

卫生统计学实习

何平平

北京大学公共卫生学院
流行病学与卫生统计学系

Tel: 82801619

实习六

数值变量资料的统计推断（三）

第237 ~ 249页

一、直线回归（linear regression）

（一）定义：用直线方程表达 **X** （自变量，independent variable）和 **Y** （应变量，dependent variable）之间的数量关系。

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

\hat{Y} ：是 **Y** （实测值）的预测值（predicted value），是直线上点的纵坐标。对于每一个 **X** 值，根据直线回归方程都可以计算出相应的 **Y** 预测值。

直线回归的重要应用之一：预测（Prediction）

一、直线回归 (linear regression)

(二) b 和 a 的意义

a : 是回归直线在 Y 轴上的截距, 即 $X=0$ 时 Y 的预测值。

b : 是回归直线的斜率, 又称为回归系数。

表示当 X 改变一个单位时, Y 的预测值平均改变 $|b|$ 个单位。

一、直线回归 (linear regression)

(三) b 和 a 的估计

最小二乘法 (the method of least squares) :
各实测点到直线的纵向距离的平方和最小。

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

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

$$\sum (X - \bar{X})(Y - \bar{Y}) = \sum XY - \frac{(\sum X \sum Y)}{n}$$
$$\sum (X - \bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n}$$

一、直线回归 (linear regression)

(四) b 的假设检验: b 为样本回归系数, 由于抽样误差, 实际工作中 b 一般都不为0。要判断直线回归方程是否成立, 需要检验总体回归系数 β 是否为0。

$$H_0: \beta=0 \quad H_1: \beta \neq 0$$

方法一: t 检验

$$t = \frac{b}{S_b}$$

方法二: F 检验

$$F = \frac{MS_{\text{回归}}}{MS_{\text{剩余}}}$$

两种方法等价,

$$\sqrt{F} = t$$

只有当 $\beta \neq 0$ 时, 才能认为直线回归方程成立 (具有统计学意义)。

一、直线回归 (linear regression)

(五) 直线回归方程的置信区间估计

1. 总体回归系数 β 的95% 置信区间估计

$$b \pm t_{0.05/2, n-2} s_b$$

$$s_b = s_{YX} / \sqrt{l_{xx}}$$

$$s_{YX} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n-2}} = \sqrt{\frac{SS_{\text{剩余}}}{n-2}}$$

s_{YX} 称为剩余标准差或者残差标准差

(the standard deviation of residual)

一、直线回归 (linear regression)

(五) 直线回归方程的置信区间估计

2. $\mu_{\hat{Y}}$ 的95% 置信区间估计

当 $X = X_0$ 时, 以95% 的概率估计 Y 的均数的置信区间为

$$\hat{Y} \pm t_{0.05/2, n-2} s_{\hat{Y}}$$

$$s_{\hat{Y}} = s_{YX} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{l_{XX}}}$$

一、直线回归 (linear regression)

(五) 直线回归方程的置信区间估计

3. 个体 Y 值的 95 % 容许区间估计

当 $X = X_0$ 时, 以 95 % 的概率估计个体 Y 值的波动范围为

$$\hat{Y} \pm t_{0.05/2, n-2} S_{Y-\hat{Y}}$$

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

二、直线相关 (linear correlation)

(一) 定义

描述具有直线关系的两个变量之间的相互关系。

r : 相关系数, **correlation coefficient**

用来衡量有直线关系的两个变量之间相关的密切程度和方向。 $-1 \leq r \leq 1$

$r > 0$, 正相关; $r = 1$ 为完全正相关

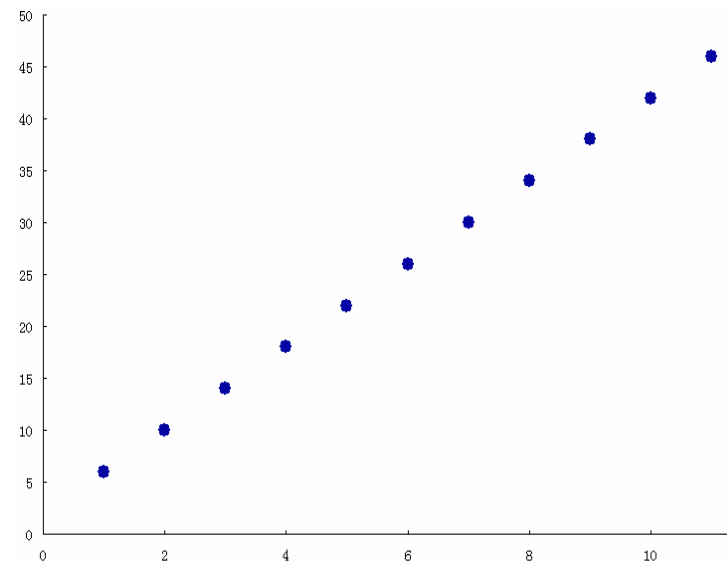
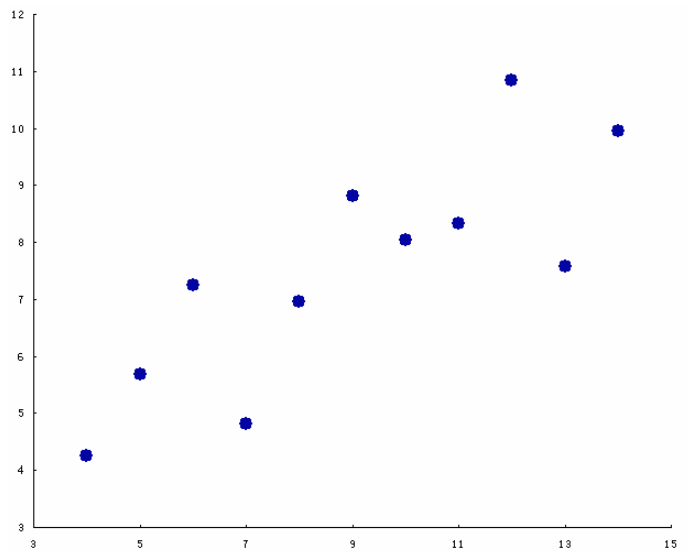
$r < 0$, 负相关; $r = -1$ 为完全负相关

$|r|$ 越大, 两变量相关越密切 (前提: r 有统计学意义)

二、直线相关（linear correlation）

（二）相关类型

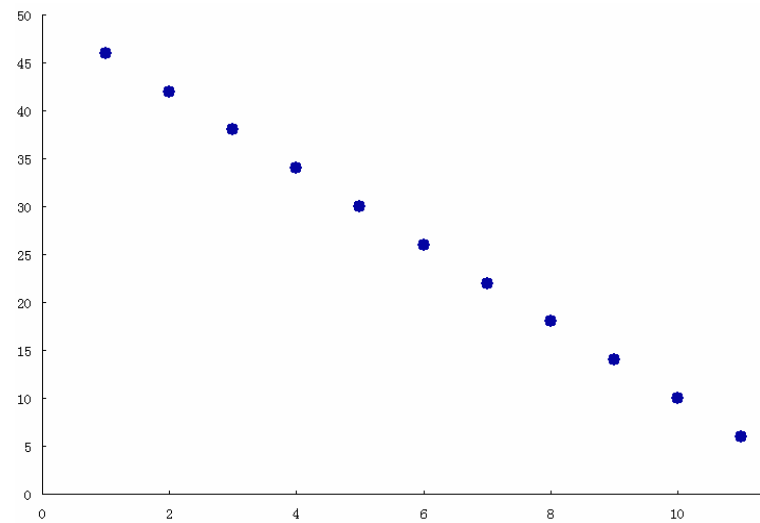
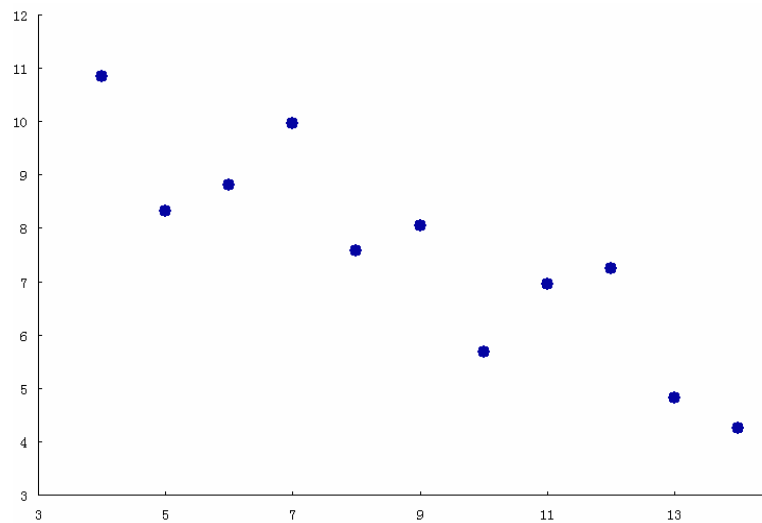
正相关： $0 < r \leq 1$



