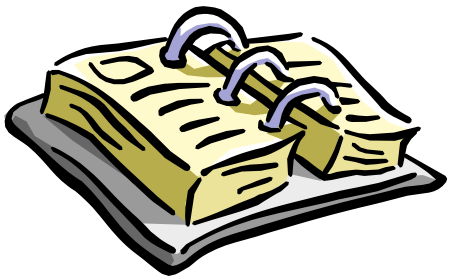
第9章 相关与一元回归分析

- 9.1 相关分析
- 9.2 一元线性回归
- 9.3 利用回归方程进行估计和预测



学习内容

1.相关系数的分析方法
2.线性回归的基本原理和参数的最小二乘估计
3.回归直线的拟合优度
4.回归方程的显著性检验
5.利用回归方程进行估计和预测

相关与回归

- 当对某种现象进行研究时，首先是观测或研究现象的变化及其分布特征；当某种现象的变化及其分布特征清楚后，须分析产生这种变化的原因，方差分析就是分析因素对因变量效应的一种方法。方差分析分析了因素对因变量的影响，得到的直接结果是不同因素不同水平处理组合下是否存在差异，如果存在则可以推断该因素对因变量有影响。进一步，我们还希望知道，这个因素怎样影响因变量？影响的程度如何？因变量怎样随因素变量变化而变化？
- 回答这些问题，需要研究两个或多个变量之间的关系。

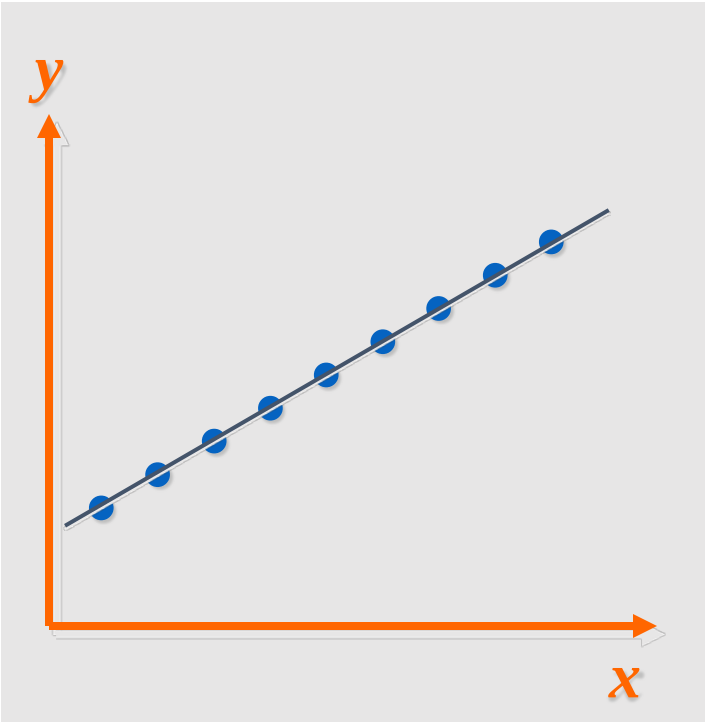
8.1 相关分析

一. 变量间的关系

二. 相关关系的描述与测度

一. 变量间的关系函数关系

1. 是一一对应的确定关系
2. 设有两个变量x和y，变量y随变量x一起变化，并完全依赖于x，当变量x取某个数值时，y依确定的关系取相应的值，则称y是x的函数，记为$y=f(x)$，其中x称为自变量，y称为因变量
3. 各观测点落在一条线上

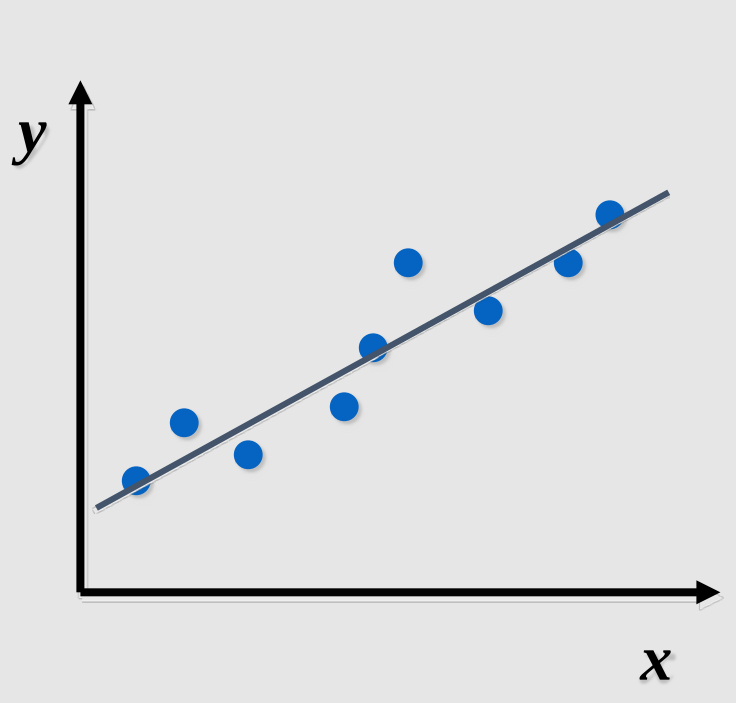


函数关系(几个例子)

- 函数关系的例子
  - 某种商品的销售额(y)与销售量(x)之间的关系可表示为 $y = px$ (p 为单价)
  - 圆的面积(S)与半径之间的关系可表示为$S=\pi R^2$
  - 企业的原材料消耗额(y)与产量(x_1)、单位产量消耗(x_2)、原材料价格(x_3)之间的关系可表示为$y = x_1\ x_2\ x_3$

相关关系(correlation)

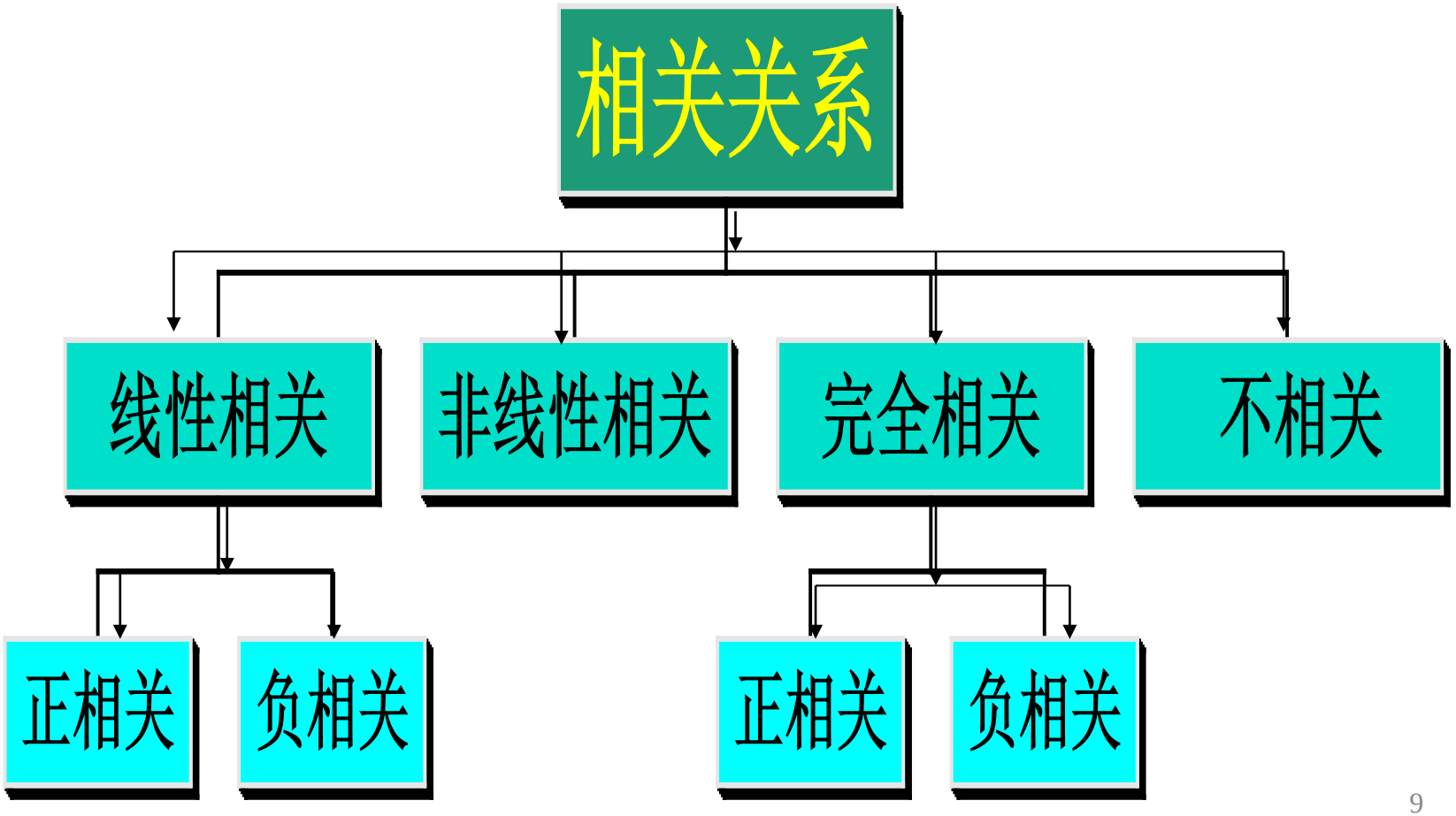
1. 变量间关系不能用函数关系精确表达
2. 一个变量的取值不能由另一个变量唯一确定
3. 当变量 x 取某个值时，变量 y 的取值可能有几个
4. 各观测点分布在直线周围



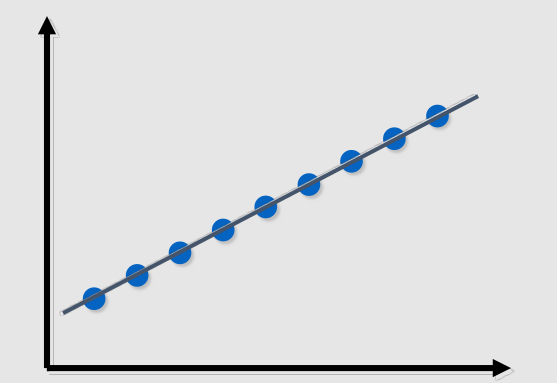
相关关系(几个例子)

