《医学统计学》第十次课

回归分析
Regression Analysis

Department of Health Statistics

变量

某市1995年104名男童身高（cm）资料如下

117.3	119.6	121.9	125.1	117.0	115.4	124.7	120.1	123.0	122.8
120.6	121.5	125.0	125.9	123.2	126.6	122.0	127.6	125.1	120.1
119.5	126.1	126.4	125.6	118.9	130.4	124.9	125.8	126.1	120.9
116.1	124.0	124.6	118.7	119.1	121.9	118.0	117.0	114.6	123.9
116.0	125.3	123.6	123.6	126.4	115.5	119.2	114.0	123.4	126.6
117.3	113.6	127.6	120.5	113.6	130.2	128.3	118.2	124.7	122.4
118.8	123.1	122.7	126.6	127.8	125.9	110.5	124.8	115.2	119.4
128.0	116.7	132.4	129.3	121.7	115.0	120.4	122.1	127.0	135.3
125.7	111.2	124.3	124.2	124.7	121.7	121.3	124.1	119.9	121.7
113.8	116.7	129.9	128.5	126.5	122.8	120.1	118.2	122.5	127.7
124.9	123.3	120.3	125.7						

例：采用配对设计将两只同窝、同性别、体重相似的大鼠配成对子，共8对，研究不同饲料大鼠肝中维生素A的含量。

变量

不同饲料大鼠肝中维生素A含量

对子号	正常饲料组	缺乏维生素E组
1	1.07	0.74
2	0.60	0.72
3	0.90	0.54
4	1.19	0.98
…	…	…
8	0.92	0.53

医学与医院管理中的其它问题

- 医学研究中：

食盐摄入量与血压的关系；吸烟量与期望寿命的关系；

• • •

- 医院管理中：

就诊人数与医院收入的关系；住院人数与病床周转率的关系；• • •

相关性技术（correlational techniques）：

相关：分析两个变量之间关系的强度和方向；

回归：用方程分析两个变量之间的关系，实现估计和预测。

回归（regression）的由来

- “维多利亚女王时代最博学的人”
- “普用回归定律”：每个人的特征是与其亲属共有的，但平均来说在程度上略差一点。
- 高尔顿对统计学的最大贡献是相关性概念的提出和回归分析方法的建立。

回归分析

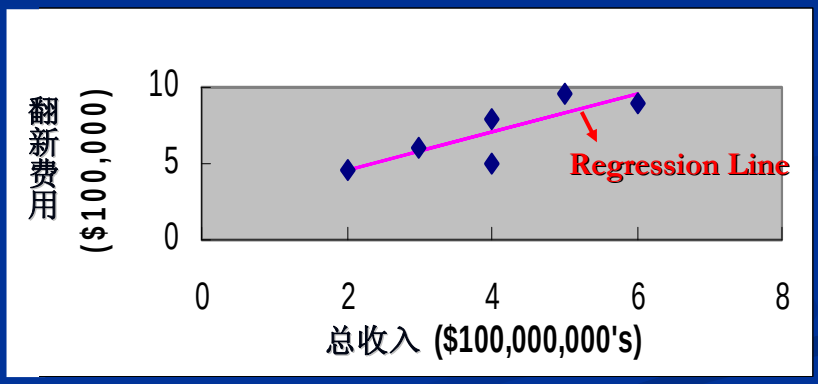
- 两相关变量的散点图
- 回归方程
- 回归系数的假设检验与区间估计
- 预测值的区间估计

例补2 某医院想翻新位于奥尔巴尼的旧病房。经调查发现翻新的费用($)与当地的总收入($)有关。如何估计翻新所需要的费用？

表2 某医院旧病房的翻新费用与当地的总收入

翻新费用 ($100,000's)	当地总收入 ($100,000,000's)
6	3
8	4
9	6
5	4
4.5	2
9.5	5

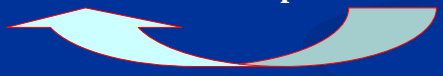
翻新费用（Y）与当地总收入（X）的散点图



线性回归模型

回归模型由一个应变量（dependent variable）和一个（或多个）自变量（independent variable）组成，它定量地表达了应变量依赖自变量的变化而变化的线性关系。

Dependent Variable = Independent Variable(s)