二、直线相关（linear correlation）

（二）相关类型

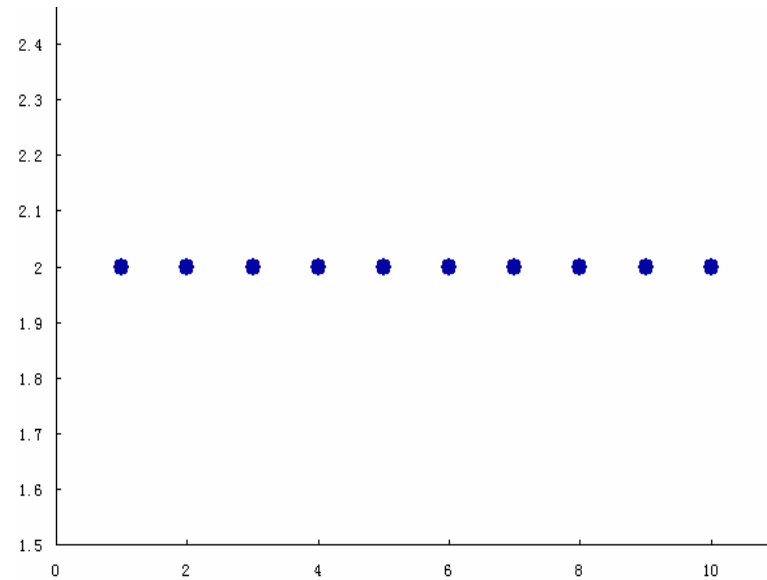
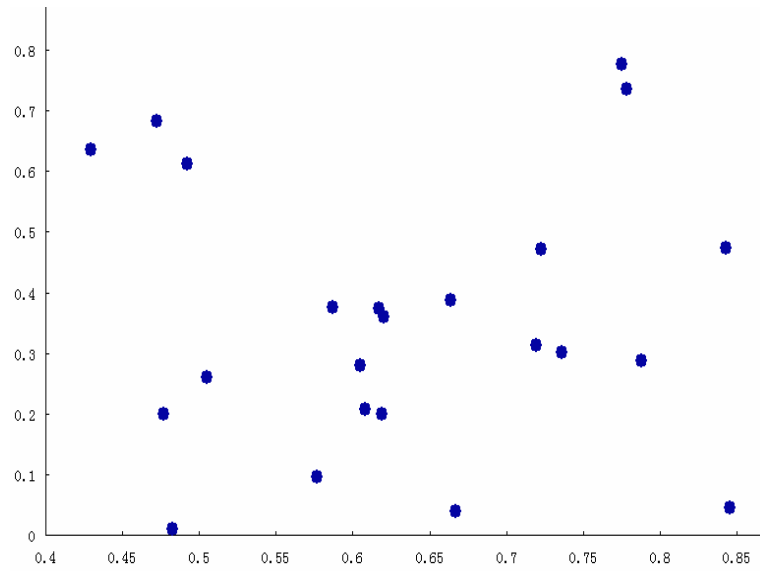
负相关 $-1 \leq r < 0$



二、直线相关（linear correlation）

（二）相关类型

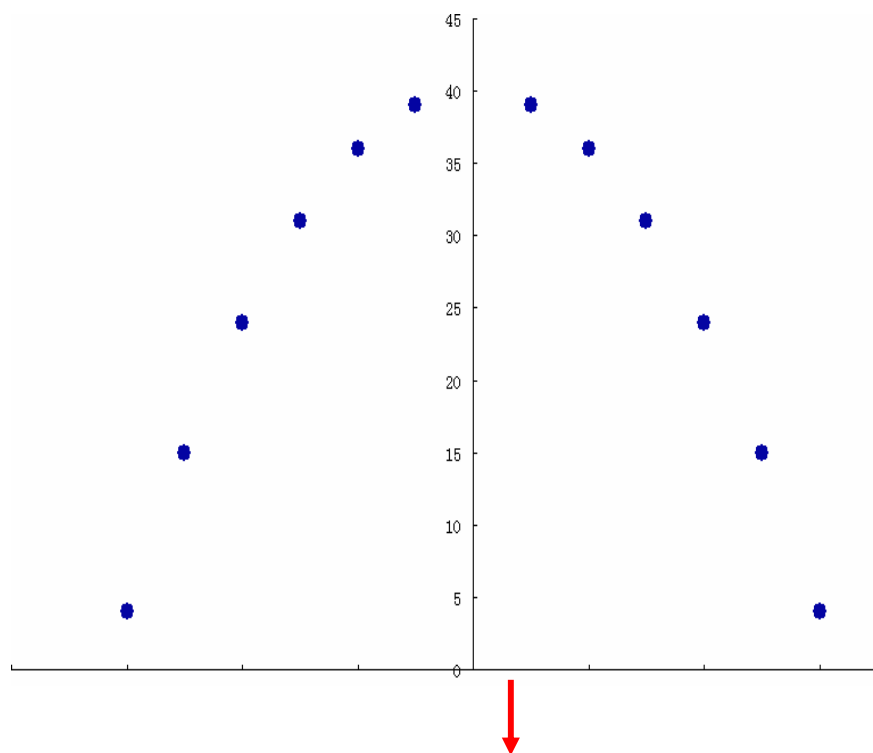
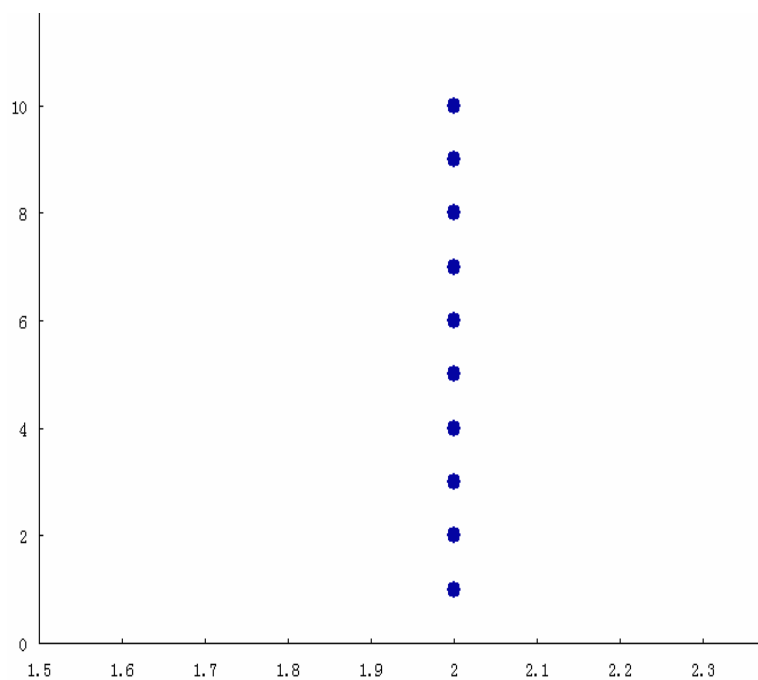
零相关 $r=0$



二、直线相关（linear correlation）

（二）相关类型

零相关 $r=0$



曲线相关

二、直线相关 (linear correlation)

