如何正确运用*t*检验——线性回归模型中参数与0比较*t*检验及SAS实现

黄慧杰[1]，刘媛媛[1*]，李长平[1,2]，胡良平[2,3]

（1. 天津医科大学公共卫生学院，天津 300070；
2. 世界中医药学会联合会临床科研统计学专业委员会，北京 100029；
3. 军事科学院研究生院，北京 100850
*通信作者：刘媛媛，E-mail：ivyuan10@126.com）

【摘要】 本文目的是介绍*t*检验在简单线性回归分析和多重线性回归分析中参数与0之间差异性比较时的理论依据和应用实践。首先介绍线性回归分析中*t*检验的基本原理和计算公式；然后用SAS程序分别对2个实例进行简单线性和多重线性回归分析，提请读者着重关注输出结果中有关“参数假设检验部分的检验统计量的名称及结果”；最后对结果进行解释和讨论。

【关键词】 简单线性回归分析；多重线性回归分析；参数；检验统计量；*t*检验
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20200717004

How to use *t* test correctly——comparison of parameter and 0 *t* test in the linear regression model and SAS implementation

Huang Huijie[1], *Liu Yuanyuan*[1*], *Li Changping*[1,2], *Hu Liangping*[2,3]

(1. *School of Public Health*, *Tianjin Medical University*, *Tianjin* 300070, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies*, *Beijing* 100029, *China*;
3. *Graduate School*, *Academy of Military Sciences PLA China*, *Beijing* 100850, *China*
**Corresponding author*: *Liu Yuanyuan*, *E-mail*: *ivyuan*10@126.*com*)

【**Abstract**】 The purpose of this paper was to introduce the theoretical basis and application practice of *t* test when comparing the difference between parameters and 0 in simple linear regression analysis and multiple linear regression analysis. Firstly, this paper introduced the basic principle and calculation formula of *t* test in linear regression analysis. Then the simple linear regression analysis and multiple linear regression analysis of two examples were carried out by SAS program. In the output results, readers should draw attention to the name and results of test statistics of the parametric hypothesis testing. Finally, the results was explained and discussed.

【**Keywords**】 Simple linear regression; Multiple linear regression; Parameter; Test statistics; *t* test

回归分析是医学研究中常用的分析方法，其中，线性回归分析是最常用、最简单的一种。线性回归分析是通过建立因变量随单个或多个自变量按线性关系变化的方程式并检验整个方程式和参数是否具有统计学意义。严格地说，线性回归分析的自变量应该是定量的，但在实际应用中，自变量的范围被拓展了，包括二分类变量、进行哑变量变换后的多分类变量及多值有序变量，而因变量必须是定量变量。含单个因变量和单个自变量的线性回归模型被称为简单线性回归模型；含单个因变量和多个自变量的线性回归模型被称为多重线性回归模型[1]。在线性回归分析中，需要对整个回归模型和模型中的各参数进行假设检验，对回归模型整体检验采用的是方差分析；对参数检验采用的是*t*检验。本文着重探讨*t*检验对简单线性回归模型和多重线性回归模型的参数与0之间差异进行假设检验的原理与应用。

基金项目：国家自然科学基金项目（项目名称：贝叶斯生存分析方法在肝细胞癌肝移植患者预后预测中的应用研究，项目编号：81803333）

1 基本概念

1.1 参数和统计量

参数是用来描述总体特征的概括性数字度量，是研究者想要了解的总体的某些特征值。依据经典统计学的观点，由于总体数据通常是未知的，所以参数是一个未知的常数[2]。设$x_1, x_2, \cdots, x_n$为取自某总体的样本，若样本函数$T(x_1, x_2, \cdots, x_n)$中不含有任何未知参数，则称T为样本统计量[3]。由于样本

是已经抽取出来的，所以样本统计量是已知的，抽样的目的就是用样本统计量去估计总体参数。也就是说，一旦选定了一个参数，就必然有一个统计量与之对应。

常用一元一重线性回归模型为：

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

上式中，β_0是直线在y轴上的截距，β_1是直线的斜率，而ε是y轴方向上的随机误差。一般假定：ε服从均值为0、方差为σ^2的正态分布。由于β_0和β_1是未知参数，一般采用最小二乘法对β_0和β_1进行估计。β_0的估计值为$\hat{\beta}_0$，β_1的估计值为$\hat{\beta}_1$：

$$\begin{cases}\hat{\beta}_1 = S_{xy}/S_{xx} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\end{cases} \tag{2}$$

