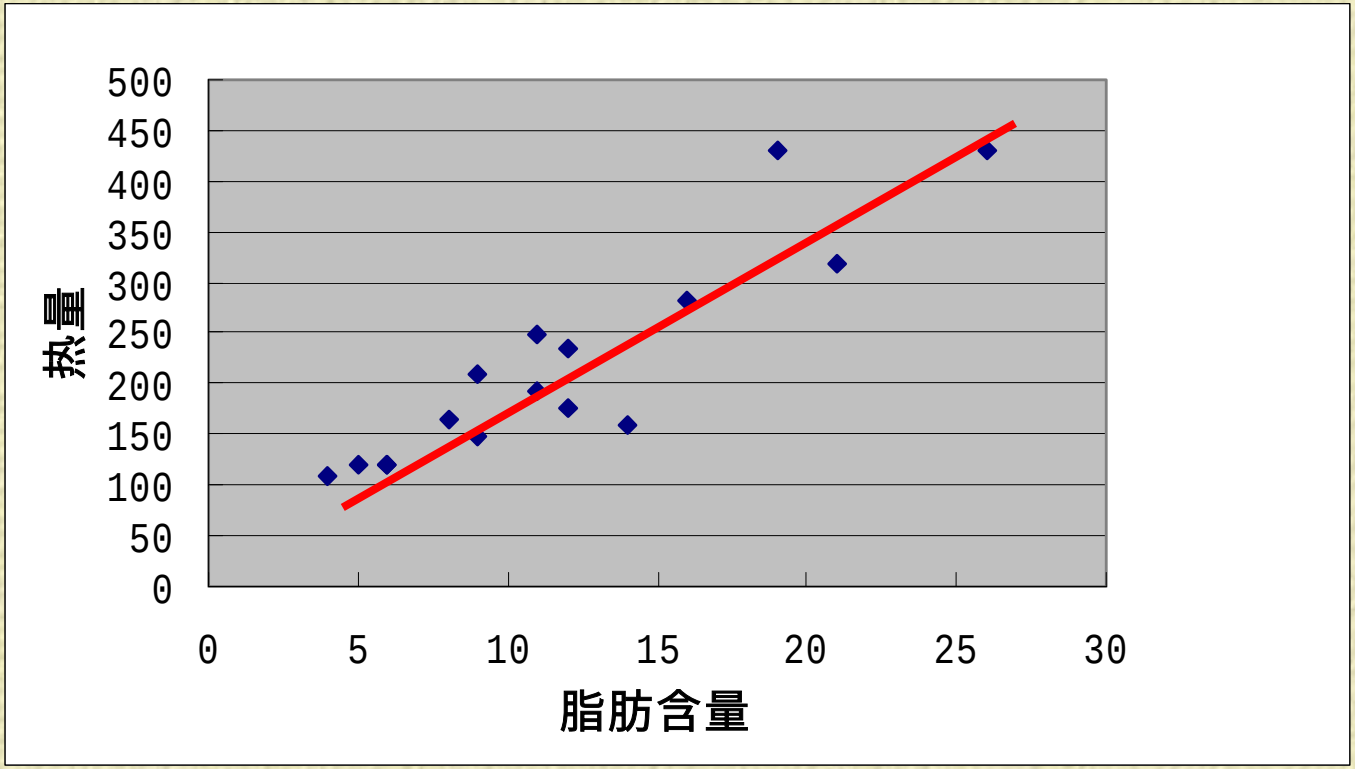
两个数值变量是否具有线性关系？

(1). 散点图的作用

对两个变量做统计分析之前首先要看散点图

如果具有线性关系，那么散点图中的点相应地应该落在某一条直线的附近。

如果散点图没有表示出线性关系，就不能直接做回归分析。


热量
500
450
400
350
300
250
200
150
100
50
0
0
5
10
15
20
25
30
脂肪含量

(2). 正相关与负相关

两个数值变量具有正相关关系，是指因变量将随着自变量的增加而增加，因此对应的直线从左下角到右上角(斜率为正)。

同理负相关是指因变量将随着自变量的增加而减小，对应直线从左上角到右下角(斜率为负)。

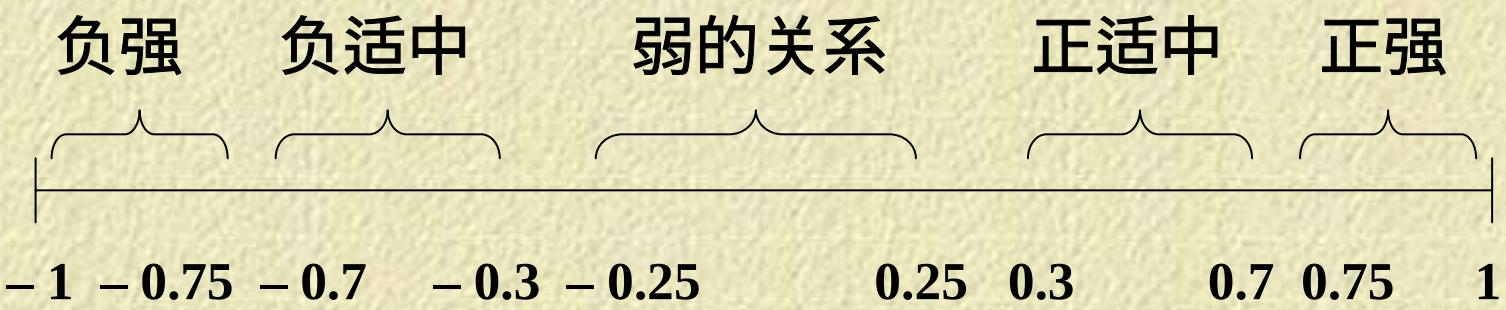
例题数据显示热量与脂肪含量具有正相关关系

两个变量的关系如何量化？

(1). 关系的强度

通过计算相关系数 *r* 来讨论

r 是介于 - 1 到 1 之间的小数，一般认为



EXCEL 计算相关系数

***r = CORREL* (自变量数据, 因变量数据)**

解析表达式：

$$CORREL = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

(2). 异常点的处理

如果有 5% 的异常点，计算出的相关系数将和去掉异常点后得出的结果有显著差异。

因此根据公式计算相关系数时，应该注意散点图中散点的分布。

(3). 相关系数 r 的确切含义

在产生因变量变化的所有因素中，自变量占据了其中$r^2 \times 100\%$ 的份额

脂肪与热量的相关系数 $r = 0.91$， $r^2 = 0.83$ □

例5.2.2　成年女性身高与腿长的关系

腿长关于身高的回归方程为：