(三) r 计算

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} = \frac{l_{XY}}{\sqrt{l_{XX} l_{YY}}}$$

$$\sum (X - \bar{X})(Y - \bar{Y}) = \sum XY - \frac{(\sum X \sum Y)}{n}$$

$$\sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$\sum (X - \bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n}$$

二、直线相关（linear correlation）

（三） r 的假设检验

r 为样本相关系数，由于抽样误差，实际工作中 r 一般都不为0。要判断两变量之间是否存在相关性，需要检验总体相关系数 ρ 是否为0。

$$H_0: \rho=0 \quad H_1: \rho \neq 0$$

$$t = \frac{r}{s_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

只有当 $\rho \neq 0$ 时，才能根据 $|r|$ 的大小判断相关的密切程度。

二、直线相关（linear correlation）

（四）相关与回归的区别和联系

1. 相关与回归的意义不同 相关表达两个变量之间相互关系的密切程度和方向。回归表达两个变量之间的数量关系，已知**X**值可以预测**Y**值。从散点图上，散点围绕回归直线的分布越密集，则两变量相关系数越大；回归直线的斜率越大，则回归系数越大。

2. r 与 b 的符号一致 同正同负。

根据公式：

$$r = \frac{l_{XY}}{\sqrt{l_{XX} l_{YY}}} \quad b = \frac{l_{XY}}{l_{XX}}$$

它们的符号取决于 l_{XY}

二、直线相关（linear correlation）

（四）相关与回归的区别和联系

3. r 与 b 的假设检验等价

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{b}{S_b}$$

意义：若 r 的假设检验拒绝 H_0 ，认为 $\rho \neq 0$ ，则 b 的假设检验也一定会拒绝 H_0 ，认为 $\beta \neq 0$ 。若 r 的假设检验接受 H_0 ，认为 $\rho = 0$ ，则 b 的假设检验也一定会接受 H_0 ，认为 $\beta = 0$ 。

二、直线相关（linear correlation）

（四）相关与回归的区别和联系

4. 可以用回归解释相关

$$r^2 = \frac{SS_{\text{回归}}}{SS_{\text{总}}}$$

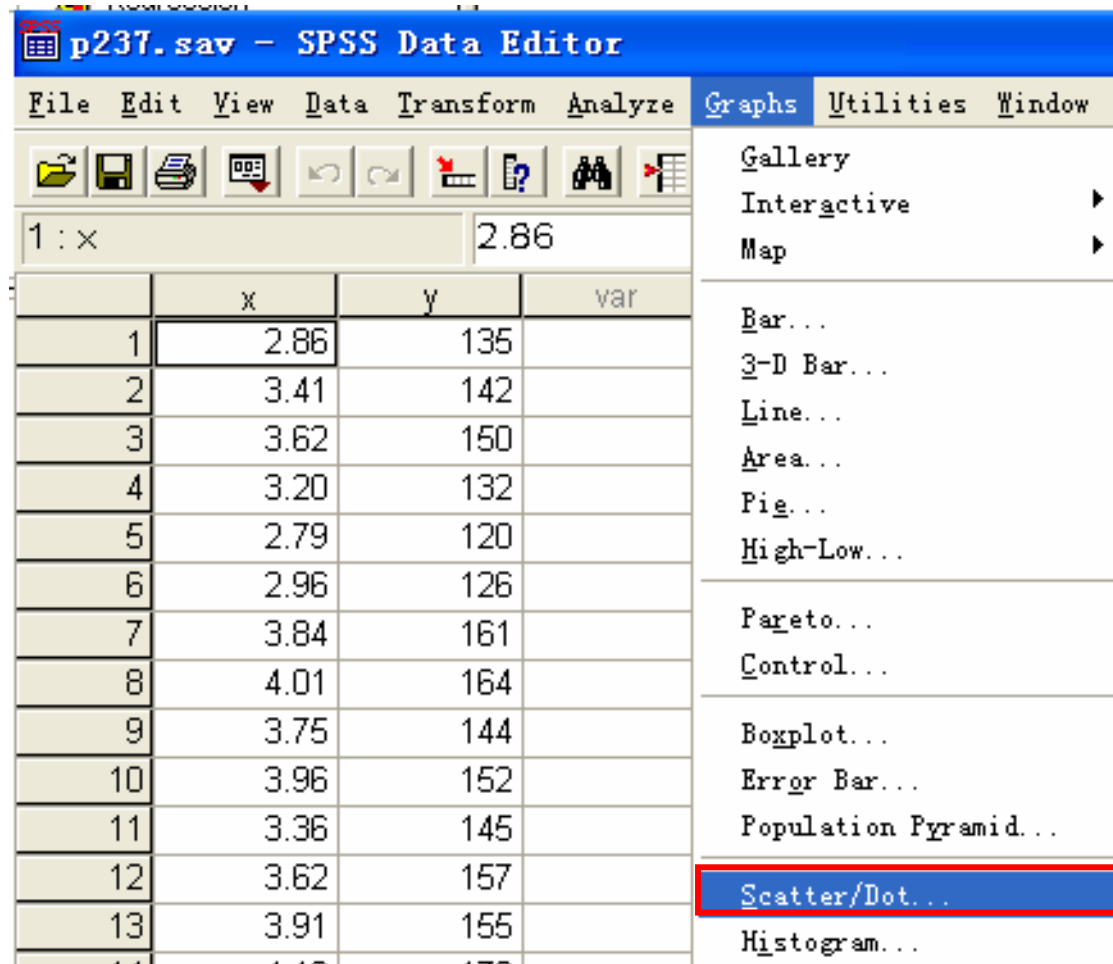
r^2 称为决定系数（**coefficient of determination**），反映了回归平方和占总平方和的比例，其越接近于**1**，回归直线拟和的效果越好。

反映回归直线拟和效果的两个指标： r^2 和 S_{YX} 。 r^2 越大， S_{YX} 越小，回归直线拟和效果越好。

三、SPSS13.0软件操作（直线回归与相关）

例1 见第237页例10-1。X: 体重指数；Y: 收缩压（mmHg）。

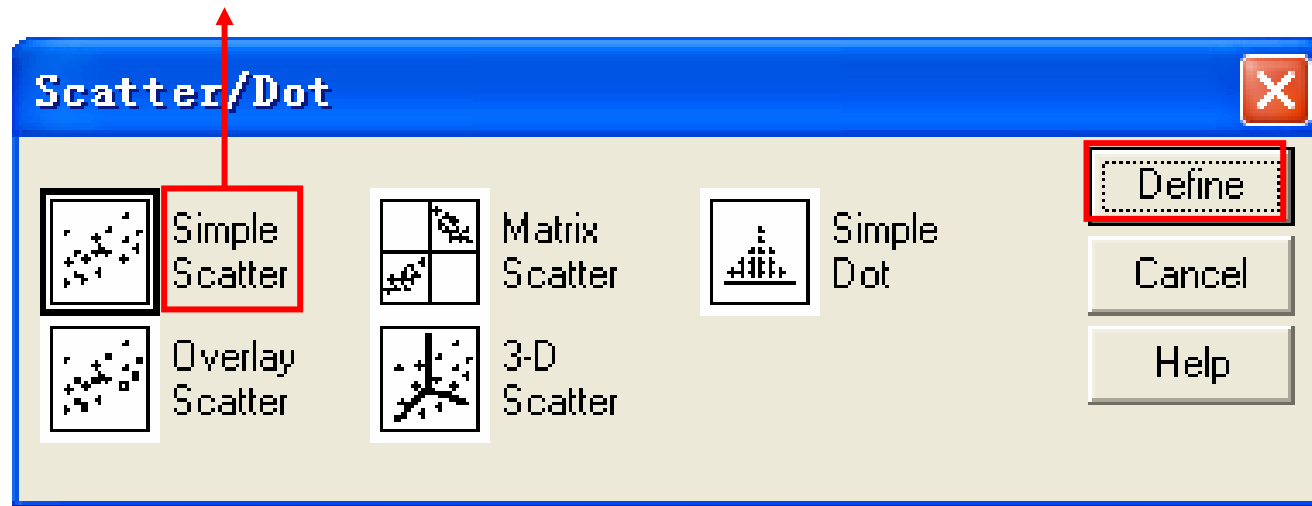
1. 绘制散点图



三、SPSS13.0软件操作 (直线回归与相关)