Prediction Relationship

根据自变量个数的多少分为：

1、只含一个自变量：简单线性回归（simple linear regression）

2、含两个（以上）个自变量：多重线性回归（multiple linear regression）

简单线性回归模型的一般形式：

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Y_i: 第i个个体的应变量值

X_i: 第i个个体的自变量值

α: 截距参数（intercept parameter）

β: 总体回归系数（population regression coefficient）

ε_i: 随机误差

本例，简单线性回归模型表示该医院第i年的旧病房翻新费用等于奥尔巴尼当地总收入的β倍加上α，再加上一点随机误差。

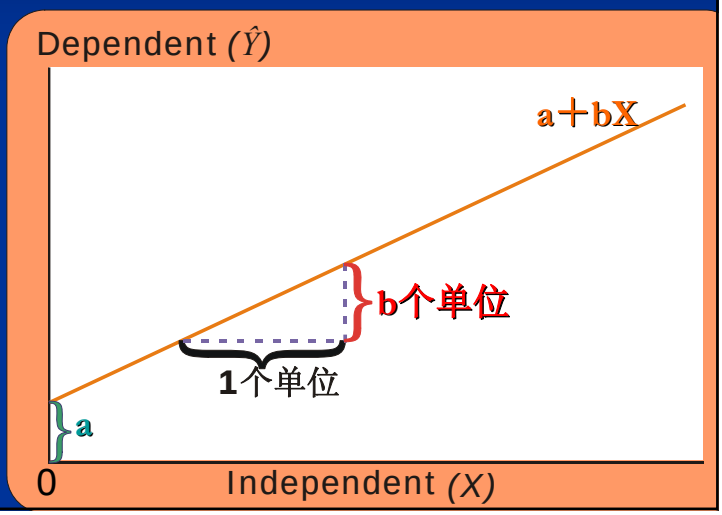
基于样本的回归方程（regression equation）：

$$\hat{Y} = a + bX$$

a：α的估计

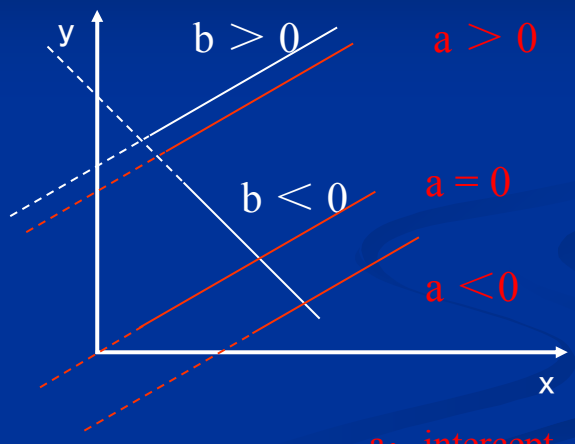
b：样本回归系数，β的估计

ŷ：与X相对应的Y的平均值



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

b：regression coefficient



a：intercept

如何估计线性回归方程？

线性回归方程： $\hat{Y}=a+bX$

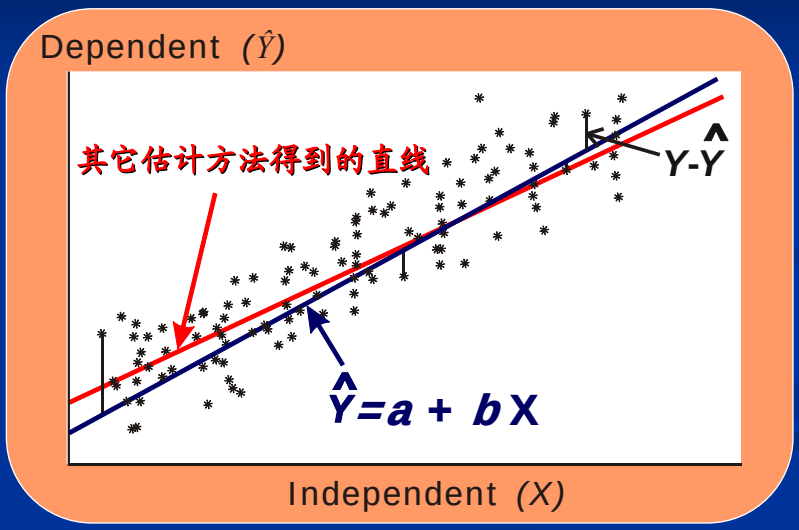
真值与估计值之差—残差(residual)：

$e=Y-\hat{Y}$

最小二乘法 (least square method)：

使$\sum e^2=\sum(Y-\hat{Y})^2$最小。

最小二乘回归



a、b的最小二乘计算公式

$$b = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2} = \frac{l_{XY}}{l_{XX}}$$

$$a = \overline{Y} - b\overline{X}$$

P187 例12-2： 用某饲料喂养12只大白鼠，得出大白鼠的进食量与体重增加量如表12-1，对大白鼠的进食量与体重增加量进行回归分析。

表12-1 12只大白鼠的进食量与体重增加量

序号	进食量（g）	体重增加量（g）
1	305.7	23.6
2	188.6	14.7
3	277.2	19.2
4	364.8	27.7
5	285.3	18.9
6	244.7	16.1
7	255.9	17.2
8	149.8	12.9
9	268.9	18.3
10	247.6	17.7
11	168.8	13.7
12	200.6	15.6
合计	2957.9	215.6