腿长 ＝ -16.073 ＋0.7194 身高；

反之，身高关于腿长的回归方程为：

身高 ＝ 31.7713 ＋1.2903 腿长。

□

5.2.2 简单的相关分析

回归分析的平方和分解

对于 $y = \beta_0 + \beta_1 x + \varepsilon$ 甚至更一般的回归模型，根据练习5.1.1显然都有：

TSS = RegSS + RSS

TSS：总(变差)平方和，因变量 y 在其均值附近总的变化；

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

RegSS：回归平方和，
自变量 x 所引起的
因变量 y 的变化；

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

RSS：残差平方和，
随机误差所引起的
因变量 y 的变化；

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

因为 RegSS / TSS 刻画了回归平方和在总平方和所占的比重，它越大也就说明自变量x对于因变量 y 的影响越大，即回归关系越显著。

所以相关系数 r (r^2 = RegSS / TSS)
刻画了 y 与 x 线性关系程度的大小

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_j - \bar{y})^2}}$$

很容易证明这里的 r^2 就是 $L_x{}^2{}_y / L_{xx} L_{yy}$

5.2.3 回归方程的检验与区间估计

1. 回归系数的假设检验

根据样本数据建立了两个数值变量间的回归方程后，首先应该检验这个方程是否成立，即利用假设检验讨论是否有 $\mathbf{H_0}$： $\beta_1 = 0$ ？

