线性相关和回归

赵耐青

在实际研究中，经常要考察两个指标之间的关系，即：相关性。现以体重与身高的关系为例，分析两个变量之间的相关性。要求身高和体重呈双正态分布，既：在身高和体重平均数的附近的频数较多，远离身高和体重平均数的频数较少。

样本相关系数计算公式(称为 Pearson 相关系数)：

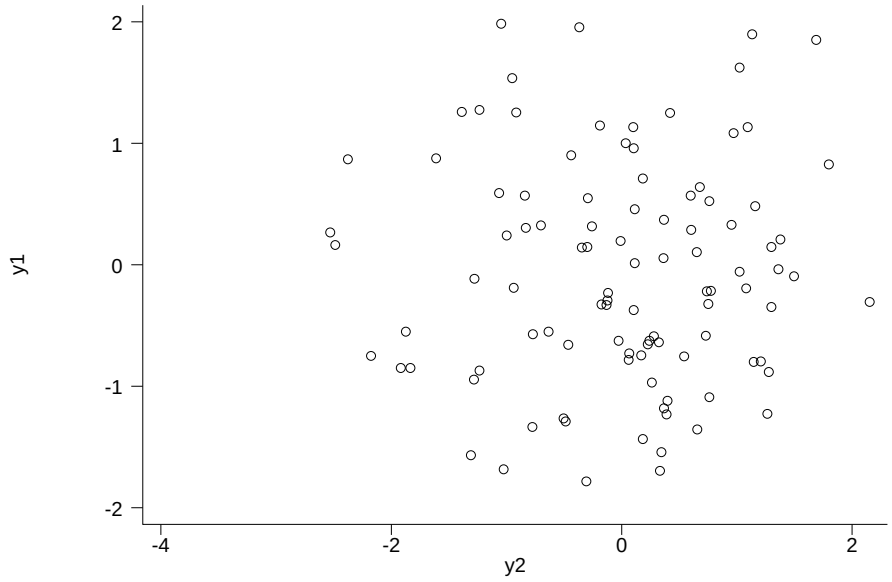
$$r=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}}=\frac{L_{XY}}{\sqrt{L_{XX}}\sqrt{L_{YY}}} \qquad (1)$$