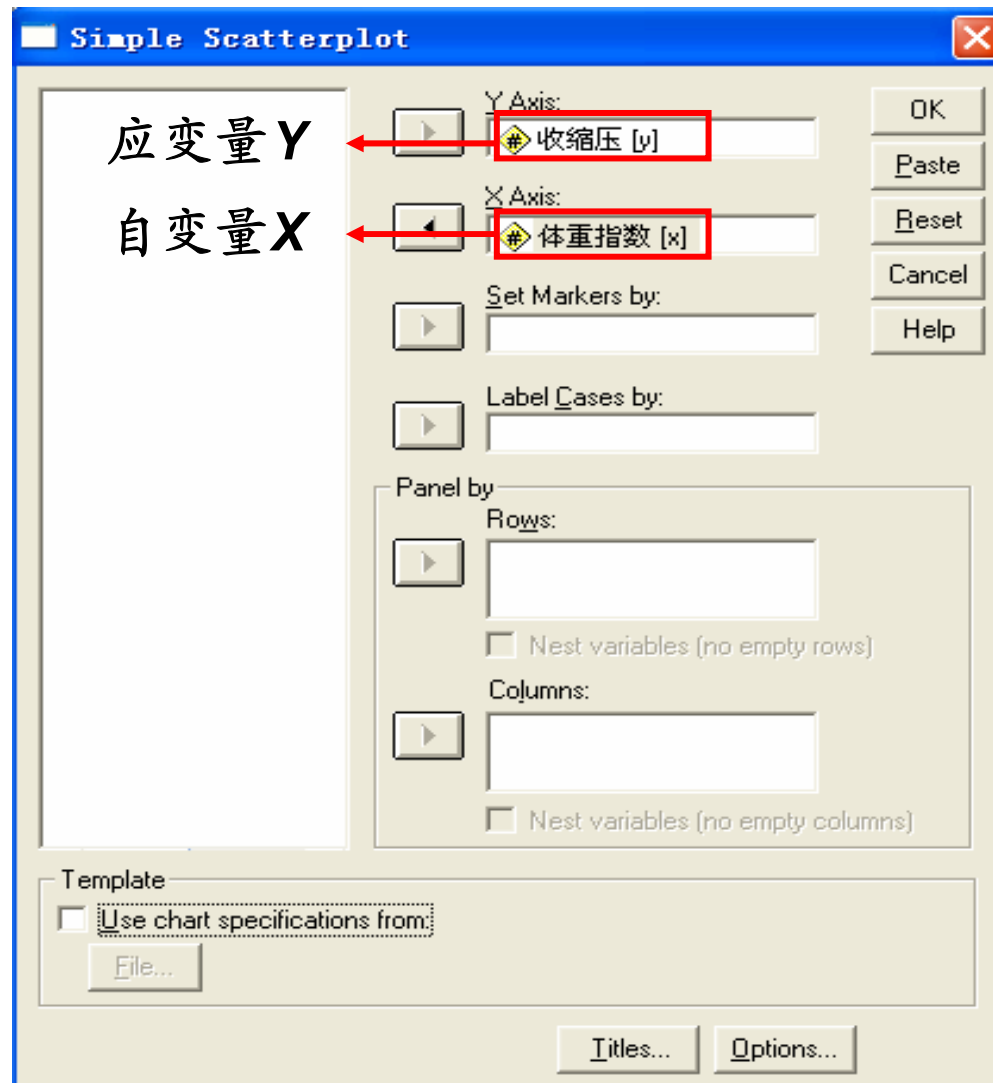
例1 1. 绘制散点图

简单散点图



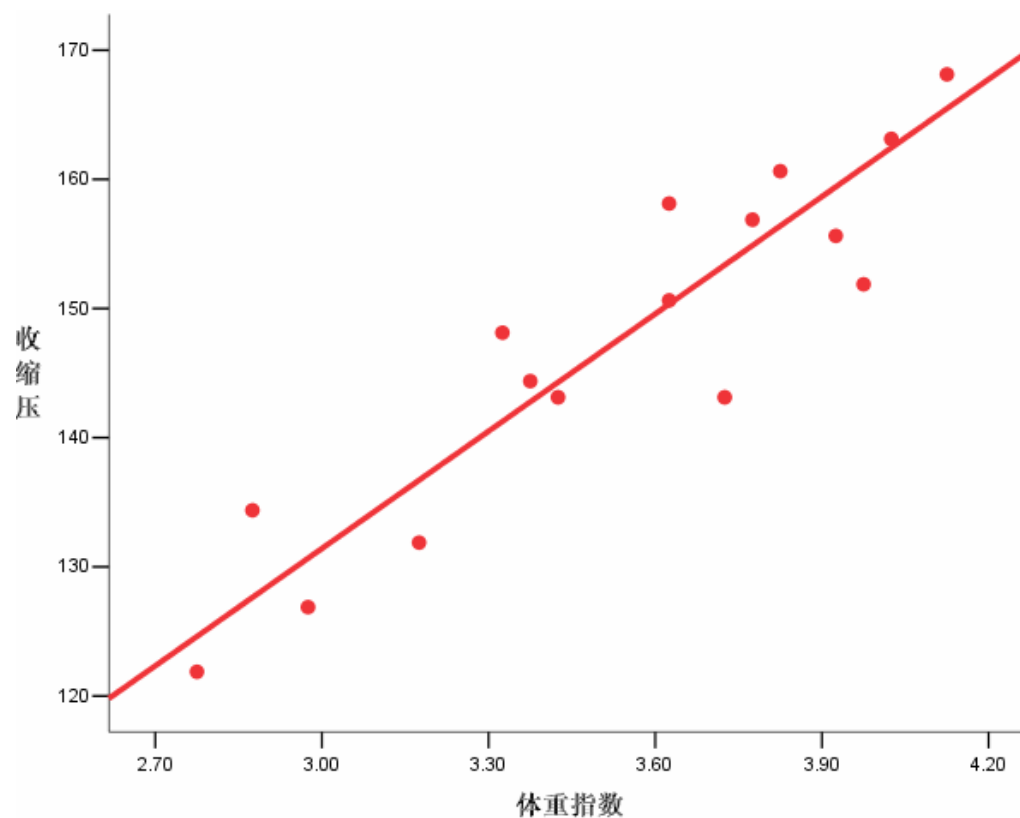
三、SPSS13.0软件操作 (直线回归与相关)

例1 1. 绘制散点图



三、SPSS13.0软件操作 (直线回归与相关)

例1 1. 绘制散点图



散点图显示：收缩压与体重指数之间有线性相关趋势，因此可以进一步做直线回归与相关

三、SPSS13.0软件操作 (直线回归与相关)

例1 2. 直线回归与相关分析

The screenshot shows the SPSS 13.0 Data Editor window with the file 'p237.sav'. The 'Analyze' menu is open, and the 'Regression' option is highlighted. A red box highlights the 'Regression' option in the 'Analyze' menu, and another red box highlights the 'Linear...' option in the 'Regression' submenu. Red arrows point from these boxes to the text 'Regression, 回归' and 'Linear, 线性' respectively.