上式中，$S_{xy} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$，$S_{xx} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)$，$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$，$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

β_0和β_1都叫做参数，而$\hat{\beta}_0$和$\hat{\beta}_1$就是两个参数的估计值，都叫做样本统计量。在多重线性回归分析中也具有类似的参数和样本统计量，只是被称为“斜率”的参数和样本统计量的数目≥2。

具有k个自变量的多重线性回归模型可表示如下：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{3}$$

上式中，β_0，β_1，β_2，…，β_k均是多重线性回归模型的参数，ε是y轴方向上的随机误差。一般假定：ε服从均值为0、方差为σ^2的正态分布。同样也采用最小二乘法来估计多重线性回归模型中的各个参数。对于n组数据，可得到以下由多重线性回归模型中参数的估计值组成的向量$\hat{\beta}$[4]：

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{4}$$

上式中，X^T是X（被称为设计矩阵）的转置矩阵，$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)^T$，$Y = (y_1, y_2, \cdots, y_n)^T$都是列向量。

$$\mathrm{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \tag{5}$$

1.2 简单线性回归模型中截距与0之间差异的t检验

在简单线性回归分析中对截距进行假设检验的原假设和备择假设分别为：

H_0:β_0=0 vs H_1:β_0≠0

在线性回归模型中，$\hat{\beta}_0$为正态变量y_1，y_2，…，y_n的线性组合，且y_1，y_2，…，y_n这些随机变量是独立同分布的，所以$\hat{\beta}_0$也服从正态分布。$\hat{\beta}_0$服从的分布为[3]：

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}}\right) \tag{6}$$

通过转换为标准正态分布可得：

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2}} \sim N(0,1) \tag{7}$$

由于残差平方和$SSE = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$，均方误差$MSE=SSE/(n-2)$[2]。而$SSE/\sigma^2$服从自由度为$n-2$的$\chi^2$分布[4]，即：

$$\frac{SSE}{\sigma^2} = \frac{MSE(n-2)}{\sigma^2} \sim \chi^2(n-2) \tag{8}$$

根据以上公式可得如下检验统计量：

$$t_0 = \frac{\dfrac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2}}}{\sqrt{\dfrac{MSE(n-2)}{\sigma^2(n-2)}}} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$$

$$= \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} \sim t(n-2) \tag{9}$$

上式中，$se(\hat{\beta}_0)$是$\hat{\beta}_0$的标准误差。

当原假设为真，即β_0=0时，t_0=$\hat{\beta}_0/se(\hat{\beta}_0)$服从自由度为$n-2$的$t$分布。当$\left|t_0\right| > t_{0.025}(n-2)$或$t_0$对应的$P$值小于0.05时，可认为$\beta_0$与0之间差异有统计学意义。

1.3 简单线性回归模型中斜率与0之间差异的t检验

在简单线性回归分析中对斜率进行假设检验的原假设和备择假设分别为：

H_0:β_1=0 vs H_1:β_1≠0

由于$\hat{\beta}_1$也是正态变量y_1，y_2，…，y_n的线性组合，故$\hat{\beta}_1$也服从正态分布[3]，为：

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2/S_{xx}\right) \tag{10}$$

转换为标准正态分布得：

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1) \tag{11}$$

由此可得检验统计量：

$$t_1 = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}}}{\sqrt{\frac{MSE(n-2)}{\sigma^2(n-2)}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t(n-2) \quad (12)$$

当原假设为真时，即β_1=0时，有t_1=$\hat{\beta}_1$/se($\hat{\beta}_1$)服从自由度为n−2的t分布。当$|t_1| > t_{0.025}(n-2)$或者t_1对应的P值小于0.05时，可认为β_1与0之间的差异有统计学意义。