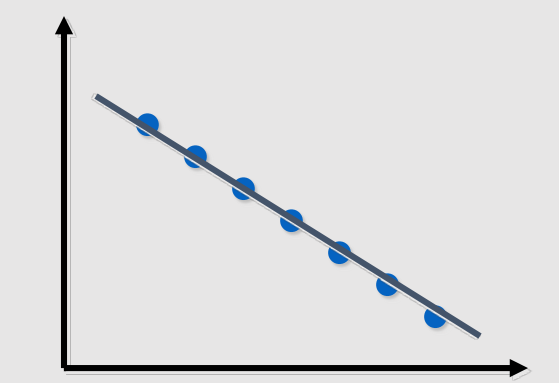
- 相关关系的例子
  - 父亲身高(y)与子女身高(x)之间的关系
  - 收入水平(y)与受教育程度(x)之间的关系
  - 粮食亩产量(y)与施肥量(x_1)、降雨量(x_2)、温度(x_3)之间的关系
  - 商品的消费量(y)与居民收入(x)之间的关系
  - 商品销售额(y)与广告费支出(x)之间的关系

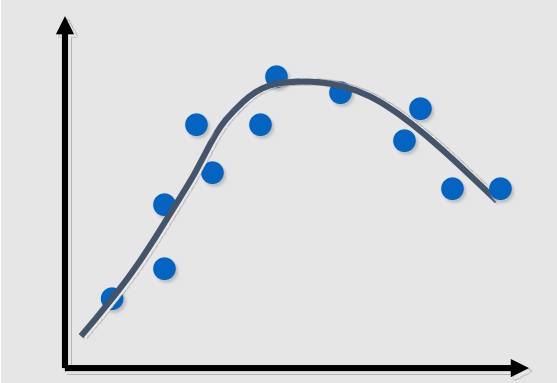
相关关系(类型)

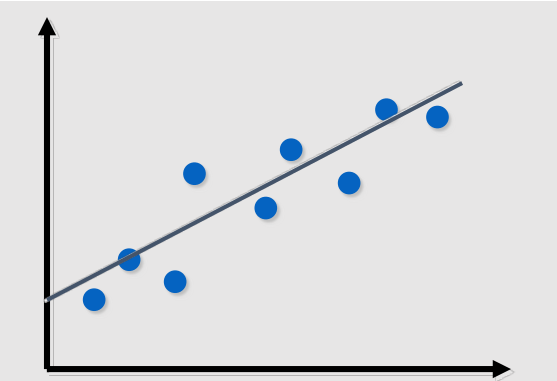


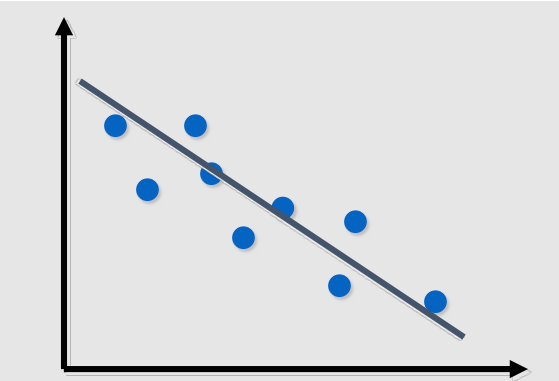
散点图(scatter diagram)

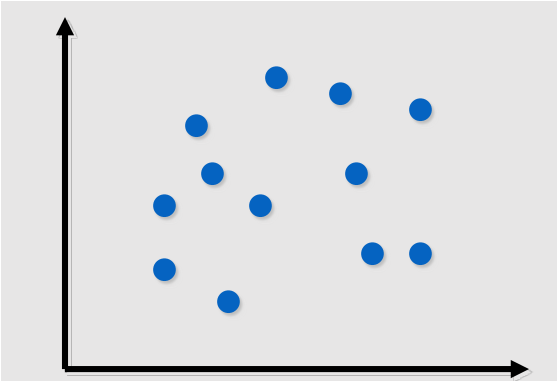

完全正线性相关


完全负线性相关


非线性相关


正线性相关


负线性相关


不相关

一、直线相关分析

进行直线相关分析的基本任务在于根据x、y的实际观测值计算表示两个相关变量x与y线性相关程度和性质的统计数——相关系数r，并进行显著性检验。

相关系数(correlation coefficient)

1. 对变量之间关系密切程度的度量
2. 对两个变量之间线性相关程度的度量称为简单相关系数
3. 若相关系数是根据总体全部数据计算的，称为总体相关系数，记为ρ
4. 若是根据样本数据计算的，则称为样本相关系数，记为r

相关系数（计算公式）

- 样本相关系数的计算公式

$$r=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-\left(\sum x\right)^2}\cdot\sqrt{n\sum y^2-\left(\sum y\right)^2}}$$

相关系数——协方差

协方差(covariance)：两个变量与其均值离差乘积的平均数，是相互关系的一种度量。

总体协方差：

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

样本协方差：

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

相关系数——协方差

协方差

协方差为大的正值时，表示强的正线性相关关系。

协方差接近于零时，表示很小，没有线性相关关系。

协方差为大的负值时，表示强的负线性相关关系。

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

相关系数——协方差

cm kg

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

大于

mm kg

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

基本结论：协方差受计量单位影响，从而不能真实反映相关的程度。

相关系数——协方差

相关系数（**correlation coefficient**）:协方差与两变量标准差乘积的比值，是没有量纲的、标准化的协方差。

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

总体相关系数

$$r = \frac{S_{xy}}{S_x S_y}$$

样本相关系数

相关系数（计算公式）

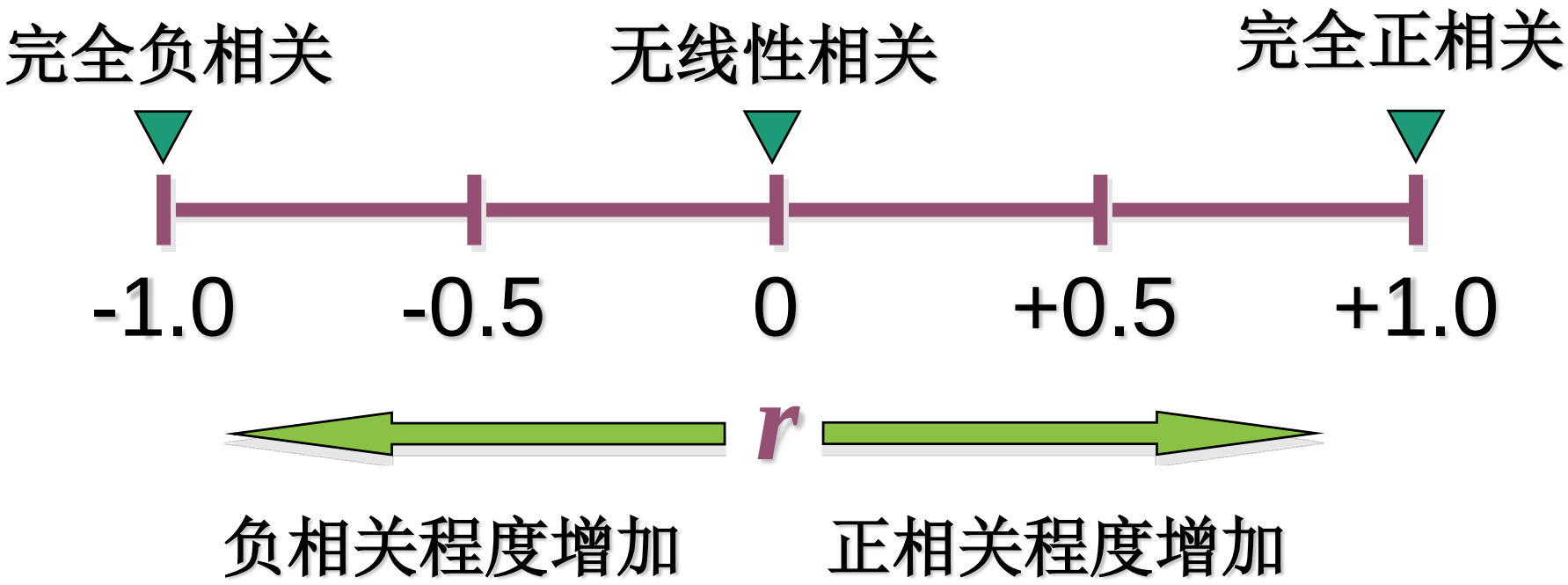
- 样本相关系数的计算公式

$$r=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-\left(\sum x\right)^2}\cdot\sqrt{n\sum y^2-\left(\sum y\right)^2}}$$

相关系数(取值及其意义)

1. r 的取值范围是 [-1, 1]
2. $|r|$=1，为完全相关
   - r =1，为完全正相关
   - r = -1，为完全负相关
3. r = 0，不存在*线性*相关关系相关
4. -1≤r<0，为负相关
5. 0<r≤1，为正相关
6. $|r|$越趋于1表示关系越密切；$|r|$越趋于0表示关系越不密切

相关系数(取值及其意义)



相关系数的性质

- **性质1**：*r*具有对称性。即*x*与*y*之间的相关系数和*y*与*x*之间的相关系数相等，即$r_{xy}= r_{yx}$
- **性质2**：*r*数值大小与*x*和*y*原点及尺度无关，即改变*x*和*y*的数据原点及计量尺度，并不改变*r*数值大小
- **性质3**：仅仅是*x*与*y*之间线性关系的一个度量，它不能用于描述非线性关系。这意味着， *r*=0只表示两个变量之间不存在线性相关关系，并不说明变量之间没有任何关系
- **性质4**：*r*虽然是两个变量之间线性关系的一个度量，却不一定意味着*x*与*y*一定有因果关系

二、相关系数的显著性检验

上述根据实际观测值计算得来的r是样本相关系数，它是双变量正态总体的总体相关系数ρ的估计值。样本相关系数r是否来自$\rho \neq 0$的总体，还须对样本相关系数r进行显著性检验。此时无效假设、备择假设为：H_0 $\rho = 0$，H_A $\neq 0$。

（1）假设　$H_0: \rho = 0$ ； $H_A: \rho \neq 0$

（2）水平　选取显著水平 α

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

（4）推断

$|t| > t_{\alpha(n-2)}$　在 α 显著水平上，否定H_0，接受H_A；推断r显著。

$|t| < t_{\alpha(n-2)}$　在 α 显著水平上，接受H_0，否定H_A；推断r不显著。

r经显著性检验的结果呈不显著时，
便推断两变数间不存在相关关系，
这时不能用r代表其相关密切程度。

例：

温度　　天数

X	Y
平均温度（℃）	历期天数（d ）
11.8	30.1
14.7	17.3
15.6	16.7
16.8	13.6
17.1	11.9
18.8	10.7
19.5	8.3
20.4	6.7