1. 考察随机模拟相关的情况。

显示两个变量相关的散点图程序 simur.ado（本教材配套程序,使用见前言）。命令为 simur　样本量　总体相关系数

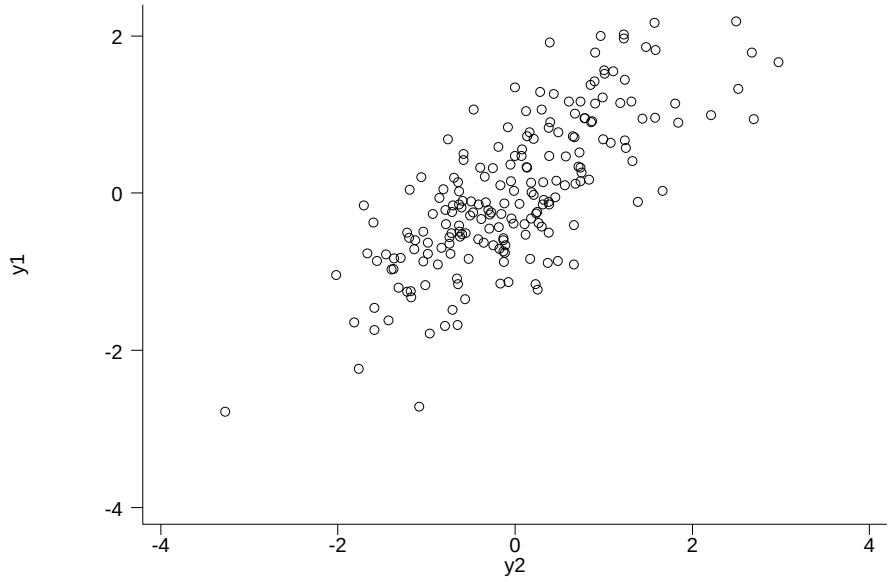
如显示样本量为 100，ρ=0 的散点图

本例命令为 **simur　100　0**



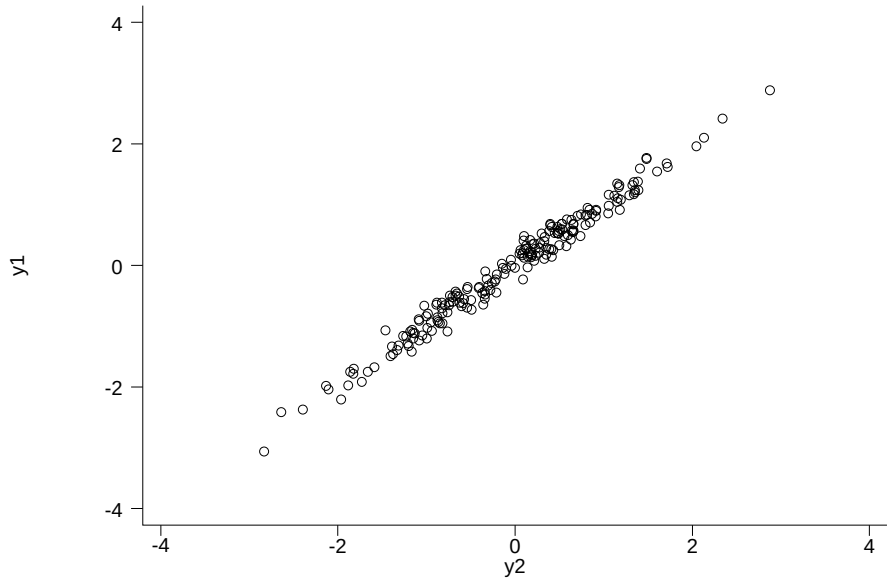
如显示样本量为 200，ρ=0.8 的散点图

本例命令为 **simur　200　0.8**



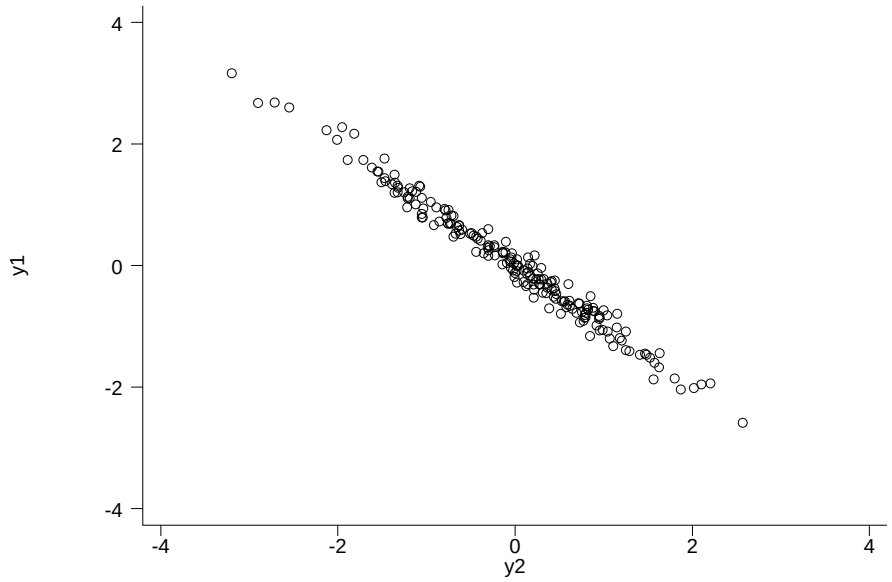
如显示样本量为 200，ρ=0.99 的散点图

本例命令为 **simur　200　0.99**



如显示样本量为 200，ρ=-0.99 的散点图

本例命令为 **simur　200　-0.99**



例 1.　测得某地 15 名正常成年男子的身高 *x*（cm）、体重 *y*（kg）如

试计算 *x* 和 *y* 之间的相关系数 *r* 并检验 H_0：ρ＝0 vs　H_1: ρ≠0。

α=0.05

数据格式为

X	Y
171.0	58.0
176.0	69.0
175.0	74.0
172.0	68.0
170.0	64.0
173.0	68.5
168.0	56.0
172.0	54.0
170.0	62.0
172.0	63.0
173.0	67.0
168.0	60.0
171.0	68.0
172.0	76.0
173.0	65.0

Stata 命令　**pwcorr　变量 1　变量 2 …　变量 m，sig**

本例命令　pwcorr　x　y,sig

pwcorr x y,sig

```
             |        x        y
-------------+------------------
           x |   1.0000
             |
             |
           y |   0.5994   1.0000
             |   0.0182
             |
```

Pearson 相关系数=0.5994，P 值=0.0182<0.05，因此可以认为身高与体重呈正线性相关。

注意：Pearson 相关系数又称为线性相关系数并且要求 X 和 Y 双正态分布，通常在检查中要求 X 服从正态分布并且 Y 服从正态分布。

如果不满足双正态分布时，可以计算 Spearman 相关系数又称为非参数相关系数。

Spearman 相关系数的计算基本思想为：用 X 和 Y 的秩代替它们的原始数据，然后代入 Pearson 相关系数的计算公式并且检验与 Pearson 相关系数类同。