1.4　多重线性回归模型中参数与0之间差异的t检验

在多重线性回归模型中对参数β_j(0≤j≤k)进行假设检验的原假设和备择假设分别为：

H_0:β_j=0 vs H_1:β_j≠0

多重线性回归模型中的参数β_j的检验统计量t_j服从自由度为n−k−1的t分布[4]：

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t(n-k-1) \quad (13)$$

上式中，C_{jj}是k阶矩阵$(X^TX)^{-1}$中第j行第j列位置上的元素。

$$\hat{\sigma}^2 = \frac{SSE}{n-k-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1} \quad (14)$$

当原假设为真时，即β_j=0时，有$t_j = \hat{\beta}_j / \mathrm{se}(\hat{\beta}_j)$服从自由度为$n$−$k$−1的$t$分布。当$|t_j| > t_{0.025}(n-k-1)$或者$t_j$对应的$P$值小于0.05时，可认为$\beta_j$与0之间的差异有统计学意义。

2　简单线性回归模型中参数与0之间差异t检验的实例

2.1　简单线性回归分析的数据结构

【例1】研究20名儿童血红蛋白(y)与血铁(x)之间的关系[5]。数据见表1。

表1　20名儿童血红蛋白(y)与血铁(x)的测定资料

n	x(ug/dL)	y(mg/dL)
1	518.7	13.5
2	467.3	13.0
3	469.8	11.0
…	…	…
18	283.3	7.8
19	312.5	7.3
20	294.7	7.0

通过t检验来判断简单线性回归模型中的截距(β_0)和斜率(β_1)与0之间差异是否有统计学意义，若斜率与0之间差异有统计学意义，则说明血红蛋白与血铁之间存在着线性依赖关系，即血红蛋白会随着血铁的变化呈线性变化趋势。

2.2　构建与求解简单线性回归模型的SAS程序

根据例1中数据进行简单线性回归分析，并对回归方程的截距和斜率进行t检验。

SAS程序如下：

```
data test1;/*1 输入数据*/
input x y @@;
cards;
518.7  13.5  467.3  13.0
469.8  11.0  456.6  14.3
448.7  12.5  424.1  12.5
405.6  11.8  446.0  11.5
416.7  11.0  430.8  10.7
409.8  10.2  384.1  10.0
356.3   9.5  388.6   9.4
325.9   8.8  292.8   6.3
332.8   7.3  283.0   7.8
312.5   7.3  294.7   7.0
;
run;
proc reg;/*2 输出简单线性回归的结果*/
model y=x;
run;
proc reg;/*3 输出删除截距项的简单线性回归的结果*/
model y=x/noint;
run;
proc sgplot NOAUTOLEGEND;/*4 生成散点图和回归直线*/
scatter x=x y=y/markerattrs=(symbol=circlefilled color=black);
reg x=x y=y/lineattrs=(color=black);
xaxis label='x(ug/dL)';
yaxis label='y(mg/dL)';
run;
```

【程序说明】以上SAS程序由1个数据步和3个过程步构成。数据步建立例1中的数据集test1，输入20例儿童的血铁(x)和血红蛋白(y)数据。第一个过程步调用REG过程，建立简单线性回归方程，并对总体和参数进行检验。第二个过程步也调用REG过程，

但通过noint语句删除了方程的截距项，是对第一个过程步的调整。第三个过程步调用SGPLOT过程，通过scatter语句绘制散点图，通过reg语句绘制回归直线。

2.3 简单线性回归分析中与t检验有关的结果

【SAS主要输出结果及解释】

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	$Pr>\|t\|$
Intercept	1	−2.06406	1.26321	−1.63	0.1196
x	1	0.03137	0.00317	9.9	<0.0001