1. 绘制散点图——观察两变量间的关系

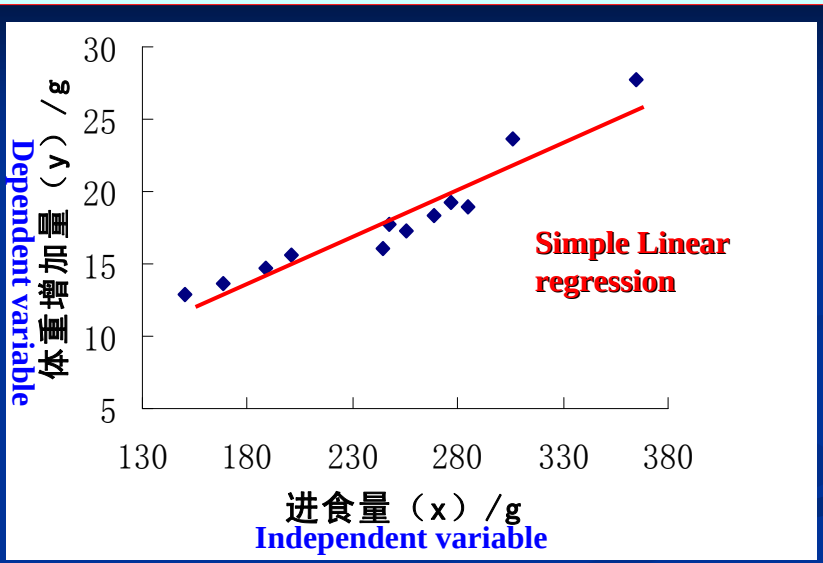


图12-1 12只大白鼠的进食量与体重增加量

2. 求回归系数b和截距a

$$b = \frac{l_{XY}}{l_{XX}} = 0.0648$$

$$a = \overline{Y} - b\overline{X} = 2.00$$

3. 列出线性回归方程

$$\hat{Y} = 2.00 + 0.0648X$$

Q: b≠0能否说明总体中进食量（X）与体重增加量（Y）有上述线性关系存在？

否

A: 需要对总体回归系数β是否为“0”进行假设检验。

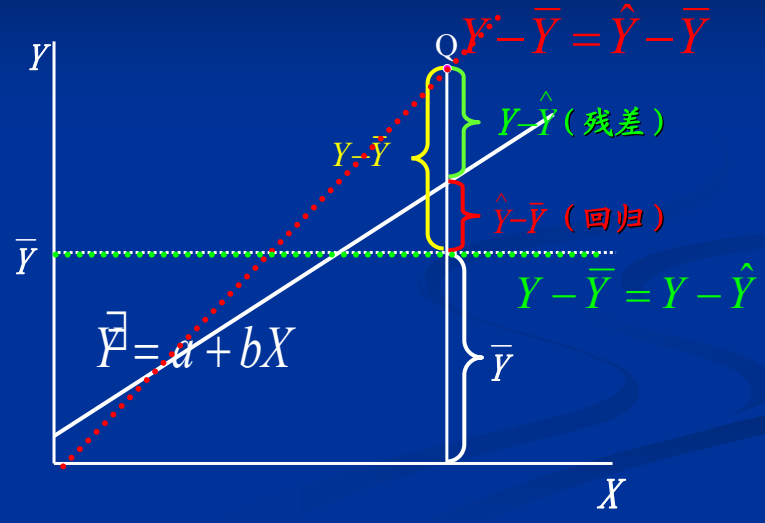
总体回归系数β的假设检验

H_0: β = 0

方差分析法:

基本思想：把总的离均差平方和(即总变异)分解为至少两个部分，其中一部分主要表示处理因素（回归）的效应，另一部分表示抽样误差（剩余）的影响，然后比较两者的均方，计算F值，若F值远大于1，可认为处理（回归）有效应，否则认为处理（回归）无效应。

应变量Y的变异分解



应变量Y的变异分解

$$Y-\bar{Y}=(\hat{Y}-\bar{Y})+(Y-\hat{Y})$$

$$\sum(Y-\bar{Y})^2=\sum[(\hat{Y}-\bar{Y})+(Y-\hat{Y})]^2$$

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

$SS_{总}$ = $SS_{回}$ + $SS_{剩（残）}$

$\nu_{总}=n-1$, $\nu_{回}=1$, $\nu_{剩}=n-2$

各离均差平方和SS的计算

$SS_{总}=l_{YY}$

$SS_{回}=b\cdot l_{XY}=l_{XY}^2/l_{XX}$

$SS_{剩}=SS_{总}-SS_{回}=l_{YY}-l_{XY}^2/l_{XX}$

$$F=\frac{SS_{回}/\nu_{回}}{SS_{剩}/\nu_{剩}}=\frac{MS_{回}}{MS_{剩}}$$

例12-3 用方差分析法对表12-1数据的总体回归系数进行假设检验。

$$\hat{Y}=2.00+0.0648X$$

H_0：$\beta=0$（大白鼠的体重增加量与进食量之间无直线回归关系）

H_1：$\beta\neq0$（大白鼠的体重增加量与进食量之间有直线回归关系）

$\alpha=0.01$

$SS_{总}=l_{YY}=193.3$

$SS_{回}=b\cdot l_{XY}=l_{XY}^2/l_{XX}=173.7$

$SS_{剩}=SS_{总}-SS_{回}=l_{YY}-l_{XY}^2/l_{XX}=19.6$

$$F=\frac{SS_{回}/\nu_{回}}{SS_{剩}/\nu_{剩}}=\frac{MS_{回}}{MS_{剩}}=88.6$$

表12-2 回归系数方差分析表

变异来源	SS	DF	MS	F	P
回归	173.7	1	173.70	88.60	P<0.01
剩余	19.6	10	1.96		
总变异	193.3	11			

按α=0.01的检验水准，拒绝H_0，接受H_1，可认为大白鼠的体重增加量与进食量之间有直线回归关系。

总体回归系数的假设检验：t检验法

$$t_b = \frac{b-0}{S_b} \qquad \upsilon = n-2$$

$$S_b = \frac{S_{YX}}{\sqrt{l_{XX}}}; \quad S_{YX} = \sqrt{\frac{\Sigma(Y-\hat{Y})^2}{n-2}} = \sqrt{\frac{SS_{剩}}{n-2}}$$