温度　　天数

$$r = \frac{SP_{xy}}{\sqrt{SS_x \times SS_y}} = \frac{-139.6937}{\sqrt{55.1788 \times 377.2688}} = -0.9682$$

$$r^2 = 0.9374$$

黏虫孵化历期平均温度与历期天数成负相关。

x和y的变异有93.74%可用二者之间的线性关系来解释。

（1）假设　$H_0: \rho = 0$ ； $H_A: \rho \neq 0$

（2）水平　选取显著水平 $\alpha = 0.01$

（3）检验　$$t = \frac{r}{s_r} = \frac{r}{\sqrt{\frac{(1-r^2)}{n-2}}} \frac{-0.9682}{0.1021} = -9.48$$

（4）推断　$$|t| > t_{0.01(6)} = 3.707$$

否定H_0，接受H_A；推断r极显著，黏虫孵化历期温度与历期天数之间存在着极显著的直线相关关系。

椰子树的产量数X(个)

X(个)	120	121	123	126	128
Y(尺)	21	23	22	25	24

椰子树的高度Y(尺)

$r = 0.7996 < r_{0.05(3)} = 0.8783$

椰子树的产果树与树高之间无直线相关关系。

当样本太小时，即使r值达到0. 7996，样本也可能来自总体相关系数 ρ =0的总体。

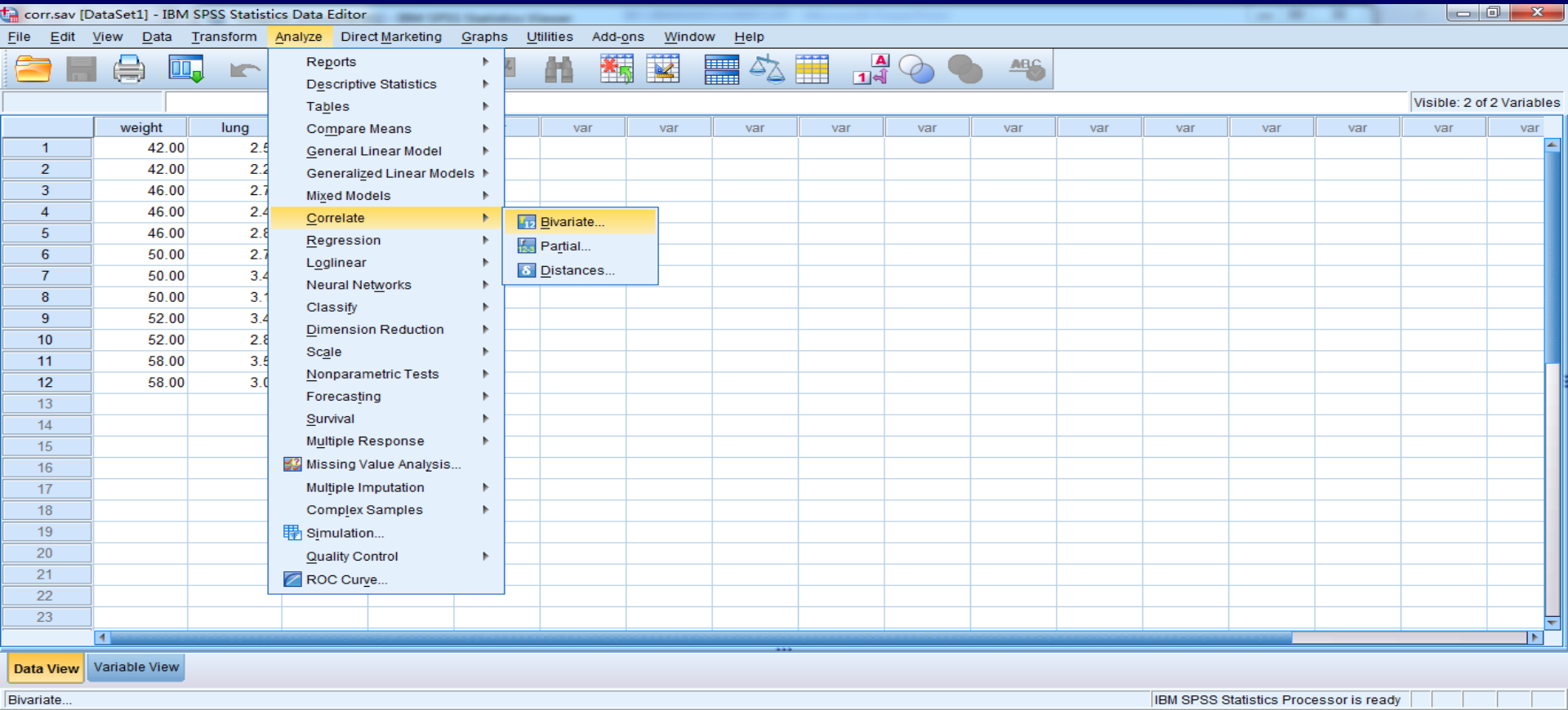
1 不能直观地由r值判断两变数间的相关密切程度。

2 试验或抽样时，所取的样本容量n大一些，由此计算出来的r值才能参考价值。

相关分析：SPSS中的Correlate过程

例：某地一年级12名女大学生的体重是以肺活量数据如下，试分析两者的直线相关关系

体重	42	42	46	46	46	50	50	50	52	52	58	58
肺活量	2.55	2.20	2.75	2.40	2.80	2.81	3.41	3.10	3.46	2.85	3.50	3.00



9.2 一元线性回归

一. 一元线性回归模型
二. 参数的最小二乘估计
三. 回归直线的拟合优度
四. 显著性检验

什么是回归分析？(Regression)

1. 从一组样本数据出发，确定变量之间的数学关系式
2. 对这些关系式的可信程度进行各种统计检验，并从影响某一特定变量的诸多变量中找出哪些变量的影响显著，哪些不显著
3. 利用所求的关系式，根据一个或几个变量的取值来预测或控制另一个特定变量的取值，并给出这种预测或控制的精确程度

回归分析的任务是揭示出呈因果关系的相关变量间的联系形式，建立它们之间的回归方程，利用所建立的回归方程，由自变量（原因）来预测、控制因变量（结果）。

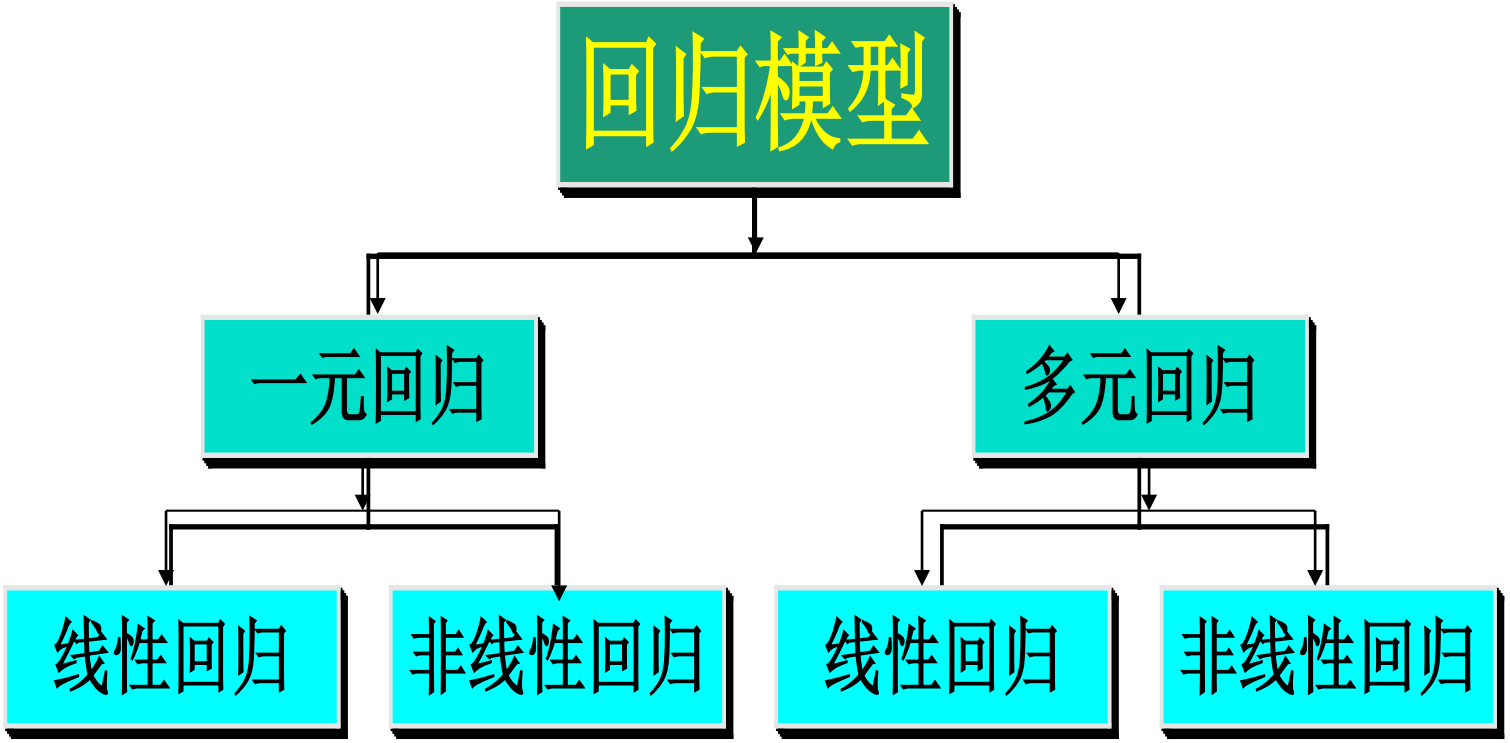
回归（Regression）

- 早在1885年，高尔顿(F.Galton)在“Regression towards mediocrity in hereditary stature”一文中发表了他关于根据父母身体特征预测子女身体特征的研究结果。其发现“身高偏高的父母，其子女平均身高要低于他们的父母的平均身高；相反的，身高矮的父母，其子女的平均身高却要高于他们的父母的平均身高。他在此预测工作的论文中利用了“Regression”一词表示此效应。也就是两极端分数会“回归”到平均数的现象。因此，将一变量去预测另一变量的方法称为回归分析。回归一词本有其特殊意义，现在已经将其一般化，用以描述两个或两个以上变量间的关系。所以回归分析是以一个或多个变量(independent variable)描述、预测或控制一特定因变量(dependent variable)的分析。

回归分析与相关分析的区别

1. 相关分析中，变量x变量y处于平等的地位；回归分析中，变量y 称为因变量，处在被解释的地位，x称为自变量，用于预测因变量的变化
2. 相关分析中所涉及的变量x和y都是随机变量；回归分析中，因变量y是随机变量，自变量x可以是随机变量，也可以是非随机的确定变量
3. 相关分析主要是描述两个变量之间线性关系的密切程度；回归分析不仅可以揭示变量x对变量y的影响大小，还可以由回归方程进行预测和控制