以上是例1中数据的简单线性回归模型参数检验的结果，采用t检验。例1共20例数据，所以截距β_0的检验统计量t_0和斜率β_1的检验统计量t_1均服从自由度为18的t分布。β_0的最小二乘估计值$\hat{\beta}_0$=−2.06406，$\hat{\beta}_0$的标准误差$se(\hat{\beta}_0)$=1.26321，β_0的检验统计量t_0=−1.63，t_0对应的P值为0.1196，所以截距β_0与0之间差异无统计学意义。β_1的最小二乘估计值$\hat{\beta}_1$=0.03137，$\hat{\beta}_1$的标准误差$se(\hat{\beta}_1)$=0.00317，β_1的检验统计量t_1=9.9，t_1对应的P值小于0.0001，故斜率β_1与0之间差异有统计学意义。由于截距β_0与0之间差异无统计学意义，所以回归方程的截距项β_0为0，从而应重新拟合下面的回归方程：

$$\hat{y}=\hat{\beta}_1 x \tag{15}$$

【SAS主要输出结果及解释】

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	$Pr>\|t\|$
x	1	0.02626	0.00055080	47.68	<0.0001

以上是根据调整后的过程步可以得到的简单线性回归参数检验的结果（删除截距项），t_1=47.68，P<0.0001，所以斜率β_1与0之间差异有统计学意义，故由例1中数据得到的线性回归方程为$\hat{y}$=0.02626x。图1是用该数据生成的散点图以及根据回归方程拟合的回归直线。

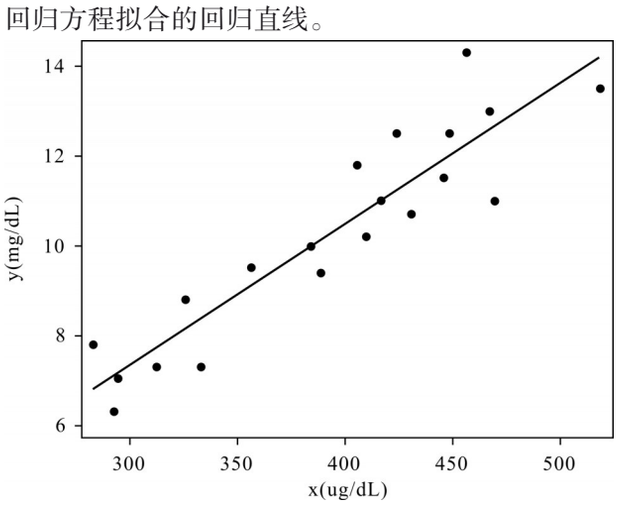

图1 (x,y)散点图及拟合的回归直线

3 多重线性回归模型的参数与0之间差异t检验的实例

3.1 多重线性回归分析的数据结构

【例2】研究26例糖尿病患者的血清总胆固醇(x_1)、甘油三酯(x_2)、空腹胰岛素(x_3)、糖化血红蛋白(x_4)与空腹血糖(y)之间的关系[6]。数据见表2。

表2 26例糖尿病患者血样中相关指标测定结果

n	x_1	x_2	x_3	x_4	y
1	5.68	1.90	4.53	8.20	11.20
2	3.97	1.64	7.32	6.90	8.80
…	…	…	…	…	…
25	11.54	10.89	1.20	10.50	20.00
26	3.84	1.20	6.54	9.60	10.40

通过t检验来判断多重线性回归模型中的总体截距和各个自变量对应的系数与0比较是否存在统计学差异，从而判断各个自变量是否有意义。本研究中，将空腹血糖设为因变量，将血清总胆固醇、甘油三酯、空腹胰岛素和糖化血红蛋白设为自变量。

3.2 构建与求解多重线性回归模型的SAS程序

根据例2的数据进行多重线性回归分析，并对回归方程的各参数进行t检验。

SAS程序如下：

```
data test2;/*1 输入数据*/
input x1 x2 x3 x4 y;
cards;
5.68  1.90  4.53  8.20  11.20
3.97  1.64  7.32  6.90  8.80
…
11.54  10.89  1.20  10.50  20.00
3.84  1.20  6.54  9.60  10.40
;
run;
```

```
proc reg;/*2 输出多重线性回归的结果(没采用变量筛选)*/
model y=x1 x2 x3 x4;
run;
proc reg;/*3 输出多重线性回归的结果(采用变量筛选)*/
model y=x1 x2 x3 x4/selection=stepwise sle=0.05 sls=0.05 stb;
run;
```