S_{YX}：剩余标准差（standard deviation of residuals），表示去除了X的影响后Y的变异

发现$\hat{Y}$依赖于X变动的实际幅度大小
——总体回归系数β的可信区间

- 总体回归系数β的点估计：b
- β的100·(1-α)%的可信区间为：

$$(b - t_{\alpha/2,\nu} s_b ,\ b + t_{\alpha/2,\nu} s_b)$$

或 $$b \pm t_{\alpha/2,\nu} s_b$$

$$\nu = n - 2,\ S_b = \frac{S_{YX}}{\sqrt{l_{XX}}}$$

P191例12-5（续12-1）试求总体回归系数β的95%可信区间。

$$\hat{Y} = 2.00 + 0.0648X$$

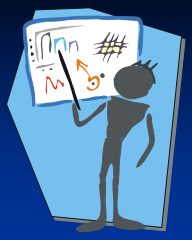
b=0.0648，ν=12–2=10，$t_{0.05/2,10}=2.228$，$S_b=0.00688$

β的95%可信区间为：

$0.0648 \pm 2.228 \times 0.00688$

=（0.0495，0.0801）

表示有95%的把握认为：大白鼠的进食量（X）每变动1g（一个单位），体重增加量（Y）随之变动的平均幅度在（0.0495g，0.0801g）之间。



H_0：β = 0的检验中P<α，或者，β的CI不包括“0”，则说明回归方程成立，但回归方程是基于样本的估计，如何反映这种估计的好坏呢？

回归方程好坏的评价：决定系数

决定系数（coefficient of determination）：Y的变异中可以被回归方程解释的部分，用R^2表示。

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (0 \le R^2 \le 1)$$

本例，R^2=173.7/193.3=89.86％，表示大白鼠体重增加量的89.86%与进食量有关。

回归方程的整体评价

- 方差分析、t检验的P值判断回归方程是否成立，即总体上X与Y的线性回归关系是否存在。
- 决定系数R^2反映了自变量X对应变量Y的影响大小。R^2值接近“1”则暗示X对Y有较强的影响。

好的回归方程应该具有低的P值（$P<\alpha$）和较高的R^2值。

运用回归方程进行预测

例补2中经线性回归分析得到该医院翻新费用（Y）与总收入（X）之间有回归方程：$\hat{y}=2+1.25X$成立，如果明年奥尔巴尼居民总收入估计为6亿美元，则该医院的翻新费用预计为多少？

翻新费用＝2＋1.25×总收入

So,

明年的翻新费用＝2＋1.25×6＝9.5

该医院明年的翻新费用预计为95万美元。

运用回归方程进行估计

例12-1中大白鼠的进食量（X）与体重增加量（Y）有回归方程：$\hat{Y}=2.0+0.0648X$成立，如果大白鼠的进食量为250g，则其体重增加量估计为多少？

体重增加量= 2.0 + 0.0648 × 进食量

So,

体重增加量 = 2.0 + 0.0648 ×250 = 18.2

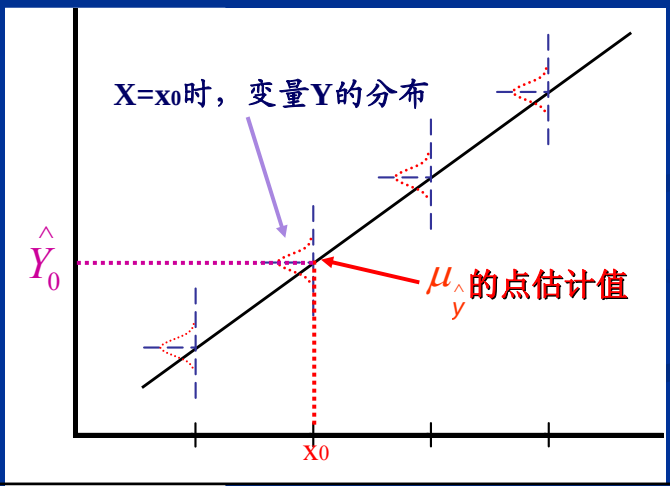
大白鼠进食量为250g时体重增加量估计为18.2g。

预测（估计）值的区间估计

- 应用回归方程进行估计和预测：
  1. X = x_0时，Y的总体均数 $\mu_{\hat{Y}}$的CI;
  2. X = x_0时，个体Y值的容许区间。

$\mu_{\hat{Y}}$ 的可信区间

$\mu_{\hat{Y}}$：X为某一定值x_0时Y的总体均数，也记为 $\mu_{Y|X}$。



- 样本条件均数 $\hat{Y}_0$ 的标准误：

$$S_{\hat{Y}_0} = S_{YX}\sqrt{\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X-\bar{X})^2}}$$