这个已经建立的(线性)回归方程的好坏，取决于相应的假设检验的 p-值。

根据定理5.1.3，估计量的分布为

(1) $\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n}+\frac{\bar{x}^2}{L_{xx}}))$

(2) $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$

(3) $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 不独立，协方差为

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{L_{xx}}$$

(4) $\hat{\sigma}^2$ 与 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 都独立，并且

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

要检验回归关系是否显著，可以利用t 分布：

$$\frac{\hat{\beta}_1}{\hat{\sigma}}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sim t(n-2)$$

更多的是采用：　$$\frac{\hat{\beta}_1^2}{\hat{\sigma}^2}L_{xx}\sim F(1,n-2)$$

即，$$\frac{(n-2)L_{xy}^2}{L_{xx}L_{yy}-L_{xy}^2}\sim F(1,n-2)$$

这个检验统计量恰好就是 $(n\text{-}2)r^2/(1\text{-}r^2)$，也就是 $(n\text{-}2)$ RegSS / RSS。

利用F 统计量 $F=\dfrac{(n-2)r^2}{(1-r^2)}$

H_0: $\beta_1=0$ 的否定域是 $\{F>F_{0.05}(1,n\text{-}2)\}$，如果零假设被否定，即认为回归方程成立。

Remark

当零假设没有被拒绝，意味着这两个数值变量之间不存在前面建立的线性回归关系，但是它们之间可能存在着其它类型的关系

因此讨论两个数值变量的回归关系时，比较恰当的做法是计算有关的F 检验的 p-值，它越小说明所建立的回归方程越好。

2. 回归系数的区间估计

从 $\hat{\beta}_1 - \dfrac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} t_{\alpha/2}(n-2)$

到 $\hat{\beta}_1 + \dfrac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} t_{\alpha/2}(n-2)$

原因是 $\dfrac{\hat{\beta}_1-\beta_1}{\hat{\sigma}}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \sim t(n-2)$

例5.2.3　哺乳动物出生以后开始玩耍的时间 p 与开始行走的时间 w 似乎有关：

类别	走动w(天)	玩耍 p(天)
人类	360	90
大猩猩	165	105
猫	21	21
家犬	23	26
挪威鼠	11	14
乌鸫	18	28
混血猕猴	18	21
黑猩猩	150	105
松鼠猴	45	68
花鼠	45	75
白脸猴	18	46