File Edit View Data Transform **Analyze** Graphs Utilities Window Help

1 : x 2.86

	x	y
1	2.86	135
2	3.41	142
3	3.62	150
4	3.20	132
5	2.79	120
6	2.96	126

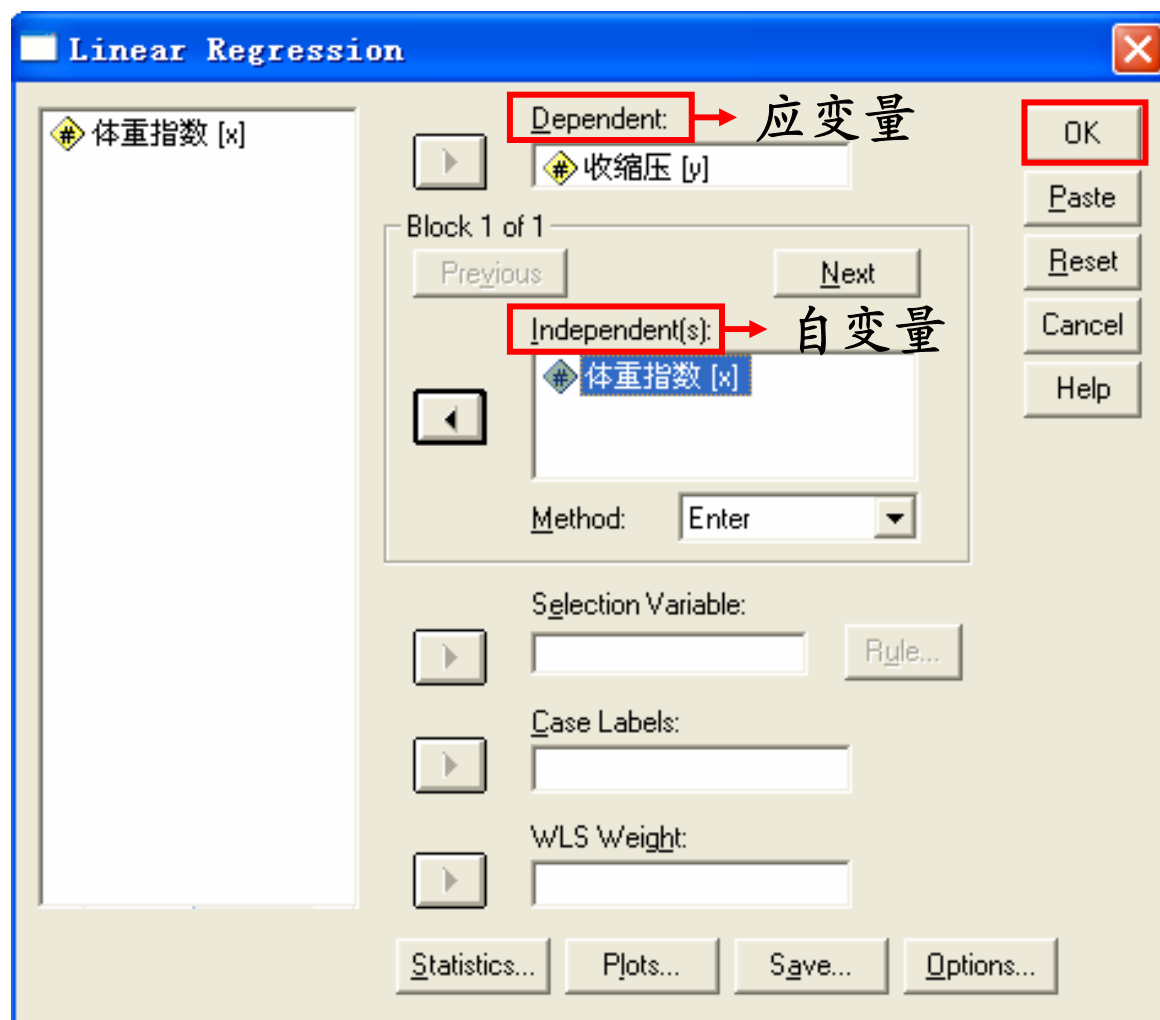
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Binary Logistic...

Linear...
Curve Estimation...
Binary Logistic...

Regression, 回归 **Linear, 线性**

三、SPSS13.0软件操作 (直线回归与相关)

例1 2. 直线回归与相关分析



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.911 ^a	.830	.818	5.947

a. Predictors: (Constant), 体重指数

相关
系数 r 决定
系数 r^2 调整 r^2 ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2414.920	1	2414.920	68.290	.000 ^a
	Residual	495.080	14	35.363		
	Total	2910.000	15			

SS
SS
SS
回归
剩余
总

a. Predictors: (Constant), 体重指数

b. Dependent Variable: 收缩压

自由度

 $MS_{\text{回归}}$ 及 $MS_{\text{剩余}}$ F 值 P 值Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	40.597	13.022		3.118	.008
	体重指数	30.274	3.663	.911	8.264	.000

a. Dependent Variable: 收缩压

截距 a 回归系数 b s_b

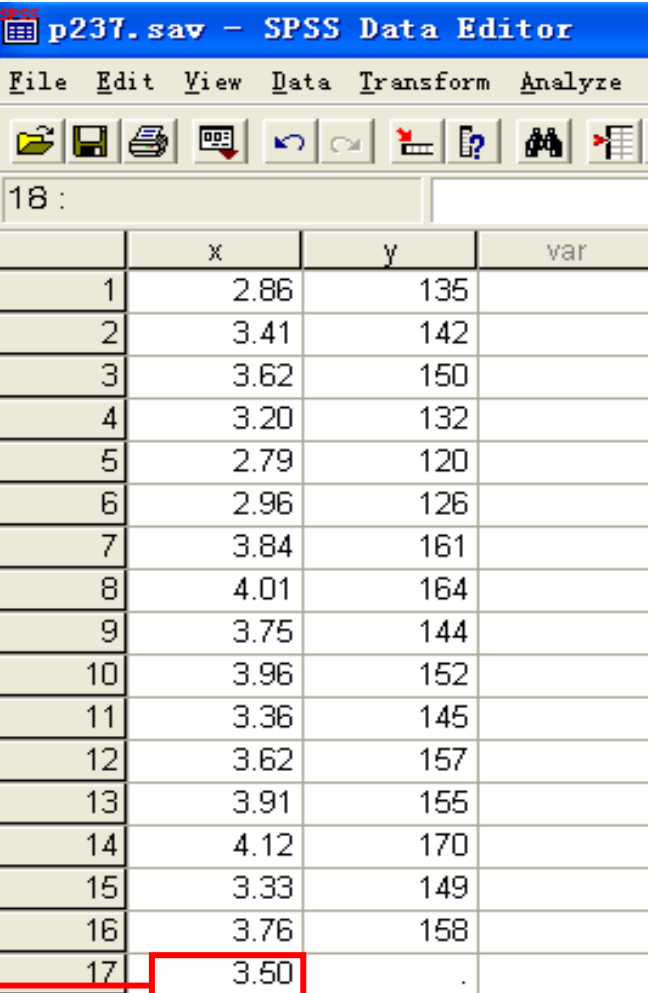
标准化回归系数

 t 值 P 值

$$F = t^2$$

三、SPSS13.0软件操作 (直线回归与相关)