- 总体条件均数$\mu_{\hat{Y}}$的100(1-α)% CI:

$$\hat{Y}_0 \pm t_{\alpha/2,n-2}S_{\hat{Y}_0}$$

P191 例12-6（续例12-1）试计算当x_0=250时$\mu_{\hat{Y}}$的95%可信区间。

$$\hat{Y}=2.00+0.0648X$$

x_0=250时，$\hat{Y}_0$=18.2

S_{YX}=1.40，l_{xx}=41389.4，$\bar{X}$=246.49

$$S_{\hat{Y}_0}=1.40\sqrt{\frac{1}{12}+\frac{(250-246.49)^2}{41389.4}}=0.405$$

ν=12–2=10，$t_{0.05/2,10}$=2.228

$\mu_{\hat{Y}}$的95%可信区间：

18.2 ± 2.228 × 0.405 =（17.30，19.10）

表示有95%的把握认为：大白鼠的进食量（X）为250g时，其体重增加量（Y）的总体均数$\mu_{\hat{Y}}$在（17.30g，19.10g）之间。

个体Y值的容许区间

个体Y值的容许区间：总体中X为某定值x_0时，Y值由于随机误差的影响在其均数$\hat{Y}_0$上下波动的范围。

常用于估计当自变量X（如：年龄）为某一定值时，应变量Y（如：尿肌酐含量）的医学参考值范围。

- $X=x_0$时，Y值的标准差：

$$S_{Y_0} = S_{YX}\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X-\bar{X})^2}}$$

- $X=x_0$时，Y值的100(1-α)% 容许区间：

$$\hat{Y}_0 \pm t_{\alpha/2,n-2}S_{Y_0}$$

P192 例12-7（续例12-1）试计算当x_0=250时个体Y值的95%容许区间。

$$S_{Y_0}=1.40\sqrt{1+\frac{1}{12}+\frac{(250-246.49)^2}{41389.4}}=1.457$$

ν=12–2=10，$t_{0.05/2,10}$=2.228

个体Y值的95%容许区间：

18.2 ± 2.228 × 1.457 =（14.95，21.44）

表示：当大白鼠的进食量（X）为250g时，有95%的大白鼠其体重增加量（Y）在（14.95g，21.44g）之间。

回归分析中涉及的几个符号辨析

S_{YX}：剩余标准差

R^2：决定系数

$S_{\hat{Y}_0}$：样本条件均数的标准误，即X=x_0时，$\hat{Y}_0$ 的标准误

S_{Y_0}：X=x_0时，应变量Y的标准差

相关与回归的区别

1. 应用与意义：

相关：分析两变量的相互关系，两个变量中，任何一个的变化都会引起另一个的变化，是一种双向变化的关系。

回归：分析两个变量的依存关系，自变量的改变会引起应变量的变化，是一种单向变化的关系。

相关与回归的区别

2. 研究性质：

相关：对两个变量之间的线性关系进行描述，看两个变量是否有关，关系是否密切，是正相关还是负相关。

回归：对两个变量定量描述，研究两个变量的数量关系，已知一个变量值可以预测出另一个变量值，可以得到定量结果。

相关与回归的区别

3. 相关系数r与回归系数b的绝对值反映的意义不同：

r的绝对值越大，散点图中的点越趋向于一条直线，表明两变量的关系越密切，相关程度越高；
b的绝对值越大，回归直线越陡，说明当X变化一个单位时，Y的平均变化就越大。

相关与回归的联系

关系：

能进行回归分析的变量之间必然存在相关关系。

相关系数r与回归系数b：

1、符号一致；

2、假设检验结果一致，即可用ρ的显著性检验代替β的显著性检验。

应用相关与回归的注意事项

1. 相关回归分析一定要有*实际意义*，强的相关关系*不一定是因果*关系；

案例一：

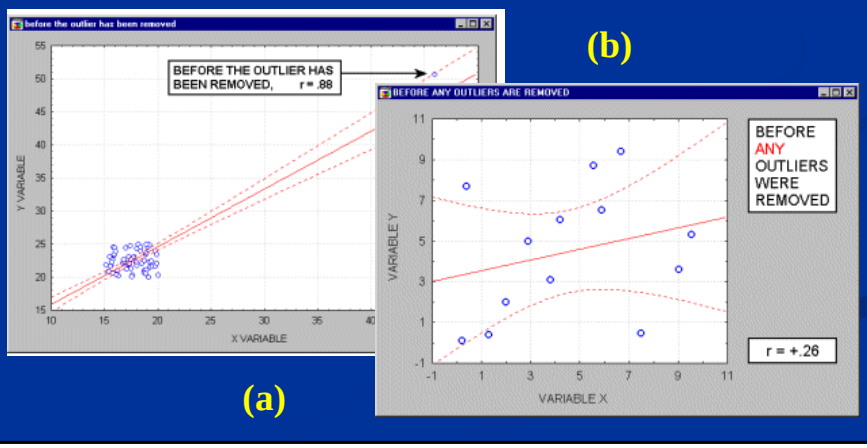
当样本足够大时，身高Y与家庭中的每月用电量X的线性回归关系具有统计学意义（β的假设检验P<0.05）。