解. 假定 p 对于 w 有线性回归关系，

回归直线为：$\hat{p} = 35.81 + 0.235\,w$ ，

对应检验统计量 $F = 9.50635$ ，而p-值只有

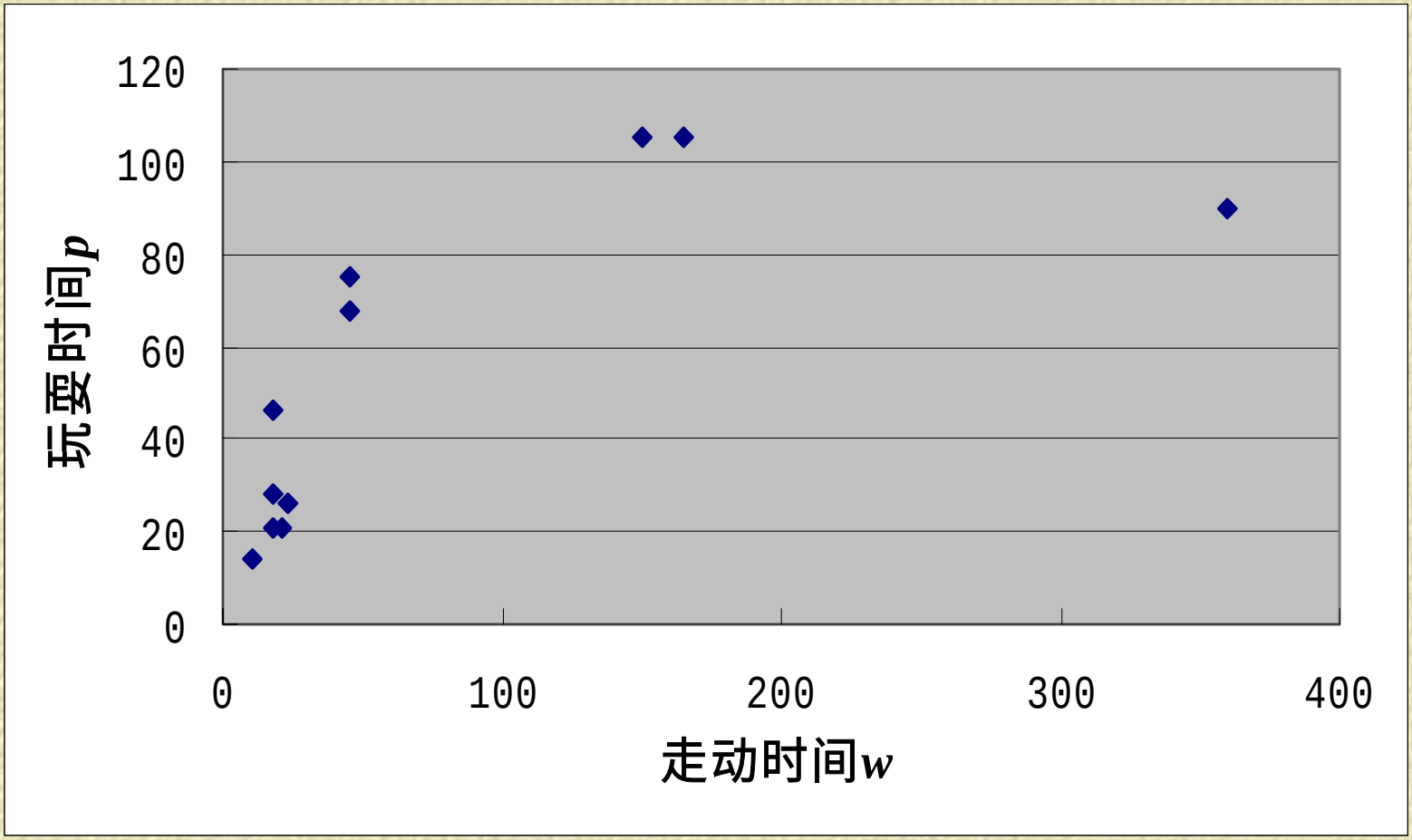
$$P(F(1,9) > 9.50635) = 0.01307,$$

即水平0.01下回归关系不显著，而在水平0.05下才有显著的回归关系。

比较恰当的解决方法：

画出散点图，数据似乎落在幂函数曲线：

$p = a\,w^b$ （$a > 0$， $b > 0$）的附近。


120
100
80
60
40
20
0
玩耍时间p
0
100
200
300
400
走动时间w

作变换：$y_i = \ln p_i$，　$x_i = \ln w_i$　；

对数据 (x_i, y_i) 作回归拟合得到回归方程：