回归模型的类型



一元线性回归

1. 涉及一个自变量的回归
2. 因变量*y*与自变量*x*之间为线性关系
   - 被预测或被解释的变量称为因变量(dependent variable)，用*y*表示
   - 用来预测或用来解释因变量的一个或多个变量称为自变量(independent variable)，用*x*表示
3. 因变量与自变量之间的关系用一条线性方程来表示

回归模型(regression model)

1. 回答“变量之间是什么样的关系？”
2. 方程中运用
   - 1 个数字的因变量(响应变量)
     - 被预测的变量
   - 1 个或多个数字的或分类的自变量 (解释变量)
     - 用于预测的变量
3. 主要用于预测和估计

一元线性回归模型

1. 描述因变量 y 如何依赖于自变量 x 和误差项ε的方程称为*回归模型*
2. 一元线性回归模型可表示为

$$y = b_0 + b_1 x + \varepsilon$$

- y 是 x 的线性函数(部分)加上误差项
- 线性部分反映了由于 x 的变化而引起的 y 的变化
- 误差项 ε 是随机变量
  - 反映了除 x 和 y 之间的线性关系之外的随机因素对 y 的影响
  - 是不能由 x 和 y 之间的线性关系所解释的变异性
- β_0 和 β_1 称为模型的参数

一元线性回归模型(基本假定)

1. 误差项 ε 是一个期望值为0的随机变量，即 $E(\varepsilon)=0$。对于一个给定的x值，y的期望值为

   $E(y)=\beta_0+\beta_1 x$

1. 对于所有的x值，ε 的方差 σ^2 都相同
2. 误差项 ε 是一个服从正态分布的随机变量，且相互独立。即 $\varepsilon-N(0,\sigma^2)$
   - 独立性意味着对于一个特定的x值，它所对应的 ε 与其他x值所对应的 ε 不相关
   - 对于一个特定的x值，它所对应的y值与其他x所对应的y值也不相关

回归方程 (regression equation)

1. 描述 y 的平均值或期望值如何依赖于 x 的方程称为回归方程
2. 一元线性回归方程的形式如下

$$E(y) = \beta_0 + \beta_1 x$$

- 方程的图示是一条直线，也称为直线回归方程
- β_0是回归直线在 y 轴上的截距，是当 x=0 时 y 的期望值
- β_1是直线的斜率，称为回归系数，表示当 x 每变动一个单位时，y 的平均变动值

估计的回归方程(estimated regression equation)

一元线性回归中估计的回归方程为:

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

$\hat{\beta}_1$

其中：$\hat{\beta}_0$ 是估计的回归直线在 y 轴上的截距， 是直线的斜率，它表示对于一个给定的 x 的值，$\hat{y}$ 是 y 的估计值，也表示 x 每变动一个单位时， y 的平均变动值

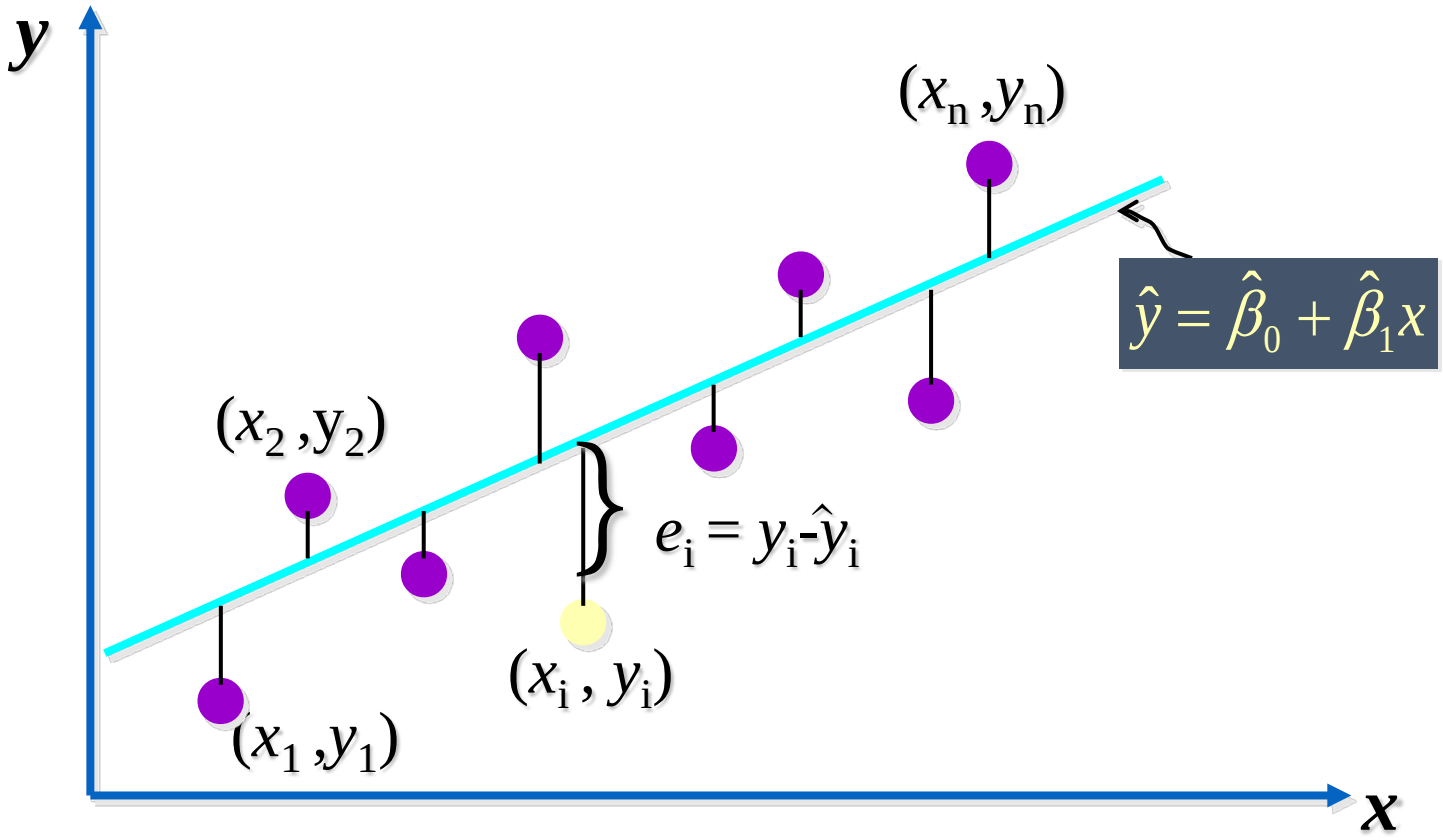
最小二乘估计

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 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方法。即

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

2. 用最小二乘法拟合的直线来代表x与y之间的关系与实际数据的误差比其他任何直线都小

最小二乘估计(图示)



最小二乘法
($\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的计算公式)

- 根据最小二乘法，可得求解 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的公式如下

$$\begin{cases} \left.\dfrac{\partial Q}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)=0 \\ \left.\dfrac{\partial Q}{\partial \beta_1}\right|_{\beta_1=\hat{\beta}_1} = -2\sum_{i=1}^{n}x_i(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)=0 \end{cases}$$

最小二乘法（$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的计算公式）

- 根据最小二乘法的要求，可得求解 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的公式如下

$$\begin{cases} \hat{\beta}_1 = \dfrac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

【例9·1】 江苏武进县测定1956-1964年间3月下旬至4月中旬平均温度累积值 （x，单位：旬·度）和一代三化螟蛾盛发期 （y，以5月10日为0）的资料如下表，建立y与x的直线回归方程。

表7-1 平均温度累积值(x)与一代三化螟盛发期（y）资料

年份	1956	1957	1958	1959	1960	1961	1962	1963	1964
累积温 x	35.5	34.1	31.7	40.3	36.8	40.2	31.7	39.2	44.2
盛发期 y	12	16	9	2	7	3	13	9	-1

估计方程的求法(例题分析)

【例】求三化螟蛾盛发期与3月下旬至4月中旬平均温度累积值的回归方程

$$\begin{cases} \hat{\beta}_1 = -1.0996 \\ \hat{\beta}_0 = 7.7778 - (-1.0996 \times 37.0778) = 48.5485 \end{cases}$$

回归方程为：$\boldsymbol{y = 48.5485 - 1.0996x}$

回归系数回归系数$\hat{\beta}_1 = -1.0996$的意义为：当3月下旬x的积温（x）每提高1旬•度时，一代三化螟盛发期将平均提早1.0996天；

回归截距 $\hat{\beta}_0 = 48.5485$的意义为：若3月下旬的积温为0，则一代三化螟盛发期为48.5485，即在6月27-28日。

注意，由于实测区间为[31.7，44.2]，当x＜31.7或x＞44.2时，y 的变化是否还符合$\hat{y} = 48.5485 - 1.0996x$的规律，还必须提供新的依据。