Stata 实现

spearman x y

```
Number of obs =        15
Spearman's rho =       0.6552
Test of Ho: x and y are independent
     Prob > |t| =      0.0080
```

stata 计算结果与手算的结果一致。结论为身高与体重呈正相关，并且有统计学意义。

直线回归

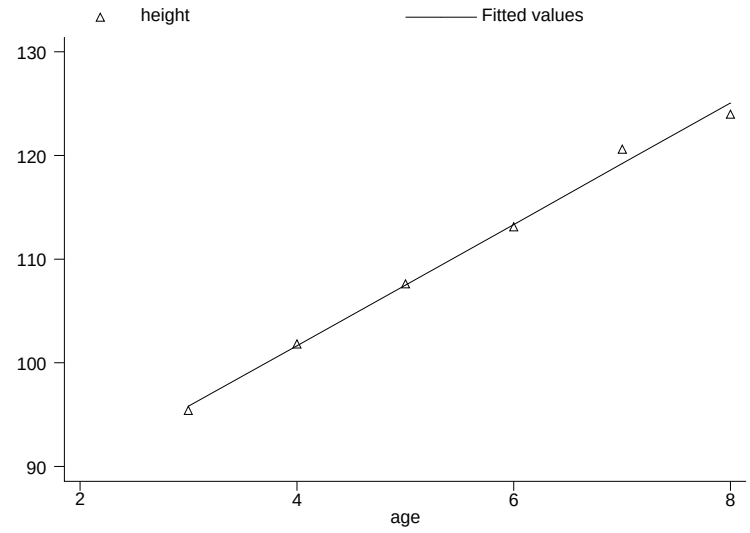
例 2 为了研究 3 岁至 8 岁男孩身高与年龄的规律，在某地区在 3 岁至 8 岁男孩中随机抽样，共分 6 个年龄层抽样：3 岁，4 岁，…，8 岁，每个层抽 10 个男孩，共抽 60 个男孩。资料如下：

60 个男孩的身高资料如下

年龄	3 岁	4 岁	5 岁	6 岁	7 岁	8 岁
身	92.5	96.5	106.0	115.5	125.5	121.5
高	97.0	101.0	104.0	115.5	117.5	128.5
	96.0	105.5	107.0	111.5	118.0	124.0

	96.5	102.0	109.5	110.0	117.0	125.5
	97.0	105.0	111.0	114.5	122.0	122.5
	92.0	99.5	107.5	112.5	119.0	123.5
	96.5	102.0	107.0	116.5	119.0	120.5
	91.0	100.0	111.5	110.0	125.5	123.0
	96.0	106.5	103.0	114.5	120.5	124.0
	99.0	100.0	109.0	110.0	122.0	126.5
平均身高	95.4	101.8	107.6	113.1	120.6	124.0

由于男孩的身高与年龄有关系，不同的年龄组的平均身高是不同的，由平均身高与年龄作图可以发现：年龄与平均身高的点在一条直线附近。



考虑到样本均数存在抽样误差，故有理由认为身高的总体均数与年龄的关系可能是一条直线关系 $\mu_y = \alpha + \beta x$，其中 y 表示身高，x 表示年龄。由于身高的总体均数与年龄有关，所以更正确地标记应为

$$\mu_{y|x} = \alpha + \beta x$$

表示在固定年龄情况下的身高总体均数。

上述公式称为直线回归方程。其中β为回归系数（regression coefficient），或称为斜率（slope）；α称为常数项（constant），或称为

截距（intercept）。回归系数β表示 x 变化一个单位 y 平均变化β个单位。当 x 和 y 都是随机的，x、y 间呈正相关时β>0，x、y 间呈负相关时β<0，x、y 间独立时β=0。

一般情况而言，参数α和β是未知的。对于本例而言，不同民族和不同地区，α和β往往是不同的，因此需要进行估计的。由于不同年龄的身高实际观察值应在对应的身高总体均数附近(即：实际观察值与总体均数之间仅存在个体变异的差异)，故可以用年龄和实际身高观察值的资料对未知参数α和β进行估计。得到样本估计的回归方程

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

二、直线回归方程的建立

直线回归分析的 Stata 实现：

数据结构：

x	y
3	92.5
3	97
3	96
3	96.5
3	97
3	92
3	96.5
3	91
3	96
3	99
4	96.5
4	101
4	105.5
4	102
4	105
4	99.5
4	102
4	100

4	106.5
4	100
5	106
5	104
5	107
5	109.5
5	111
5	107.5
5	107
5	111.5
5	103
5	109
6	115.5
6	115.5
6	111.5
6	110
6	114.5
6	112.5
6	116.5
6	110
6	114.5
6	110
7	125.5
7	117.5
7	118
7	117
7	122
7	119
7	119
7	125.5
7	120.5
7	122
8	121.5
8	128.5
8	124
8	125.5
8	122.5
8	123.5
8	120.5
8	123
8	124
8	126.5