例1 3. 直线回归的预测及置信区间估计

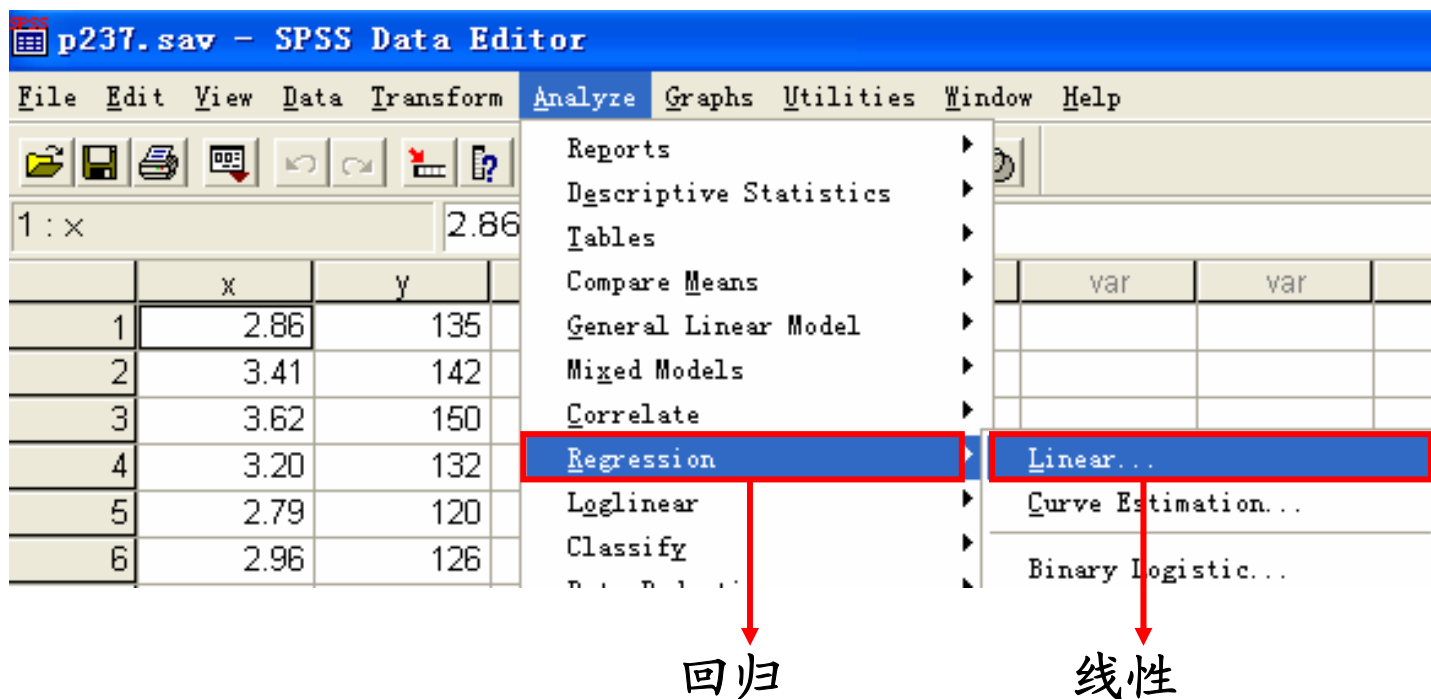


	x	y	var
1	2.86	135	
2	3.41	142	
3	3.62	150	
4	3.20	132	
5	2.79	120	
6	2.96	126	
7	3.84	161	
8	4.01	164	
9	3.75	144	
10	3.96	152	
11	3.36	145	
12	3.62	157	
13	3.91	155	
14	4.12	170	
15	3.33	149	
16	3.76	158	
17	3.50	.	

给定 $X = X_0$,
预测 Y

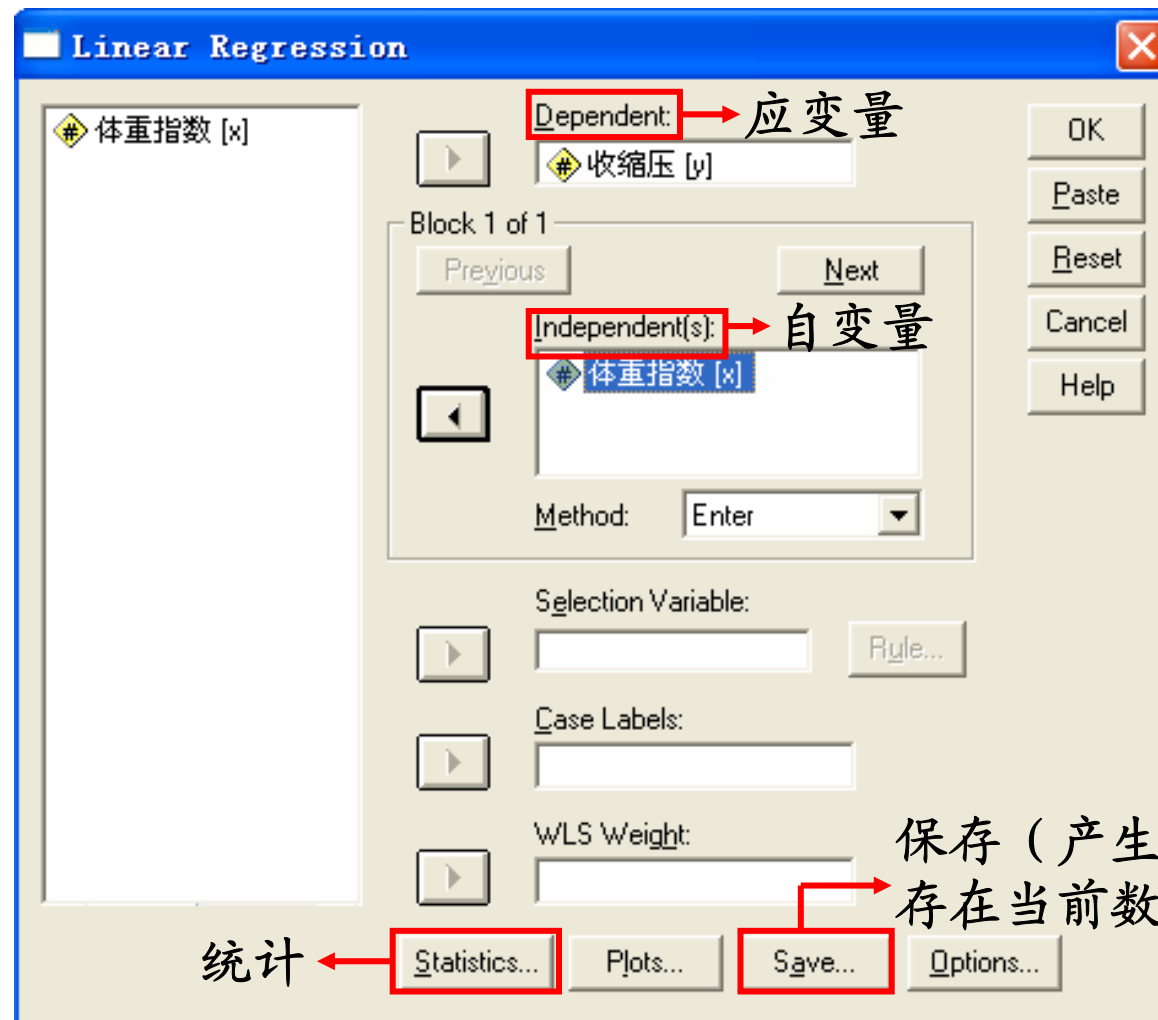
三、SPSS13.0软件操作 (直线回归与相关)

例1 3. 直线回归的预测及置信区间估计



三、SPSS13.0软件操作 (直线回归与相关)

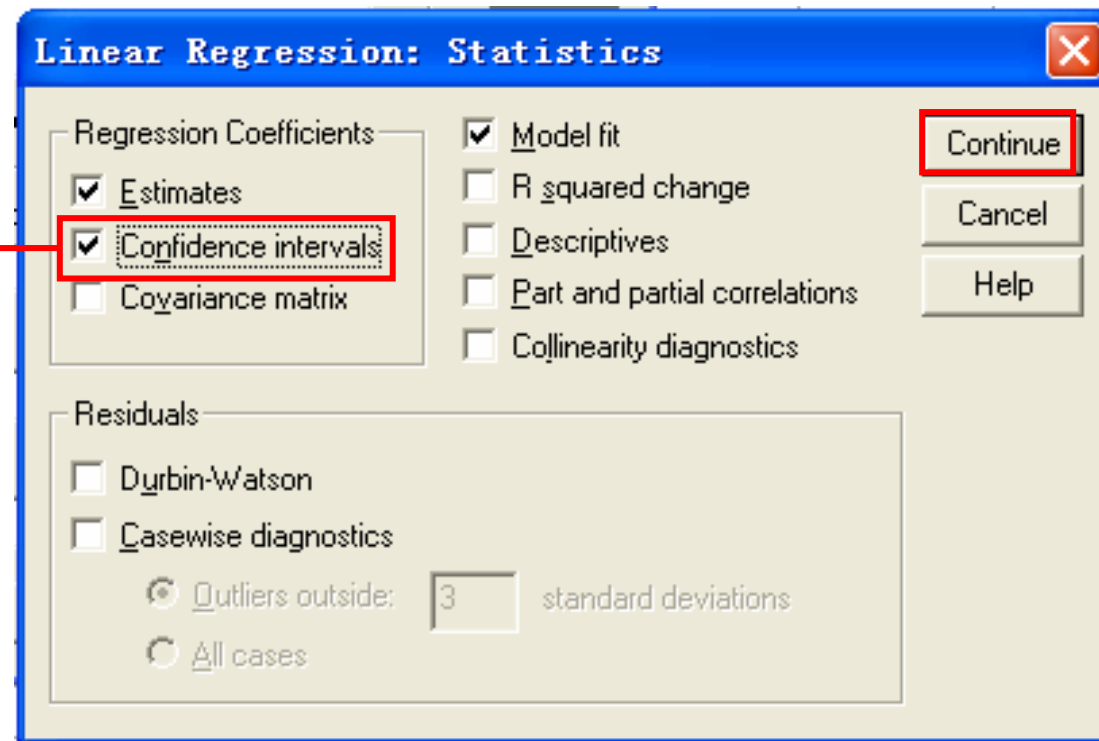
例1 3. 直线回归的预测及置信区间估计



三、SPSS13.0软件操作 (直线回归与相关)