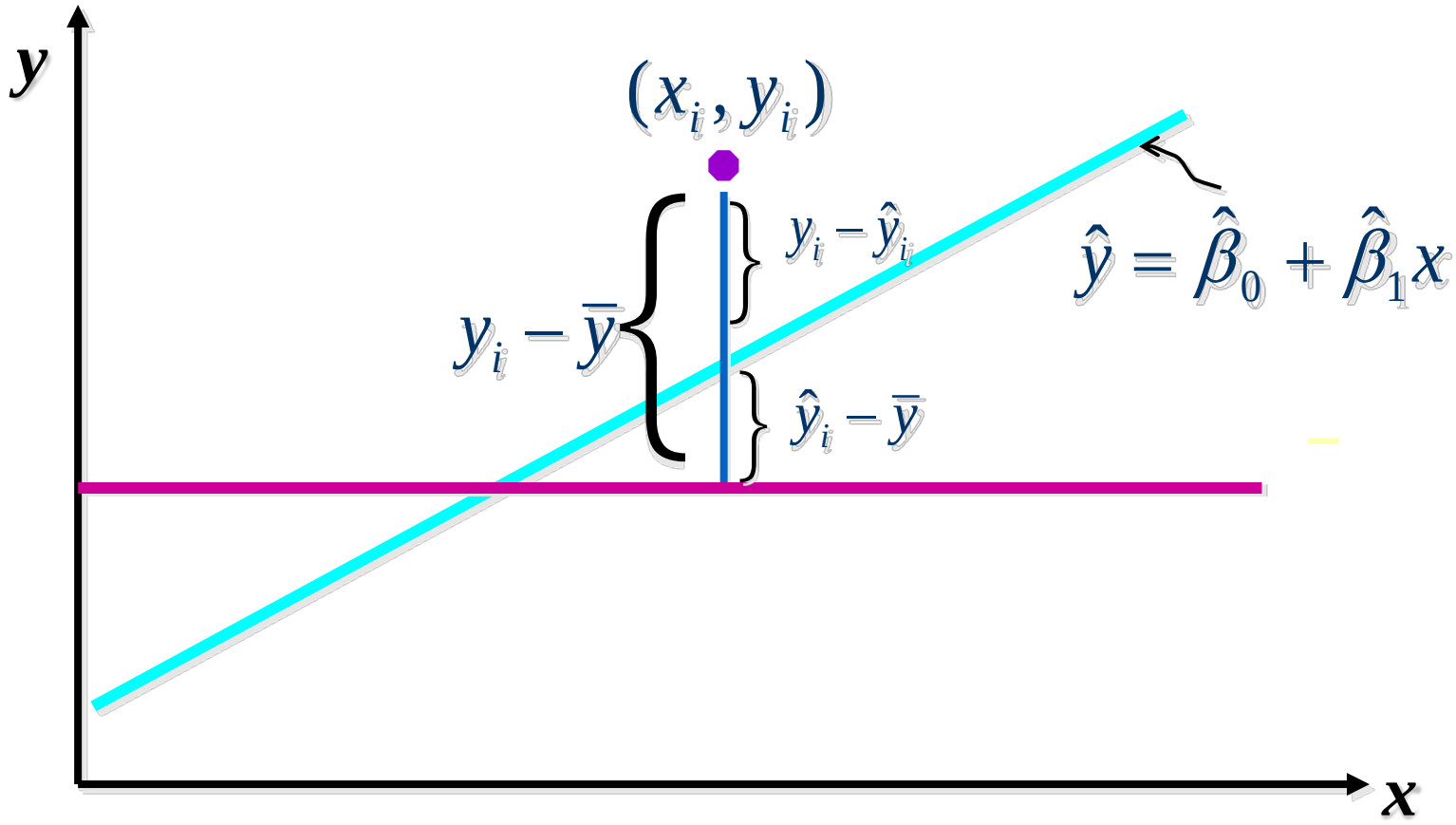
基本假定

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

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- x是没有误差的固定变量，或其误差可以忽略，而y是随机变量，且有随机误差。
- x是的任一值对应着一个y总体，且作正态分布，其平均数$\boldsymbol{\mu = \beta_0 + \beta_1 x}$，方差受偶然因素的影响，不因x的变化而改变。
- 随机误差 ε 是相互独立的，呈正态分布。

直线回归的变异来源



$y-\hat{y}$　实际值与估计值之差，剩余或残差。

$\hat{y}-\bar{y}$　估计值与均值之差，它与回归系数的大小有关。

离差平方和的分解（三个平方和的关系）

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

总平方和 (*SST*)　　回归平方和 (*SSR*)　　残差平方和 (*SSE*)

$$SST = SSR + SSE$$

离差平方和的分解（三个平方和的意义）

1. 总平方和(*SST*)
   - 反映因变量的 n 个观察值与其均值的总离差
2. 回归平方和(*SSR*)
   - 反映自变量 x 的变化对因变量 y 取值变化的影响，或者说，是由于 x 与 y 之间的线性关系引起的 y 的取值变化，也称为可解释的平方和
3. 残差平方和(*SSE*)
   - 反映除 x 以外的其他因素对 y 取值的影响，也称为不可解释的平方和或剩余平方和

判定系数r^2(coefficient of determination)

1. 回归平方和占总离差平方和的比例

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

2. 反映回归直线的拟合程度
3. 取值范围在 [0,1]之间
4. $R^2 \to 1$，说明回归方程拟合的越好；$R^2 \to 0$，说明回归方程拟合的越差
5. 判定系数等于相关系数的平方，即$R^2 = (r)^2$

估计标准误差(standard error of estimate)

1. 实际观察值与回归估计值离差平方和的均方根
2. 反映实际观察值在回归直线周围的分散状况
3. 对误差项ε的标准差σ的估计，是在排除了*x*对*y*的线性影响后，*y*随机波动大小的一个估计量
4. 反映用估计的回归方程预测*y*时预测误差的大小
5. 计算公式为

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

线性关系检验

1. 检验自变量与因变量之间的线性关系是否显著
2. 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
   - 回归均方：回归平方和*SSR*除以相应的自由度(自变量的个数*p*)
   - 残差均方：残差平方和*SSE*除以相应的自由度($n-p-1$)

线性关系检验（检验的步骤）

1. 提出假设
   - H_0：$\beta_1=0$　　线性关系不显著

2. 计算检验统计量F

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

3. 确定显著性水平α，并根据分子自由度1和分母自由度n-2找出临界值F_α
4. 作出决策：若$F>F_\alpha$，拒绝H_0；若$F<F_\alpha$，不能拒绝H_0

回归系数检验

1. 检验 x 与 y 之间是否具有线性关系，或者说，检验自变量 x 对因变量 y 的影响是否显著
2. 理论基础是回归系数 $\hat{\beta}_1$ 的抽样分布
3. 在一元线性回归中，等价于线性关系的显著性检验

回归系数检验(检验步骤)

1. 提出假设
   - H_0：$b_1 = 0$（没有线性关系）
   - H_1：$b_1 \neq 0$（有线性关系）
2. 计算检验的统计量

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} \sim t(n-2)$$

3. 确定显著性水平α，并进行决策
   - $|t| > t_{\alpha/2}$，拒绝H_0；$|t| < t_{\alpha/2}$，不能拒绝H_0

9.3 利用回归方程进行估计和预测

一. 点估计

二. 区间估计

利用回归方程进行估计和预测

1. 根据自变量 x 的取值估计或预测因变量 y的取值
2. 估计或预测的类型
   - 点估计
     - y 的平均值的点估计
     - y 的个别值的点估计
   - 区间估计
     - y 的平均值的*置信区间*估计
     - y 的个别值的*预测区间*估计

点估计

1. 对于自变量 x 的一个给定值x_0 ，根据回归方程得到因变量 y 的一个估计值 $\hat{y}_0$

2. 点估计值有
   - y 的*平均值*的点估计
   - y 的*个别值*的点估计

3. 在点估计条件下，平均值的点估计和个别值的的点估计是一样的，但在区间估计中则不同

区间估计

1. 点估计不能给出估计的精度，点估计值与实际值之间是有误差的，因此需要进行区间估计
2. 对于自变量 x 的一个给定值 x_0，根据回归方程得到因变量 y 的一个估计区间
3. 区间估计有两种类型
   - 置信区间估计(confidence interval estimate)
   - 预测区间估计(prediction interval estimate)

置信区间估计

1. 利用估计的回归方程，对于自变量 x 的一个给定值 x_0 ，求出因变量 y 的平均值的估计区间 ，这一估计区间称为*置信区间*
2. $E(y_0)$ 在1-α置信水平下的置信区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)s_y\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

式中：s_y为估计标准误差

预测区间估计

1. 利用估计的回归方程，对于自变量 x 的一个给定值 x_0 ，求出因变量 y 的一个个别值的估计区间，这一区间称为*预测区间*

2. y_0在1-α置信水平下的预测区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)S_y \sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

注意！#1

表9-2　回归截距α，回归系数β，总体平均数（$\alpha+\beta x$）和单个观测值　y　置信度为（$1-\alpha$）的置信区间

	标准误	置信度为（$1-\alpha$）置信区间
回归截距 α	$S_a = S_{yx}\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{SS_x}}$	$[a - t_{\alpha(n-2)}S_a, a + t_{\alpha(n-2)}S_a]$
回归系数 β	$S_b = S_{yx} / \sqrt{SS_x}$	$[b - t_{\alpha(n-2)}S_b, b + t_{\alpha(n-2)}S_b]$
y 总体平均数 $\alpha + \beta x$	$S_{\hat{y}} = S_{yx}\sqrt{\frac{1}{n}+\frac{(x-\overline{x})^2}{SS_x}}$	$[\hat{y} - t_{\alpha(n-2)}S_{\hat{y}}, \hat{y} + t_{\alpha(n-2)}S_{\hat{y}}]$
单个观测值 y	$S_y = S_{yx}\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{SS_x}}$	$[\hat{y} - t_{\alpha(n-2)}S_y, \hat{y} + t_{\alpha(n-2)}S_y]$

【例7·2】 根据【例7·1】的资料估计：

（1）当3月下旬至4月中旬的积温为40旬·度时，历年的一代三化螟蛾平均盛发期在何时（置信度为95%）？

（2）某年3月下旬至4月中旬的积温为40旬·度时，该年的一代三化螟蛾盛发期在何时（置信度为95%）？

利用直线回归方程 $\hat{y}=48.5485-1.0996x$ 计算当 $x=40$ 时的 $\hat{y}$，

$$\hat{y}=48.5485-1.0996\times 40=4.56$$

因为

$$S_{\hat{y}}=3.266\times\sqrt{\frac{1}{9}+\frac{(40-37.0778)^2}{144.6356}}=1.35$$

$$S_{y}=3.266\times\sqrt{1+\frac{1}{9}+\frac{(40-37.0778)^2}{144.6356}}=3.53$$

所以 （1）在置信度为95%时，$x=40$ 的 y 总体平均数（$\alpha+\beta x$）的置信区间为：

$$\hat{y}-t_{0.05(7)}S_{\hat{y}} \le (\alpha+\beta x) \le \hat{y}+t_{0.05(7)}S_{\hat{y}}$$

将 $\hat{y}$=4. 56、$S_{\hat{y}}=1.35$、 $t_{0.05(7)}=$ 2. 36代入，得

$$4.56-(2.36\times 1.35) \le (\alpha+\beta x) \le 4.56+(2.36\times 1.35)$$

$$1.4 \le (\alpha+\beta x) \le 7.7$$

即当3月下旬至4月中旬的积温为40旬•度时，历年的一代三化螟蛾平均盛发期在[1.4，7.7]或5月12—18日，置信度为95%。

（2）在置信度为95%时，$x = 40$ 的单个观测值 y 的置信区间为：

$$\hat{y} - t_{0.05(7)} S_y \leq y \leq \hat{y} + t_{0.05(7)} S_y$$

将 $\hat{y}$ =4. 56、$s_y = 3.53$、$t_{0.05(7)} = 2.36$代入，得

$$4.56-(2.36\times3.53)\le y\le 4.56+(2.36\times3.53)$$

$$-3.8\le y\le 19.9$$

即当某年3月下旬至4月中旬的积温为40旬•度时，该年的一代三化螟蛾盛发期在[-3. 8，19. 9]或5月6—30日，置信度为95%。

类似地可求出 x 取其它值时 y 总体平均数（$\alpha+\beta x$）和单个观测值　的95%置信区间，列于表7-3。

表7-3 一代三化螟蛾盛发期95%置信区间

x_i	$\hat{y}_i$	$\alpha+\beta x$ 的 95%置信区间		y 的 95%置信区间	
		置信下限	置信上限	置信下限	置信上限
30	15.6	10.3	20.8	6.2	24.9
32	13.4	9.2	17.5	4.6	22.1
34	11.2	7.9	14.4	2.8	19.5
36	9.0	6.3	11.6	0.8	17.1
38	6.8	4.1	9.4	-1.4	14.9
40	4.6	1.4	7.8	-3.8	12.9
42	2.4	-1.7	6.4	-6.4	11.1
44	0.2	-5.0	5.3	-9.1	9.4
46	-2.0	-8.3	4.2	-12.0	7.9