$$\hat{y} = 1.689 + 0.561\, x$$ 。

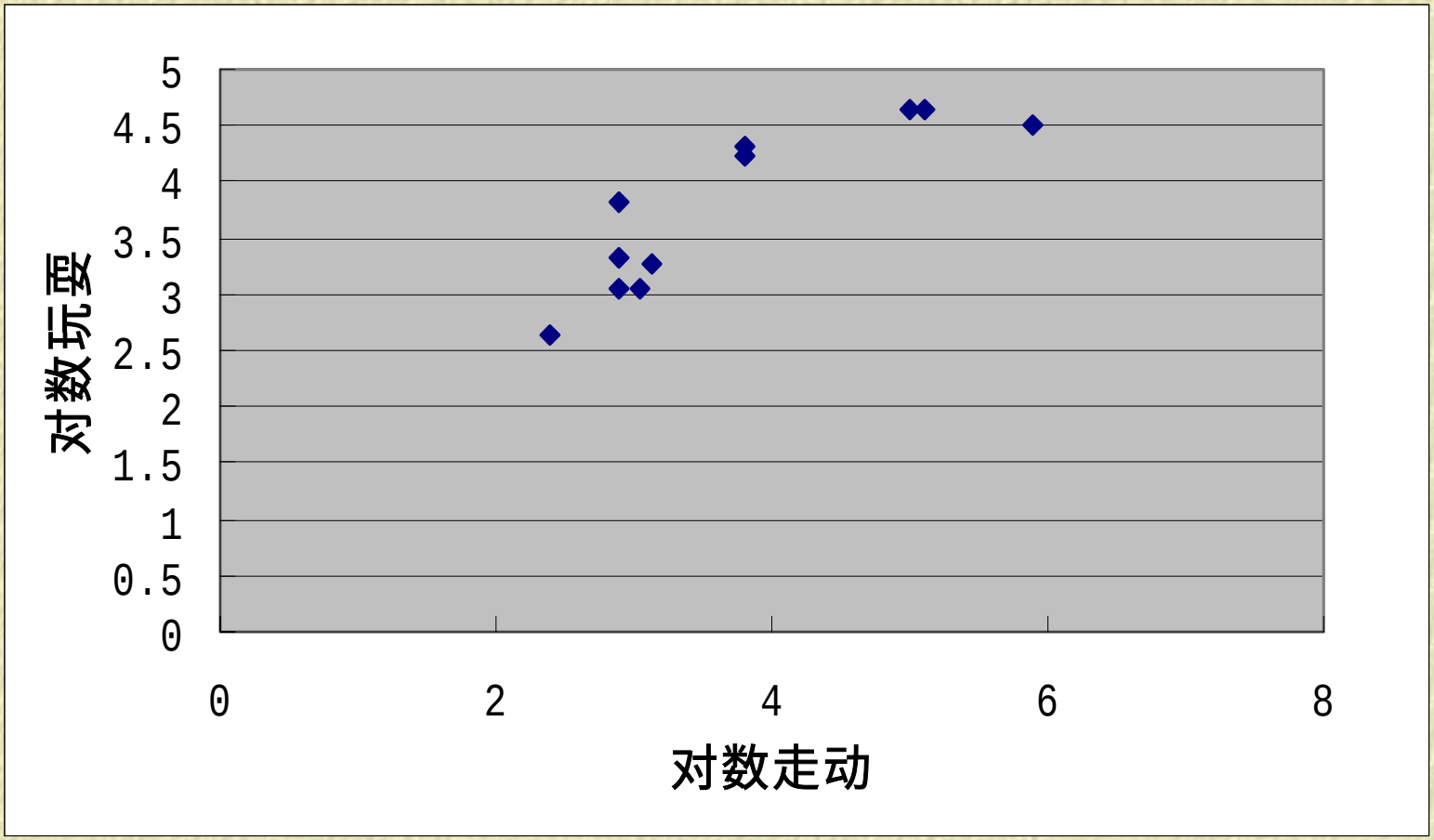
此时对应的检验统计量 $F = 27.438$，而

p-值　$P(F(1,9) > 27.438) = 0.00054$，

即在水平0.001下回归关系也是显著的。

哺乳动物出生以后开始玩耍的时间 p 关于开始走动的时间 w 的回归关系更可能应该是：

$$p = 5.42\, w^{0.561}$$ 。


对数玩耍
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
2
4
6
8
对数走动

5.2.4 回归方程的预测与控制

假定对回归模型 $y = \beta_0 + \beta_1 x + \varepsilon$，我们已经观察到了一组数据$(x_i, y_i)$，$1 \le i \le n$。