【程序说明】以上SAS程序由3步构成(实际使用只需要第1步和第3步),包含1个数据步和2个过程步。数据步建立例2中的数据集test2,输入26例糖尿病患者血清总胆固醇(x_1)、甘油三酯(x_2)、空腹胰岛素(x_3)、糖化血红蛋白(x_4)和空腹血糖(y)的数据。第一个过程步调用REG过程,但本过程没有采用变量筛选,因此即使某个变量不具有统计学意义也会被纳入多重线性回归模型。第二个过程步也调用了REG过程对回归方程进行总体检验和参数检验,但为了避免多重共线性,该过程步采用逐步回归法(stepwise)进行变量筛选,只有具有统计学意义的变量会被纳入多重线性回归模型。

3.3　多重线性回归分析中与t检验有关的结果

【SAS主要输出结果及解释】

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	$Pr>\|t\|$	Standardized Estimate
Intercept	1	4.91480	2.14919	2.29	0.0322	0
x2	1	0.43796	0.13425	3.26	0.0036	0.38281
x3	1	−0.29949	0.09699	−3.09	0.0054	−0.37405
x4	1	0.81267	0.20624	3.94	0.0007	0.48561

以上是例2中数据进行多重线性回归分析的参数检验的结果(采用变量筛选),采用的是t检验。例2包括26例数据,经变量筛选后只留下3个自变量,故β_j的检验统计量t_j均服从自由度为22的t分布。自变量血清总胆固醇(x_1)在变量筛选过程中被剔除。总体截距β_0对应的检验统计量t_0=2.29,P=0.0322,说明β_0与0之间差异有统计学意义;甘油三酯(x_2)的系数β_2对应的检验统计量t_2=3.26,P=0.0036,说明β_2与0之间差异有统计学意义;空腹胰岛素(x_3)的系数β_3对应的检验统计量t_3=−3.09,P=0.0054,说明β_3与0之间差异有统计学意义;糖化血红蛋白(x_4)的系数β_4对应的检验统计量t_4=3.94,P=0.0007,说明β_4与0之间差异有统计学意义。总体截距和三个自变量的回归系数与0之间差异都有统计学意义,多重线性回归方程如下:

$$\hat{y} = 4.9148 + 0.43796x_2 - 0.29949x_3 + 0.81267x_4 \quad (16)$$

4　讨论与小结

4.1　讨论

常规t检验(定量资料均值比较)在SAS中是用TTEST过程步实现。而本文是通过SAS中的REG过程对简单线性回归模型和多重线性回归模型中参数与0之间的差异进行t检验。若自变量的回归系数与0之间差异无统计学意义,则说明该自变量对因变量的影响可忽略不计;反之,则说明该自变量对因变量的影响有统计学意义。此外,还需对截距项与0之间的差异进行t检验,若检验结果为差异无统计学意义,则构建的回归方程中截距项为0。进行线性回归分析时应注意:①数据应满足使用线性回归分析的前提条件;②与0之间差异无统计学意义的参数可以在SAS程序中使用相应的语句进行调整,使其不出现在最终构建的线性回归方程中。

4.2　小结

综上所述,线性回归模型中参数与0比较t检验与常规t检验在SAS实现上虽有差异,但检验的原理是相同的,都是根据样本数据建立相应服从t分布的检验统计量,并对检验统计量进行检验。

参考文献

[1] 冯国双, 罗凤基. 医学案例统计分析与SAS应用[M]. 北京: 北京大学医学出版社, 2015: 117-119.

[2] 贾俊平, 何晓群, 金勇进. 统计学[M]. 6版. 北京: 中国人民大学出版社, 2015: 134-314.

[3] 茆诗松, 程依明, 濮晓龙. 概率论与数理统计教程[M]. 北京: 高等教育出版社, 2011: 252-473.

[4] Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis[M]. New Jersey: John Wiley & Sons, 2012: 12-128.

[5] 谷恒明, 胡良平. 简单线性回归分析及其应用[J]. 四川精神卫生, 2017, 30(6): 494-497.

[6] 胡良平. 多重线性回归分析的核心内容与关键技术概述[J]. 四川精神卫生, 2018, 31(1): 1-6.

(收稿日期:2020-07-17)

(本文编辑:戴浩然)