从$S_{\hat{y}}$和S_y的计算公式看出，x越接近$\bar{x}$，$S_{\hat{y}}$和S_y越小，置信区间的置信距也越小，预测越精确。

残差分析：检验模型假设

回归分析中的每一个观测值都有残差，其值为因变量的观察值y_i与由回归方程式预测而得的值$\hat{y}_i$值的的差，第i个观察值残差$y_i - \hat{y}_i$是以估计回归方程式预测值y_i所产生的误差的估计值。残差分析可以用来检验回归分析的前提假设是否成立

残差分析：检验模型假设

残差$y_i - \hat{y}_i$是ε的估计值；回归分析中如有n个观察值，就有n个残差。残差图可帮助我们判断有关的前提假设是否满足。一种最常见的残差图为

1. 残差对自变量X的图：x放横轴，残差放纵轴。对每个观察值，画出$(x_i, y_i - \hat{y}_i)$. 如果对所有x而言，ε的方差均相等这个假设成立的话，残差图将呈水平带状.
2. 残差对因变量的预测值$\hat{y}$的图: 预测值$\hat{y}_i$放横轴，残差$y_i - \hat{y}_i$放纵轴。
3. 将残差标准化（即减去平均值，然后除以标准差），再画出标准残差图。

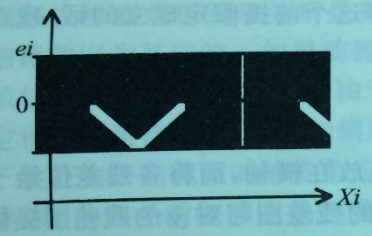
残差分析：检验模型假设

残差图的功能诊断

$E(\varepsilon_i) = 0$及

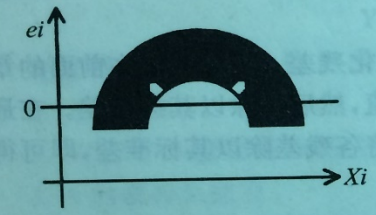
$Var(\varepsilon_i) = \sigma^2$

(i)



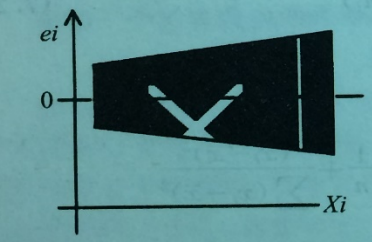
＊这样的残差图就好象符合 $E(\varepsilon_i)=0$ 及 $Var(\varepsilon_i)=\sigma^2$ 之假设，又叫做误差的同方差性(homoscedasticity)。

(ii)



＊利用线性回归不适合，应利用曲线回归(curvilinear)可能比较合适。

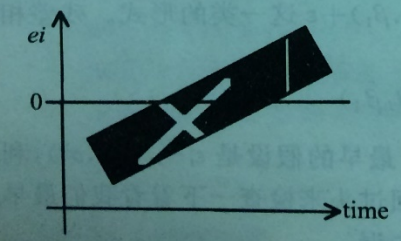
(iii)



＊表示 Var(ε_i)会随 X_i 的增加而增加，违反误差同方差性。应利用加权最小二乘方法重新估计参数。

＊heteroscedasticity(异方差性)

(iv)

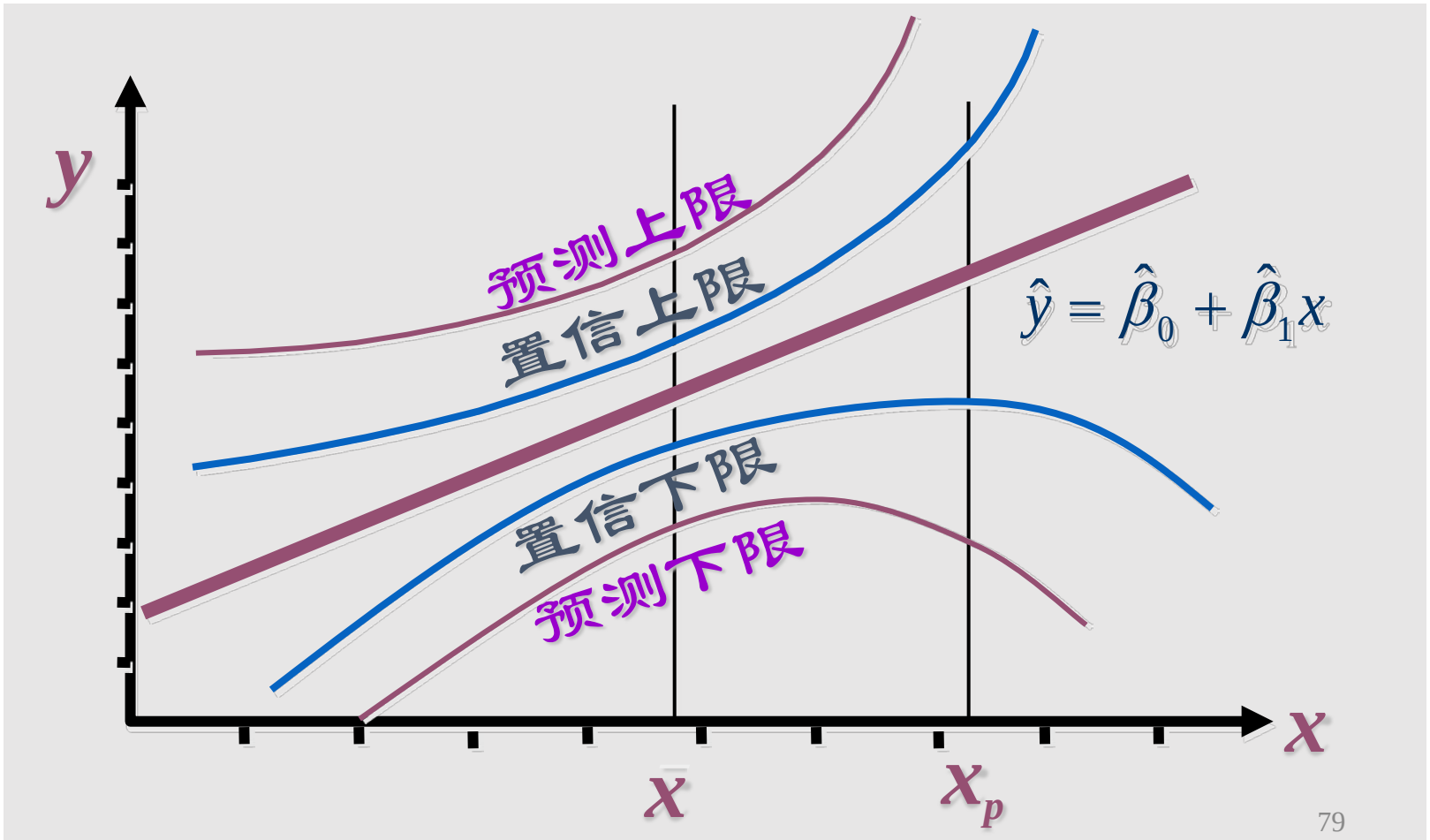


＊表示 e_i 会随时间增加而加大，故应考虑时间因素。

影响区间宽度的因素

1. 置信水平($1-\alpha$)
   - 区间宽度随置信水平的增大而增大
2. 数据的离散程度(s)
   - 区间宽度随离散程度的增大而增大
3. 样本容量
   - 区间宽度随样本容量的增大而减小
4. 用于预测的x_p与x的差异程度
   - 区间宽度随x_p与x 的差异程度的增大而增大

置信区间、预测区间、回归方程



决定系数和相关系数

已经证明了等式：

$$\sum(y-\bar{y})^2=\sum(\hat{y}-\bar{y})^2+\sum(y-\hat{y})^2\ 。$$

从这个等式不难看到：y与x直线回归效果的好坏取决于回归平方$\sum(\hat{y}-\bar{y})^2$和与离回归平方和$\sum(y-\hat{y})^2$的大小，或者说取决于回归平方和$\sum(\hat{y}-\bar{y})^2$在y的总平方和$\sum(y-\bar{y})^2$中所占比例的大小。这个比例越大，y与x的直线回归效果就越好，反之则差。

比值 $\sum(\hat{y}-\bar{y})^2 / \sum(y-\bar{y})^2$ 叫做x对y的决定系数，记为 r^2，即

$$r^2=\frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}$$

决定系数的大小表示了回归方程估测可靠程度的高低，或者说表示了回归直线拟合度的高低，显然$0\leqslant r^2\leqslant 1$。

因为

$$r^2=\frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}=\frac{SP_{xy}{}^2}{SS_xSS_y}=\frac{SP_{xy}}{SS_x}\frac{SP_{xy}}{SS_y}=b_{yx}b_{xy}$$

而SP_{xy}/SS_x是以x为自变量、y为因变量时的回归系数b_{yx}。

若把y作为自变量、x作为因变量，则回归系数$b_{xy}=SP_{xy}/SS_y$。

所以决定系数r^2等于y对x的回归系数与x对y的回归系数的乘积。

这就是说，决定系数反应了x为自变量、y为因变量和y为自变量、x为因变量时两个相关变量x与y直线相关的信息，即决定系数表示了两个互为因果关系的相关变量间直线相关的程度。但决定系数介于0和1之间，不能反应直线关系的性质——是同向增减或是异向增减。

若求r^2的平方根，且取平方根的符号与乘积和SP_{xy}的符号一致，即与b_{xy} 、b_{yx}的符号一致，这样求出的平方根既可表示y与x的直线相关的程度，也可表示y与x直线相关的性质。

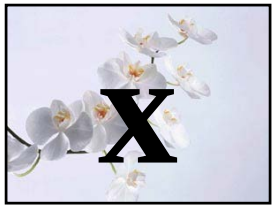
统计学上把这样计算所得的统计数称为x与y的相关系数，记为r，即

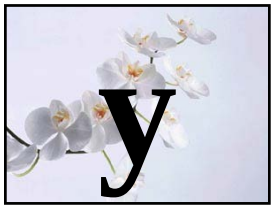
$$r=\frac{SP_{xy}}{\sqrt{SS_xSS_y}} \qquad (7\text{-}17)$$

$$=\frac{\sum xy-\frac{(\sum x)(\sum y)}{n}}{\sqrt{\left[\sum x^2-\frac{(\sum x)^2}{n}\right]\left[\sum y^2-\frac{(\sum y)^2}{n}\right]}} \qquad (7\text{-}18)$$