案例二：

你的年薪与汽车的价格之间可能有很强的相关关系，但是两者彼此之间不会相互影响。

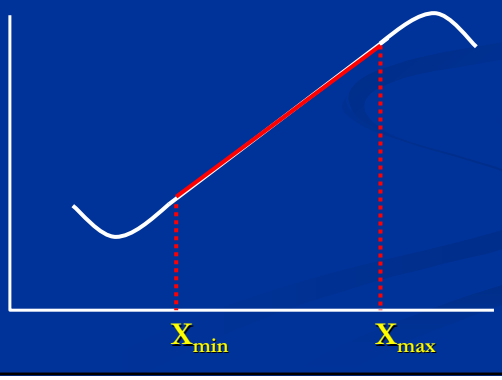
应用相关与回归的注意事项

2. 应绘制散点图：发现两变量之间的关系以及离群值（outliers）。



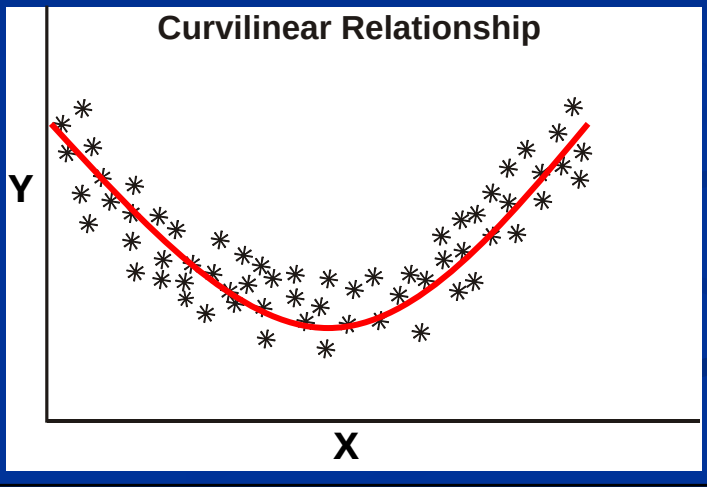
应用相关与回归的注意事项

3. 直线回归方程的适用范围以自变量的取值范围为限，应避免外延。



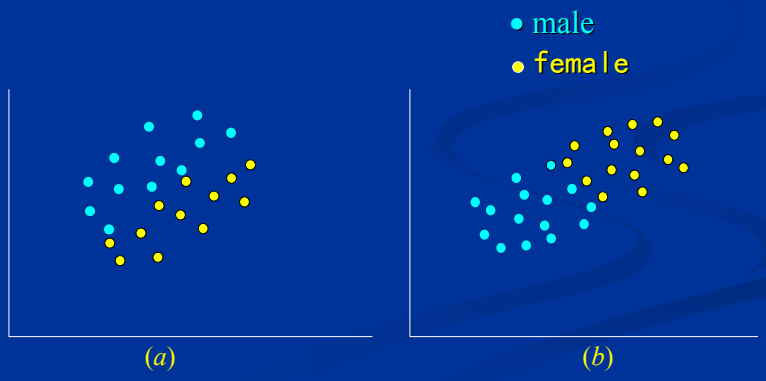
应用相关与回归的注意事项

4. 无线性相关关系（P>α）不代表不相关。



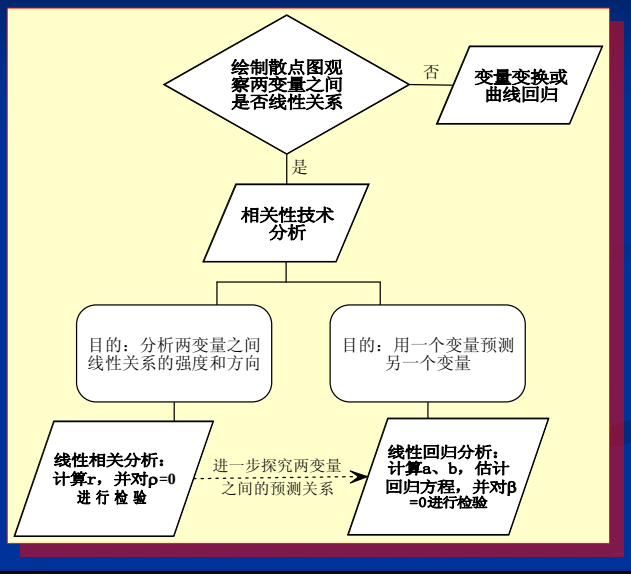
应用相关与回归的注意事项

5. 相关分析时，应审慎对待样本的*合并与分层*。



小 结

一、应用相关与回归的一般步骤。



小 结

二、线性相关分析两变量之间线性关系的强度和方向，要求X、Y均服从正态分布。线性相关系数r的大小表示两变量数量上线性关系的强度（$|r|\leq 1$），正负表示方向。样本相关系数r是总体相关系数ρ的点估计。