多重线性回归命令为

regress　因变量　自变量 1　自变量 2 ……自变量 m

直线回归命令 **regress　因变量　自变量**

本例为 **regress　y　x**，得到下列结果：

```
      Source |       SS       df       MS              Number of obs =      60
-------------+------------------------------           F(  1,    58) =  777.41
       Model |  5997.71571     1  5997.71571           Prob > F      =  0.0000
    Residual |  447.467619    58  7.71495895           R-squared     =  0.9306
-------------+------------------------------           Adj R-squared =  0.9294
       Total |  6445.18333    59  109.240395           Root MSE      =  2.7776

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    5.854286   .2099654    27.88   0.000     5.433994    6.274577
       _cons |    78.18476   1.209202    64.66   0.000     75.76428    80.60524
------------------------------------------------------------------------------
```

得到回归系数 b=5.854286，常数项 a=78.18746，回归系数的检验统计量 t_b=27.88，P 值<0.0001，可以认为 Y 与 X 呈直线回归关系。

来源	平方和 SS	自由度 df	均方 MS	F	P 值
回归	5997.71571	1	5997.71571	777.41	<0.0001
残差	447.467619	58	7.71495895		
合计	6445.18333	59			

称 $R^2=1-\frac{SS_{残差}}{SS_{合计}}$ 为决定系数(本例 Stata 计算结果 R-squared=0.9306)，因此 $0\le R^2\le 1$，因此残差平方和 SSE 越小，决定系数 R^2 就越接近 1。特别当所有的残差为 0 时，SSE=0，相应的决定系数 $R^2=1$。决定系数 R^2 表示 y 被 x 所解释的部分所占的百分比，R^2 越接近于 1 说明 x 对 y 的解释越充分。

残差=应变量观察值（y）-预测值($\hat{y}$)

Stata 的残差计算命令

在输入回归命令 regress　y　x 后，再

输入 **predict　e,residual** 计算残差并用变量 e 表示残差

输入 **sktest　e**　　残差的正态性检验

输入 **predict　yy**　　计算预测值。

残差正态性检验(H_0:残差正态分布,α=0.05)

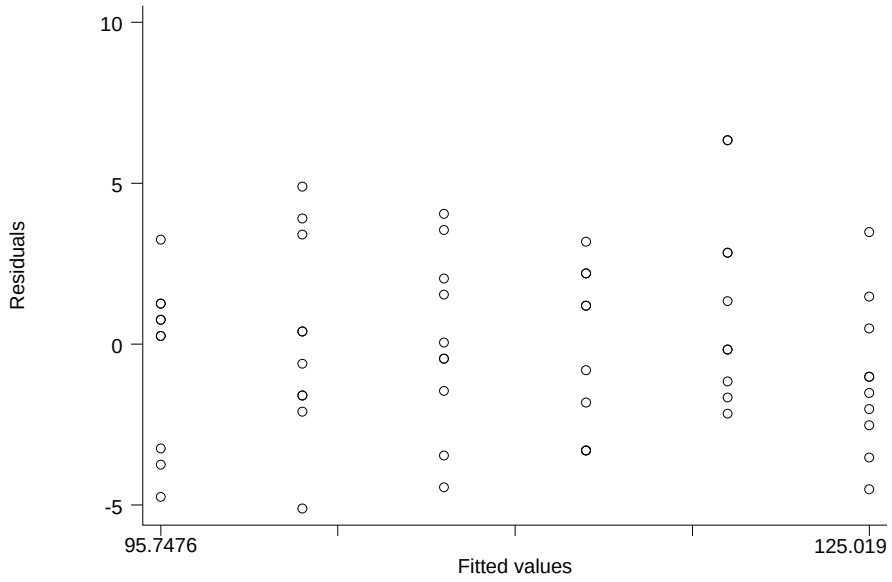
```
sktest e

                    Skewness/Kurtosis tests for Normality
                                                    ------- joint ------
    Variable | Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)     Prob>chi2
-------------+-------------------------------------------------------
           e |       0.459          0.441          1.18        0.5534
```

P 值=0.5534>>0.05，可以认为残差呈正态分布。

所建立的回归方程是否有意义，仅凭借假设检验的结论或 R^2 的大小还不能充分说明问题。残差 $\boldsymbol{e = Y - \hat{Y}}$ 的大小直接反应回归方程的优劣，经常采用图示的方法，以 e 做纵轴，$\hat{\boldsymbol{Y}}$ 为横轴作图来考察残差的变化，如果残差比较均匀地散布在 e=0 的周围，没有明显的散布趋势和明显的离群点，则说明所建回归方程比较理想，否则要借助统计软件做进一步诊断。

graph　残差　预测值

本例　graph　e　yy



说明残差比较均匀地散布在 e=0 的周围，没有明显的散布趋势和明显的离群点，故说明所建回归方程比较理想。