显然 $-1\leqslant r\leqslant 1$。当$r<0$时，相关变量x与y异向增减，称为x与y负相关；当$r>0$时，相关变量x与y同向增减，称为x与y正相关。

相关与回归的联系





回归方程的显著性

$$\hat{y} = a + bx$$

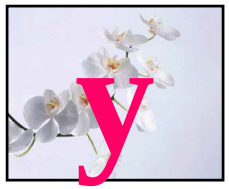
回归系数的显著性

一致

相关系数的显著性

相关与回归的区别

应用



回归 两变量间依存变化的数量关系

相关 两变量间相关关系

回归系数与相关系数的正负号都由两变量离均差积之和的符号决定，所以同一资料的b与其r的符号相同。

回归系数有单位，形式为（应变量单位/自变量单位），相关系数没有单位。

相关系数的范围在-1～+1之间，而回归系数没有这种限制。

有些资料用相关表示较适宜，比如兄弟与姐妹间的身长关系、人的身长与前臂长之间的关系等资料。

有些资料用相关和回归都适宜，此时须视研究需要而定。

就一般计算程序来说，是先求出相关系数r并对其进行假设检验，如果r显著并有进行回归分析之必要，再建立回归方程。

注意问题

** 作相关与回归分析要有实际意义。

不要把毫无关联的两个事物或现象用来作相关或回归分析。

如儿童身高的增长与小树的增长，作相关分析是没有实际意义的，如果计算由儿童身高推算小树高的回归方程则更无实际意义。也许算得的r、b是显著的，也是没有意义的。

注意问题

* 对相关分析的作用要正确理解。

相关分析只是以相关系数来描述两个变量间相互关系的密切程度和方向，并不能阐明两事物或现象间存在联系的本质。

相关并不一定就是因果关系，切不可单纯依靠相关系数或回归系数的显著性“证明”因果关系之存在。

要证明两事物间的因果关系，必须凭籍专业知识从理论上加以阐明。但是，当事物间的因果关系未被认识前，相关分析可为理论研究提供线索。

注意问题

** 适合相关和回归分析的资料通常有两种

1 一个变量X是选定的，另一个变Y是从正态分布的总体中随机抽取的。

回归分析

注意问题

2

两变量X、Y（或X_1、X_2）都是从正态分布的总体中随机抽取的，即是正态双变量中的随机样本。

由一个变量推算另一个变量 → 回归分析

说明两变量间的相互关系 → 相关分析

注意问题

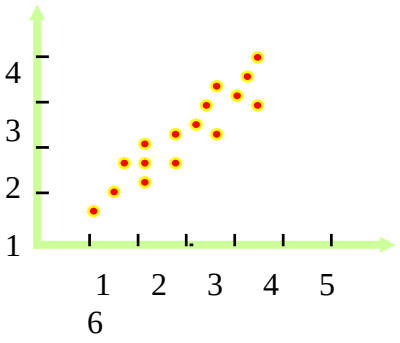
在回归分析中，由X推算Y与由Y推算X的回归方程是不同的，不可混淆。

必须正确选定自变量与应变量。

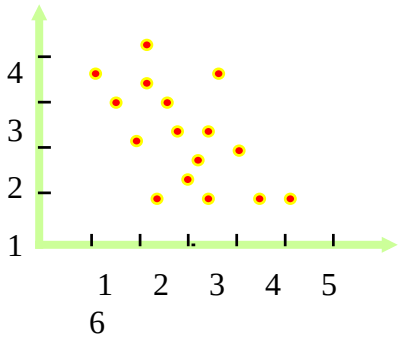
一般说，事物的原因作自变量X，当事物的因果关系不很明确时，选误差较小的即个体变异小的变量作自变量X，以推算应变量Y。

注意问题

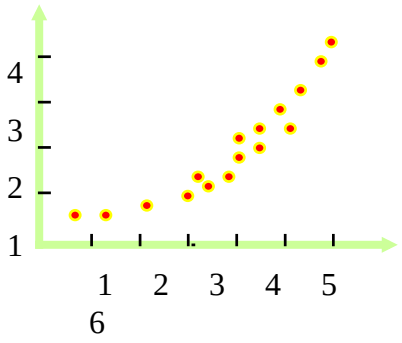
回归方程的适用范围有其限度，一般仅适用于自变量X的原数据范围内，而不能任意外推。因为我们并不知道在这些观察值的范围之外，两变量间是否也呈同样的直线关系。



正向直线关系



负向直线关系



曲线关系

直线关系是两变量间最简单的一种关系。

这种关系仅在变量的一定取值范围内可用，范围过大，散点图就偏离直线，需要借助于曲线描述。

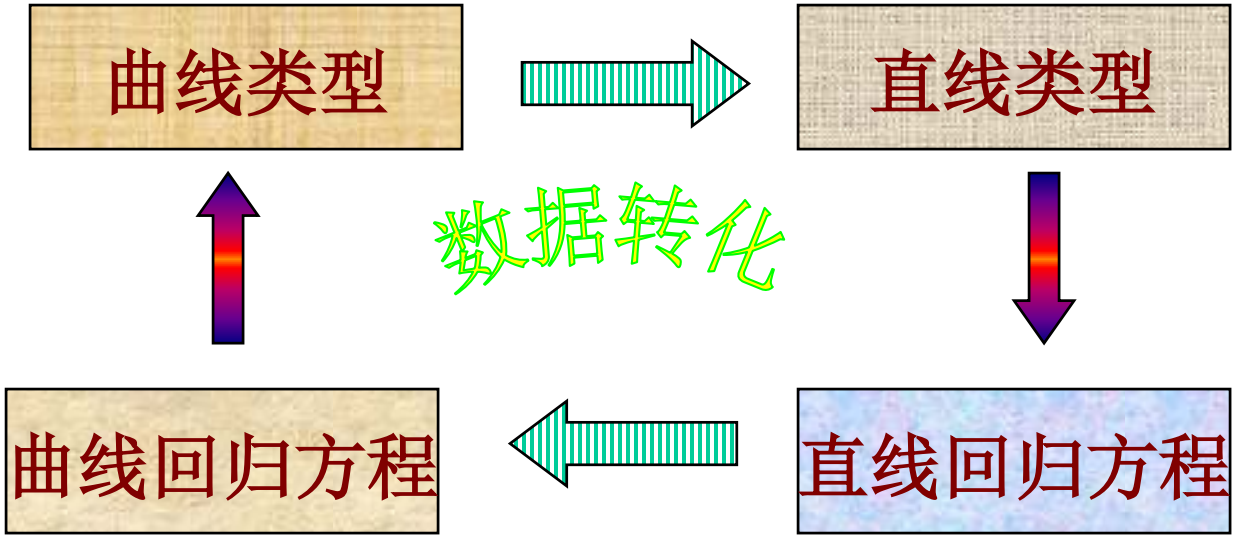
如果缩小研究范围，则任意非直线关系最后都可以用线性关系来近似，但范围过小，使用上不方便。

(1)不能对变量间的关系有一个整体上的认识。

(2)在不同取值范围内还要换用不同的方程。

两变量间的非线性关系

用来表示双变量间的关系有多种曲线。



SPSS中的Linear regression过程

有人研究了黏虫孵化历期平均温度$(x, ℃)$与历期天数(y, d)之间的关系，试验资料如下表。试建立直线回归方程。

X	**11.8**	**14.7**	**15.6**	**16.8**	**17.1**	**18.8**	**19.5**	**20.4**
Y	30.1	17.3	16.7	13.6	11.9	10.7	8.3	6.7

Descriptive Statistics

	Mean	Std. Deviation	N
历期y	14.4125	7.34136	8
平均温度x	16.8375	2.80761	8

Correlations

		历期y	平均温度x
Pearson Correlation	历期y	1.000	-.968
	平均温度x	-.968	1.000
Sig. (1-tailed)	历期y	.	.000
	平均温度x	.000	.
N	历期y	8	8
	平均温度x	8	8

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.968[a]	.937	.927	1.98376	.937	89.868	1	6	.000	1.863

a. Predictors: (Constant), 平均温度x

b. Dependent Variable: 历期y

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	353.657	1	353.657	89.868	.000[b]
	Residual	23.612	6	3.935		
	Total	377.269	7			

a. Dependent Variable: 历期y

b. Predictors: (Constant), 平均温度x

Coefficients[a]

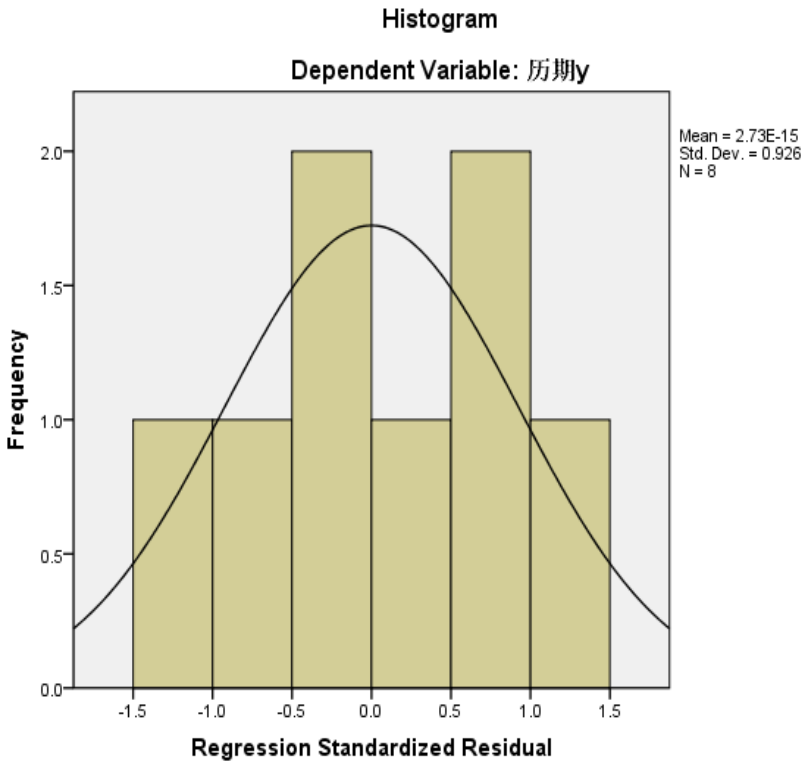
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	57.039	4.551		12.534	.000	45.904	68.175
	平均温度x	-2.532	.267	-.968	-9.480	.000	-3.185	-1.878