现在希望了解 $x = x_0$ 时对应的 $y = y_0$ 的情况。

很自然的，应该有关系：

$$y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$$

1. 回归方程的预测

如果只需要 y_0 的一个点估计，显然有：

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

如果需要预测 y_0 的一个范围，则应该求出一个区间估计，但必须知道与 y_0 有关的分布。

记 $y_0^* = \hat{\beta}_0 + \hat{\beta}_1 x_0$ (可以计算出来)

这里 y_0^* 只可能和随机变量 ε_1，…，ε_n 有关，

并且根据定理 5.1.3，($\hat{\beta}_0$，$\hat{\beta}_1$) 服从二维正态分布，因此 y_0^* 是一个只与 ε_1，…，ε_n 有关的服从一维正态分布的随机变量。

现在 $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$，只和随机变量 ε_0 有关。因此，

$y_0 - y_0^*$是两个独立正态随机变量的差，仍然服从正态分布。

有 $y_0 - y_0^* \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\right)$

把 $y_0 - y_0^*$ 的分布中心标准化为：

$$\frac{y_0 - y_0^*}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim N(0,1)$$

要消去未知参数σ，使用总体方差σ^2 的估计 $\hat{\sigma}^2$；
它与 $\hat{\beta}_0$、$\hat{\beta}_1$ 都独立，自然也与 $y_0 - y_0^*$ 独立。

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

$$\frac{y_0 - y_0^*}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim t(n-2)$$

$$y_0^* = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

从而得到 y_0 的一个置信度 1 - α 区间估计，或者说给出一个 x_0，则相应的因变量 y_0 以 1 - α 的概率在如下的一个范围内变化：

从 ($\hat{\beta}_0 + \hat{\beta}_1 x_0$ - h) 到 ($\hat{\beta}_0 + \hat{\beta}_1 x_0$ + h)

这里 $h = t_{\alpha/2}(n-2)\,\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

2. 回归方程的控制

这是预测问题的逆问题，即需要 y_0 以$1-\alpha$的概率落在一个范围(A,B)内，问 x_0 的变化范围应该是什么？

只需要取 x_0 使得：

$$A \le y_0^* - h \text{ 以及 } y_0^* + h \le B$$

同时成立，即可解出 x_0 相应的变化范围

这里 $y_0^* = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$$h = t_{\alpha/2}(n-2)\,\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

例5.2.4　给出例5.2.2中成年女性当身高 $x_0=170$ 时腿长 y_0 的预测区间。

解： 回归方程的 p-值为 2.13×10^{-9}，即

腿长 ＝ -16.073 ＋0.7194 身高

具有高度显著的统计意义。

身高 $x_0=170$ 时腿长 y_0 的预测区间为：

(102.746 , 109.6882)

□

例5.2.5　对某种型号钢的抗拉强度 Y 与硬度 X 观察了20 个数据，建立回归方程、检验并预测 $x = 230$ 时的 y 值。

x	277	257	255	278	306	268	285	286	272	285
y	103	99.5	93	105	110	98	103.5	103	104	103
x	286	269	246	255	253	255	269	297	257	250
y	108	100	96.5	92	94	94	99	109	95.5	91

□

回归分析需要注意的几点

(1) 实际问题中回归模型的建立要依赖于专业知识，并且注意散点图的使用；

(2) 即使回归模型通过了检验也只能认为所研究的变量是统计相关的；

(3) 回归分析一般需要与相关分析结合起来；

(4) 异方差性、序列相关性、多重共线性问题。

练习5.2.6　　讨论发展中国家受教育人口比例与人均收入的关系。

国家	1	2	3	4	5	6	7	8	9	10
教育	6	43	50	87	80	71	30	77	9	77
收入	61	165	125	645	398	208	289	311	246	86
国家	11	12	13	14	15	16	17	18	19	20
教育	10	6	6	22	80	6	46	11	30	6
收入	46	72	73	107	246	158	600	174	92	66

练习5.2.7　美国结婚人数与离婚人数随时间变化的趋势，数据从1890年(记为1)开始每隔5年直到1980年(记为19)。

年份	1	2	3	4	5	6	7	8	9	10
结婚	570	620	709	842	948	1008	1274	1188	1127	1327
离婚	33	40	56	68	83	104	170	175	196	218