三、当X或Y不满足正态分布的条件时，可计算Spearman秩相关系数r_s描述两变量间的等级相关关系，需对ρ_s进行检验。

小 结

四、线性回归分析两变量间的依存关系。主要用途是预测和估计。

五、线性回归有3种区间估计和1个重要指标：

1、β的CI：估计β值所在的范围；

2、μ的CI：估计当X为某定值x_0时，Y的总体均数所在的范围；

3、个体Y值的容许区间：估计总体中，当X为某定值x_0时，个体Y值的波动范围，常用于确定医学参考值范围。

4、决定系数R^2：Y的变异中可以由回归直线解释的部分。

1、借助计算器计算相关与回归

2、作业

P201：三-1

P212：三-1

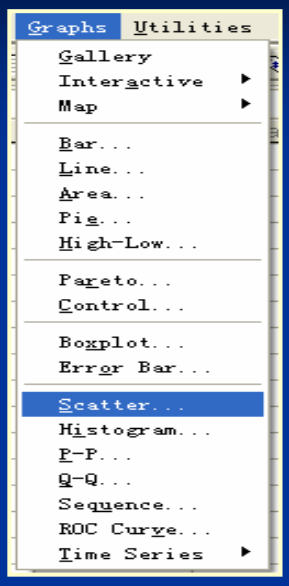
回归分析的SPSS实现

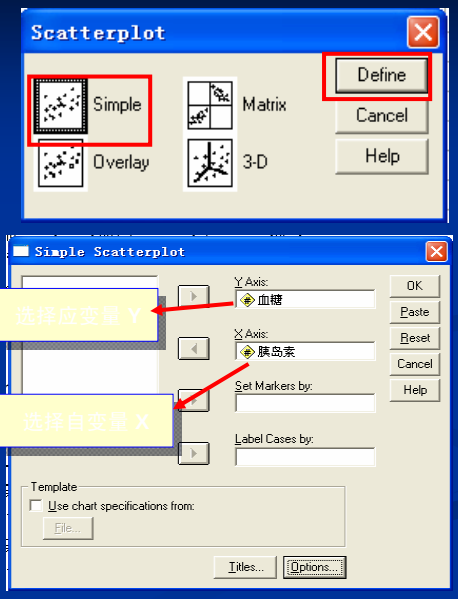
1. 建立数据集：

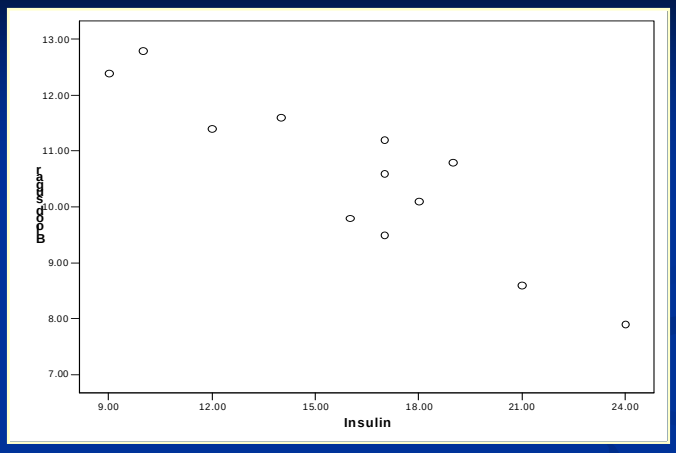
12名糖尿病患者血糖和胰岛素的测量结果

	胰岛素	血糖
1	17.00	9.50
2	14.00	11.60
3	19.00	10.80
4	12.00	11.40
5	9.00	12.40
6	16.00	9.80
7	18.00	10.10
8	21.00	8.60
9	24.00	7.90
10	17.00	11.20
11	17.00	10.60
12	10.00	12.80
13		

2. 绘制散点图：

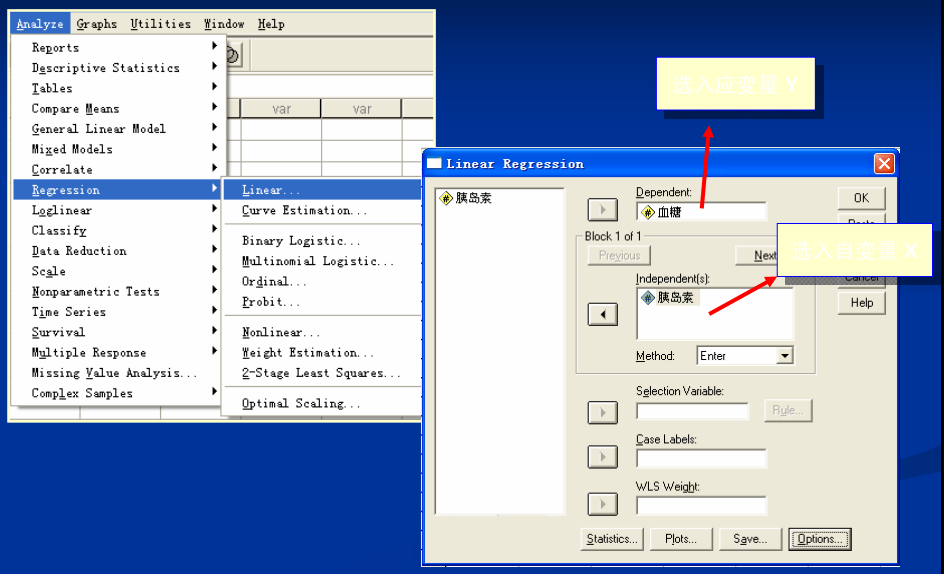






由该散点图可以看出，血糖与胰岛素之间有明显的直线回归关系

3. 回归分析:



4. 结果分析:

回归系数方差分析表

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	19.181	1	19.181	44.526	.000[a]
	Residual	4.308	10	.431		
	Total	23.489	11			