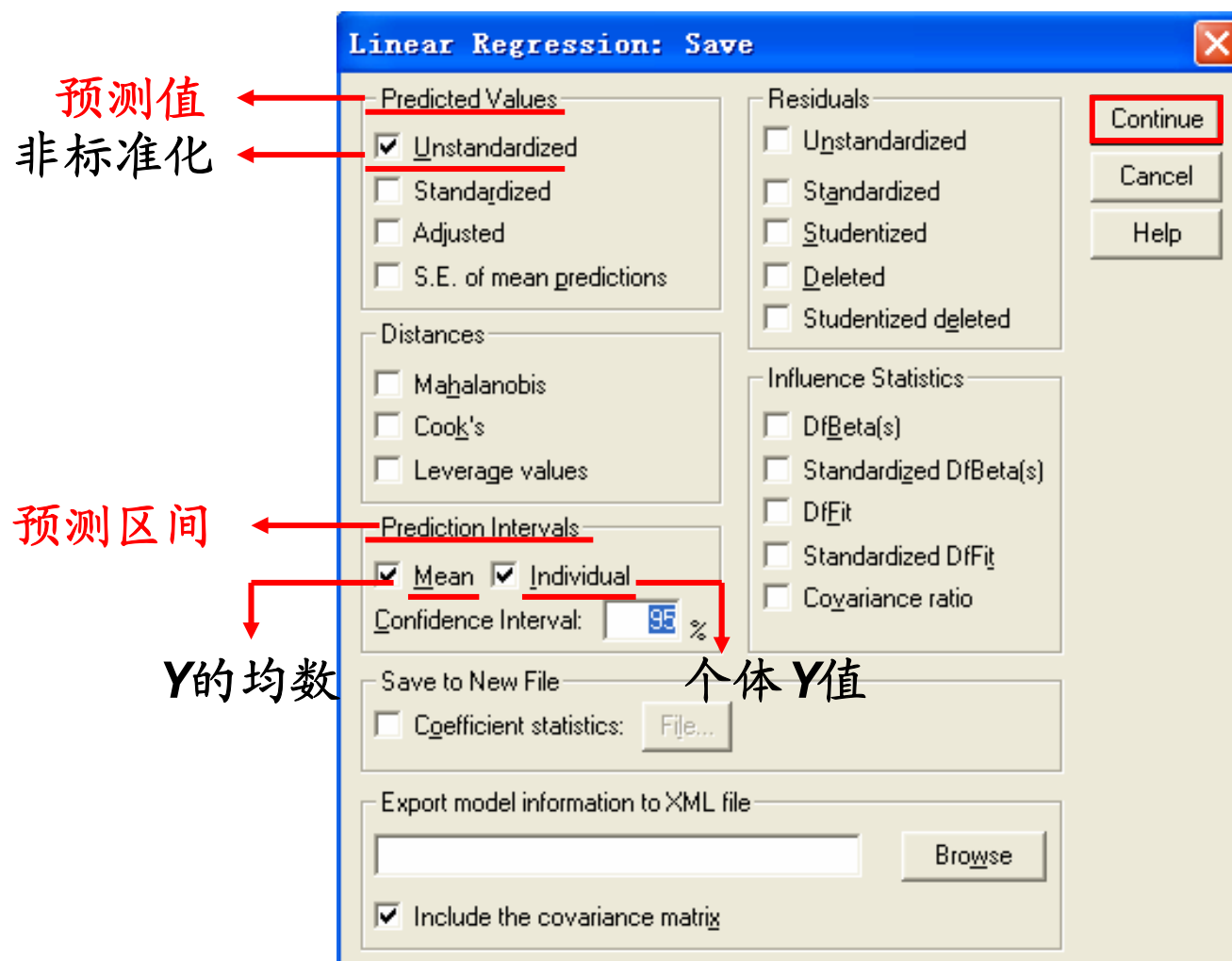
例1 3. 直线回归的预测及置信区间估计

总体回归系
数的置信区
间估计



三、SPSS13.0软件操作 (直线回归与相关)

例1 3. 直线回归的预测及置信区间估计



三、SPSS13.0软件操作 (直线回归与相关)

例1 3. 直线回归的预测及置信区间估计

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	40.597	13.022		3.118	.008	12.668	68.525
	体重指数	30.274	3.663	.911	8.264	.000	22.416	38.131

a. Dependent Variable: 收缩压

总体回归系数的
95% 置信区间

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
<u>Predicted Value</u>	125.06	165.32	147.50	12.688	16
Std. Predicted Value	-1.769	1.405	.000	1.000	16
Standard Error of Predicted Value	1.522	3.096	2.044	.509	16
Adjusted Predicted Value	124.80	164.20	147.52	12.643	16
<u>Residual</u>	-10.122	7.821	.000	5.745	16
<u>Std. Residual</u>	-1.702	1.315	.000	.966	16
Stud. Residual	-1.775	1.502	-.002	1.037	16
Deleted Residual	-11.010	10.204	-.023	6.644	16
Stud. Deleted Residual	-1.943	1.581	-.010	1.074	16
Mahal. Distance	.045	3.128	.938	.968	16
Cook's Distance	.000	.344	.081	.092	16
Centered Leverage Value	.003	.209	.063	.065	16

a. Dependent Variable: 收缩压

预测值

残差
残差标准差

三、SPSS13.0软件操作 (直线回归与相关)

例1 3. 直线回归的预测及置信区间估计

SPSS Data Editor window: p237.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

21:

	x	y	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
1	2.86	135	127.17889	121.01578	133.34200	113.01354	141.34424
2	3.41	142	143.82933	140.50147	147.15720	130.64799	157.01068
3	3.62	150	150.18678	146.92283	153.45072	137.02142	163.35213
4	3.20	132	137.47189	133.35593	141.58786	124.06987	150.87392
5	2.79	120	125.05974	118.41986	131.69963	110.68055	139.43894
6	2.96	126	130.20624	124.70050	135.71199	116.31429	144.09820
7	3.84	161	156.84695	152.84044	160.85347	143.47813	170.21577
8	4.01	164	161.99345	157.06222	166.92469	148.31902	175.66789
9	3.75	144	154.12234	150.50001	157.74466	140.86359	167.38108
10	3.96	152	160.47978	155.84126	165.11829	146.90815	174.05140
11	3.36	145	142.31566	138.85479	145.77652	129.10011	155.53121
12	3.62	157	150.18678	146.92283	153.45072	137.02142	163.35213
13	3.91	155	158.96610	154.60454	163.32766	145.48662	172.44558
14	4.12	170	165.32354	159.70514	170.94194	151.38656	179.26053
15	3.33	149	141.40745	137.84831	144.96659	128.16583	154.64908
16	3.76	158	154.42507	150.76481	158.08533	141.15591	167.69423
17	3.50	.	146.55395	143.35593	149.75198	133.40479	159.70312

Annotations for row 17:

- x_0 points to the value 3.50 in the x column.
- Y的预测值 points to the value 146.55395 in the PRE_1 column.
- Y的均数的置信区间的下限及上限 points to the values 143.35593 (LMCI_1) and 149.75198 (UMCI_1).
- 个体Y值的容许区间的下限及上限 points to the values 133.40479 (LICI_1) and 159.70312 (UICI_1).