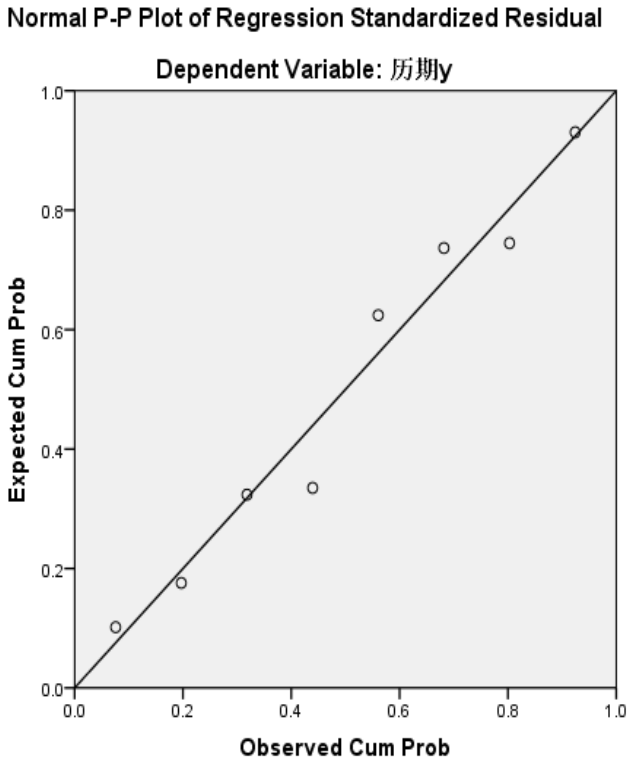
a. Dependent Variable: 历期y

Residuals Statistics[a]

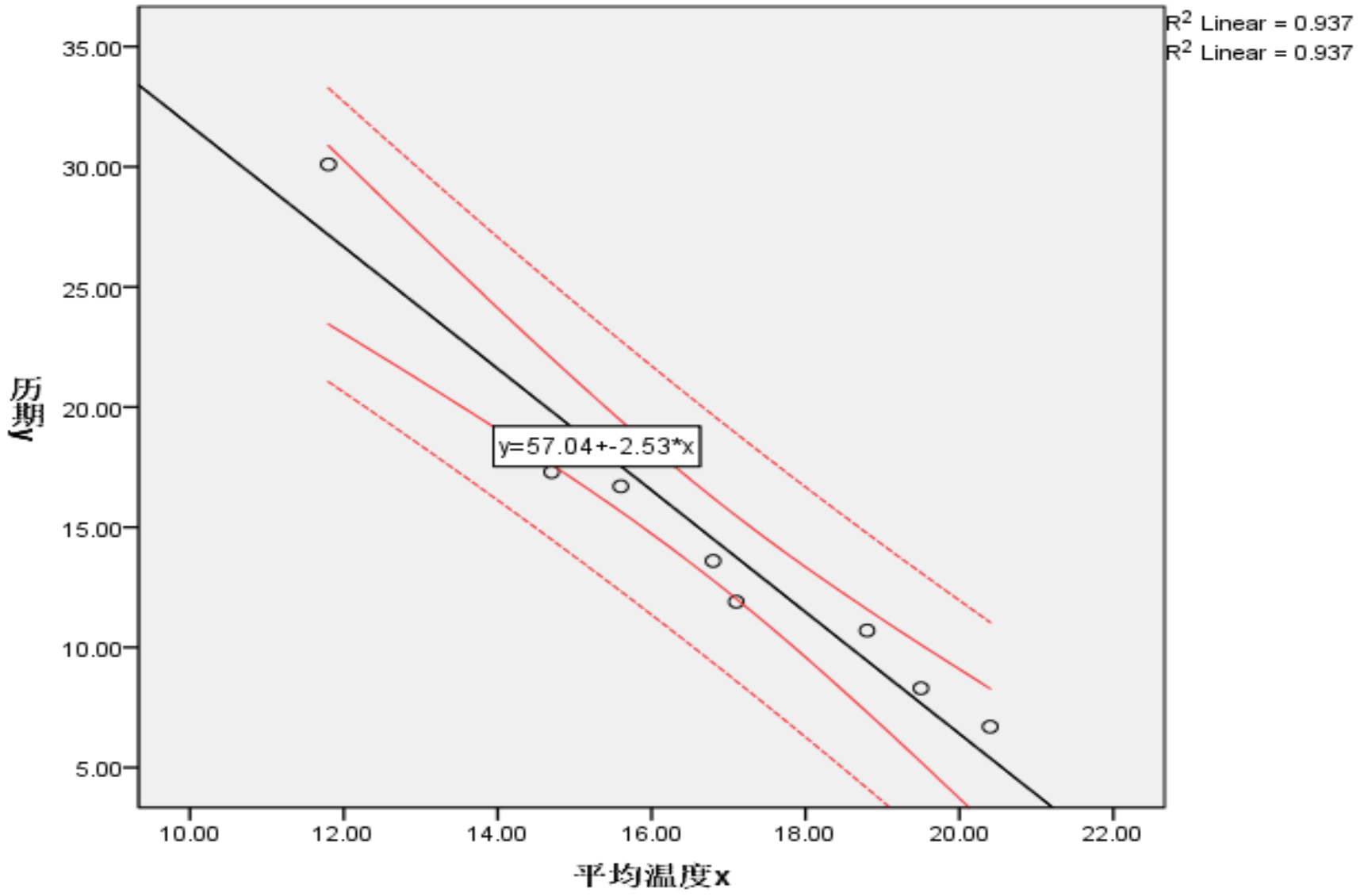
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	5.3935	27.1657	14.4125	7.10791	8
Std. Predicted Value	-1.269	1.794	.000	1.000	8
Standard Error of Predicted Value	.701	1.517	.957	.277	8
Adjusted Predicted Value	4.6744	23.0313	13.8926	6.44751	8
Residual	-2.52392	2.93427	.00000	1.83661	8
Std. Residual	-1.272	1.479	.000	.926	8
Stud. Residual	-1.429	2.296	.101	1.194	8
Deleted Residual	-3.18597	7.06874	.51994	3.20140	8
Stud. Deleted Residual	-1.607	6.011	.541	2.359	8
Mahal. Distance	.000	3.219	.875	1.086	8
Cook's Distance	.017	3.713	.545	1.283	8
Centered Leverage Value	.000	.460	.125	.155	8

a. Dependent Variable: 历期y





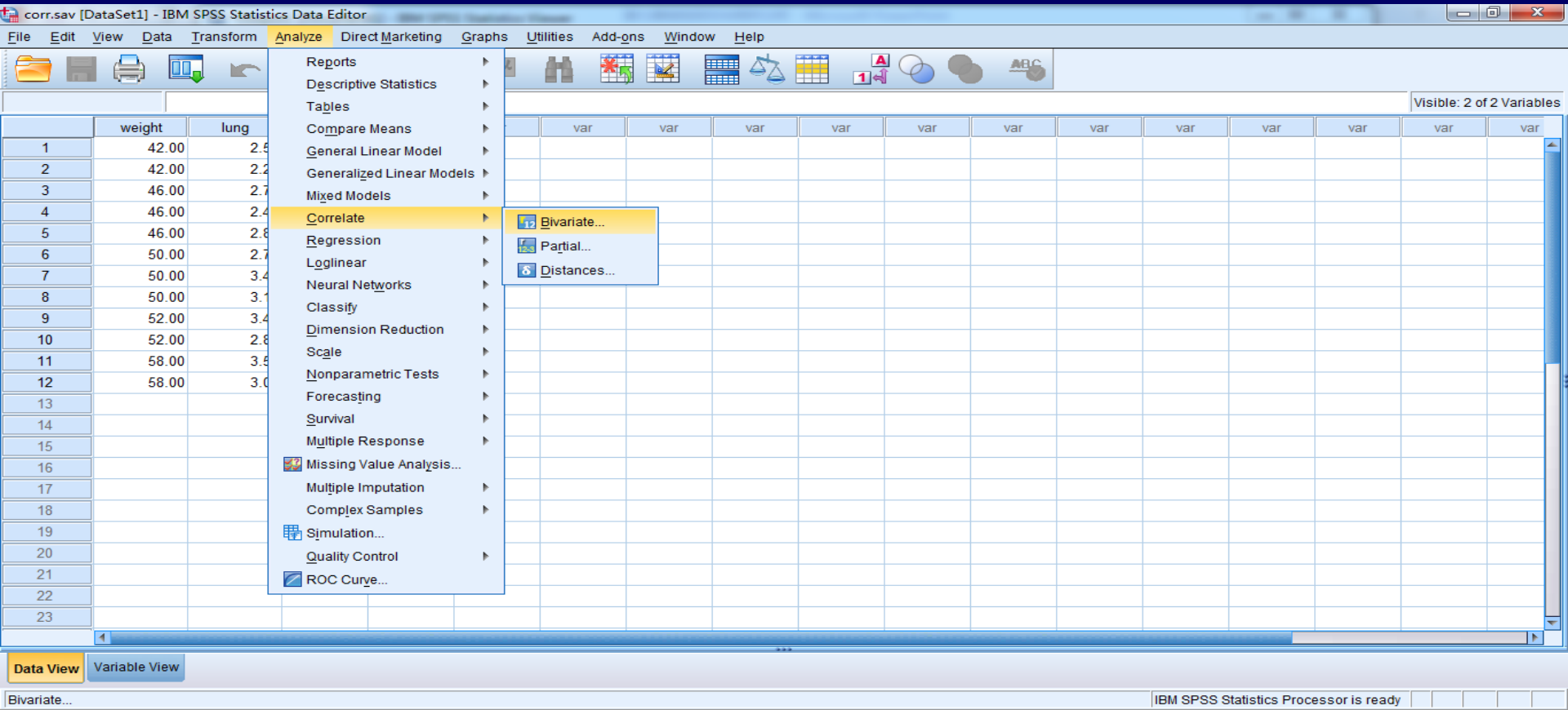
个体与均值的置信区间


R2 Linear = 0.937
R2 Linear = 0.937
y=57.04+-2.53*x
历期y
35.00
30.00
25.00
20.00
15.00
10.00
5.00
10.00
12.00
14.00
16.00
18.00
20.00
22.00
平均温度x

相关分析：SPSS中的Correlate过程

例：某地一年级12名女大学生的体重是以肺活量数据如下，试分析两者的直线相关关系

体重	42	42	46	46	46	50	50	50	52	52	58	58
肺活量	2.55	2.20	2.75	2.40	2.80	2.81	3.41	3.10	3.46	2.85	3.50	3.00



参考书

- 生物统计学（第四版），杜荣骞，高等教育出版社，2015
- 生物统计学，谢邦昌，赵雅婷，邬宏潘，耿直，中国统计出版社，2003
- 生物统计（第二版），北京师范大学数学科学学院主编，李仲来，刘来福，程书肖编著，北京师范大学出版社，2012