年份	11	12	13	14	15	16	17	18	19
结婚	1596	1613	1667	1531	1523	1800	2159	2153	2413
离婚	264	485	385	377	393	479	708	1036	1182

(1) 画出结婚数为自变量，离婚数为因变量的散点图；

(2) 画出结婚数的对数为自变量，离婚数对数为因变量的散点图；

(3) 画出离婚比例随时间变化趋势的散点图；

(4) 从这些散点图你能够发现什么？根据这些数据做一个适当的回归与相关分析。

练习5.2.8

对例题5.2.3寻找一个可能更好的回归方程。

第5.3节 多元回归分析

对于一般的多元回归模型：

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \beta \in \mathbf{R}^{k+1}$$

5.3.1 未知参数的估计

常数项 β_0、各回归系数 β_i 以及误差方差 σ^2 的估计自然就采用定理 5.1.1 的估计。

5.3.2 回归模型的检验

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

如果拒绝了零假设，则说明 y 的期望与自变量 $x_1, \ldots, x_k$ 具有显著的线性关系，回归方程成立；否则说明我们建立的回归关系不显著。

同理如练习5.1.1定义总平方和TSS、回归平方和RegSS 以及 残差平方和RSS，

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

$$\mathrm{RegSS}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \qquad \mathrm{RSS}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

仍然可以证明 RegSS 与 RSS 独立，

根据定理5.1.3，

$$\frac{\mathrm{RSS}}{\sigma^2}\sim\chi^2(n-k-1)$$

并且零假设 H_0： $\beta_1 = \ldots = \beta_k = 0$ 成立时，有

$$\frac{\text{RegSS}}{\sigma^2} \sim \chi^2(k)$$

所以由F 统计量的构造，

$$F = \frac{n-k-1}{k} \frac{\text{RegSS}}{\text{RSS}} \sim F(k, n-k-1)$$

故零假设(即回归模型不成立) 的一个水平α 的拒绝域为 $\{ F \geq F_\alpha(k, n\text{-}k\text{-}1) \}$

检验的p-值是 $P\{F(k, n\text{-}k\text{-}1) > F$ 比$\}$

*EXCEL*函数*LINEST*检验多元回归模型

以 $y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \varepsilon$ 为例，

$\hat{\beta}_3$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_0$
$\sqrt{c_{33}}\hat{\sigma}$	$\sqrt{c_{22}}\hat{\sigma}$	$\sqrt{c_{11}}\hat{\sigma}$	$\sqrt{c_{00}}\hat{\sigma}$
r^2	$\hat{\sigma}$		
F 比	n-3-1		
RegSS	RSS		

5.3.3 回归因子的挑选

逐步回归的想法

讨论假设检验的问题：

$$H_{0i}：\beta_i = 0 \quad \Leftrightarrow \quad H_{1i}：\beta_i \neq 0$$

如果接受了这个零假设，就可以把因子 x_i 从模型中剔除。

(有时候也采用逐一添加回归因子的方法)

对于 H_{0i} ： $\beta_i = 0$ ， $i = 1, 2, \ldots, k$

如果采用 t 检验，则检验统计量为

$$T_i = \frac{\hat{\beta}_i}{\sqrt{c_{ii}}\hat{\sigma}} \quad ，自由度 \ n\text{-}k\text{-}1$$

检验的p-值是 $P\{ t(n\text{-}k\text{-}1) > T_i \}$

如果采用 F 检验，则检验统计量为

$$F_i = \frac{\hat{\beta}_i^2}{c_{ii}\hat{\sigma}^2} \quad ,$$

检验的p-值是 $P\{ F(1, \ n\text{-}k\text{-}1) > F_i \}$

例5.3.1　钢的去碳量 Y 与两种矿石 X_1、X_2 以及溶化时间 X_3 有关，实际测量了49组数据，作多元回归分析。

解. 见相关数据文件。

□