四、附录：SPSS13.0软件操作 (Spearman等级相关)

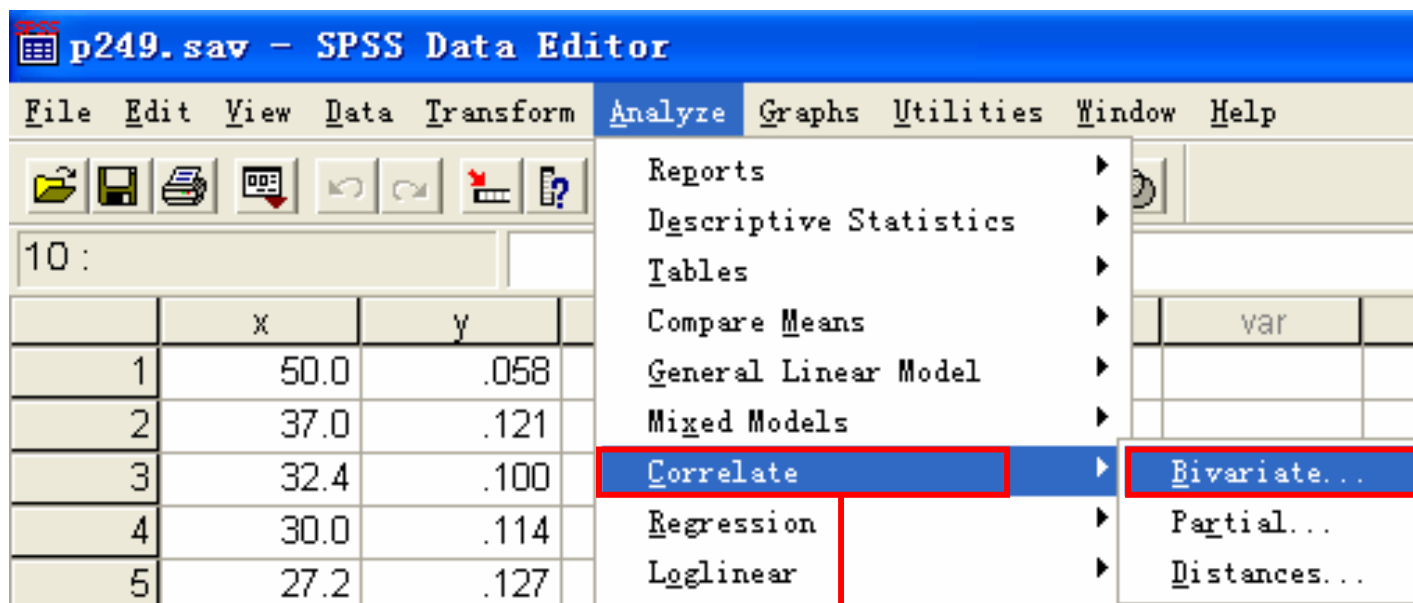
Spearman等级相关是基于秩次的非参数相关分析。

主要适用于以下情况：

1. 对于数值型变量，**X**及**Y**严重偏离正态分布；
2. 等级资料的相关分析。

四、附录：SPSS13.0软件操作 (Spearman等级相关)

例2 见第249页例10-11。X: 大骨节病阳性率；Y: 发硒。

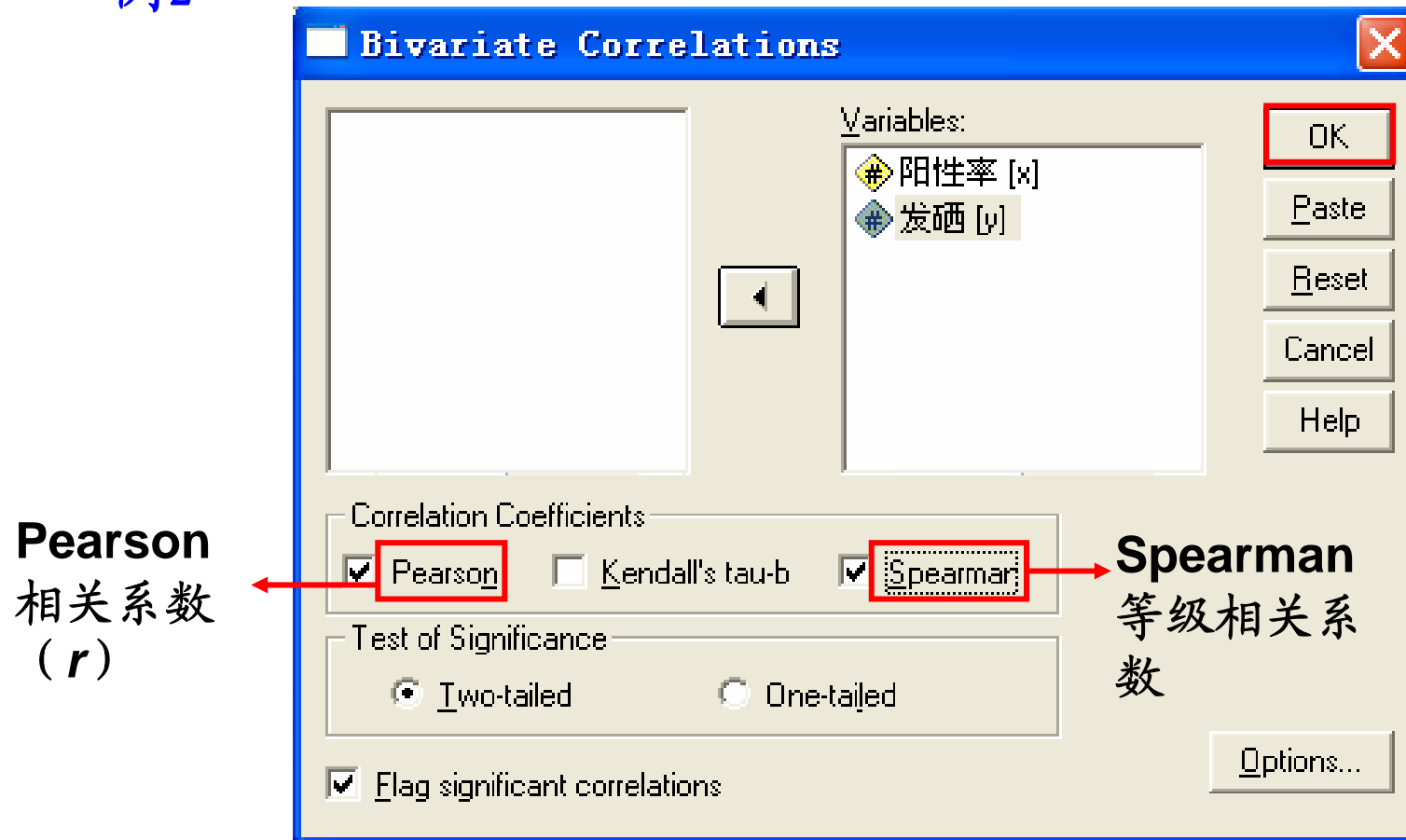


相关

两变量

四、附录：SPSS13.0软件操作 (Spearman等级相关)

例2



四、附录：SPSS13.0软件操作 (Spearman等级相关)

例2

Correlations

Correlations

		阳性率	发硒
阳性率	Pearson Correlation	1	-.912**
	Sig. (2-tailed)		.001
	N	9	9
发硒	Pearson Correlation	-.912**	1
	Sig. (2-tailed)	.001	
	N	9	9

** . Correlation is significant at the 0.01 level

Pearson相关系数 (r)

P 值

Nonparametric Correlations

Correlations

			阳性率	发硒
Spearman's rho	阳性率	Correlation Coefficient	1.000	-.917**
		Sig. (2-tailed)	.	.001
		N	9	9
Spearman's rho	发硒	Correlation Coefficient	-.917**	1.000
		Sig. (2-tailed)	.001	.
		N	9	9

** . Correlation is significant at the 0.01 level (2-tailed).

Spearman相关系数

P 值