回归

剩余

a. Predictors: (Constant), 胰岛素

b. Dependent Variable: 血糖

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	15.448	.757		20.410	.000
	胰岛素	-.302	.045	-.904	-6.673	.000

常数项a

a. Dependent Variable: 血糖

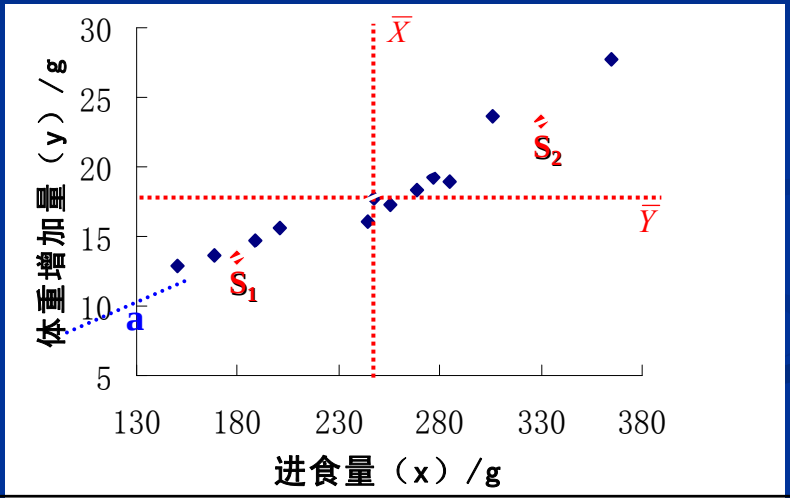
未经标准化的回归系数b

回归直线：$\hat{Y}$ =15.448-0.302 X

回归方程的图示

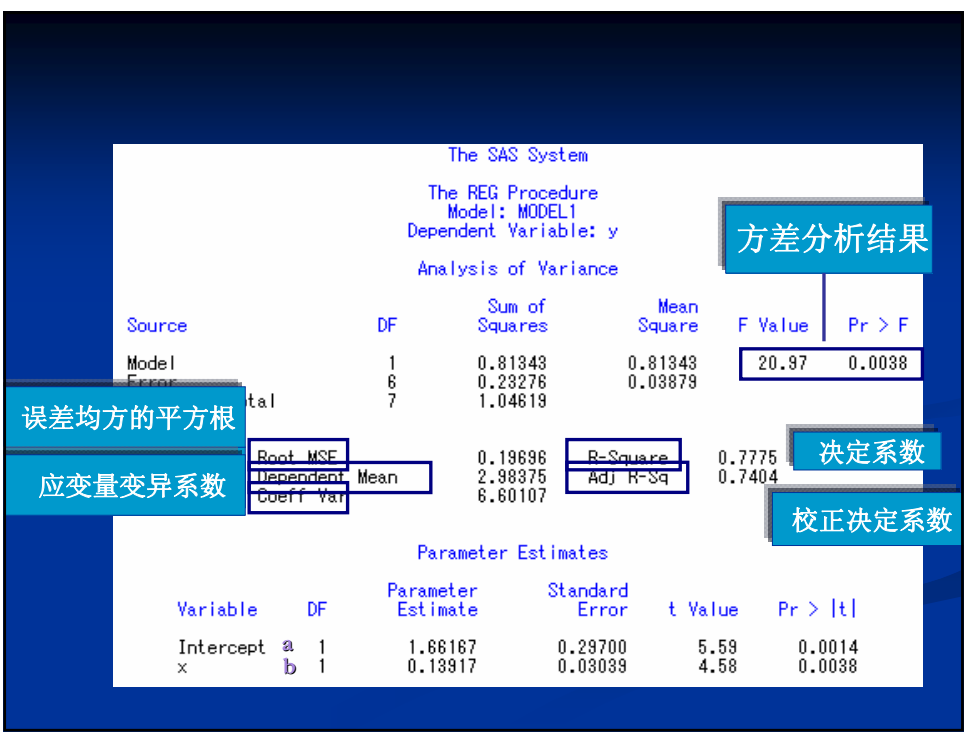
X：149.8~364.8

如：S_1（180，13.66）、S_2（330，23.38）



SAS程序 直线回归分析

```
data prg001;
  input x y @@;
cards;
13 3.54 11 3.01 9 3.09 6 2.48 8 2.56 10 3.36 12 3.18 7 2.65
;
proc reg;
  model y=x;
run;
```



回归方程为：

$$\hat{y} = 1.661667 + 0.139167x$$

常用选项：

- stb：输出标准化偏回归系数，语句为“model y=x/stb;”

Standardized Estimate
0
0.88177

常用选项:

- p: 输出每个观测的实际值、预测值和残差。语句为:

  “model y=x/p;”

应变量原始值

原始值的预测值

残　差

Obs	Dep Var y	Predicted Value	Residual
1	3.5400	3.4708	0.0692
2	3.0100	3.1925	-0.1825
3	3.0900	2.9142	0.1758
4	2.4800	2.4967	-0.0167
5	2.5600	2.7750	-0.2150
6	3.3600	3.0533	0.3067
7	3.1800	3.3317	-0.1517
8	2.6500	2.6358	0.0142

Sum of Residuals	0
Sum of Squared Residuals	0.23276
Predicted Residual SS (PRESS)	0.34220