
利用SPSS软件实现岭回归分析

作者：陶立元 赵一鸣 来源：临床流行病学和循证医学

本文原地址：<https://www.iikx.com/news/statistics/6230.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

利用SPSS软件实现岭回归分析

。近日有医院的小伙伴问起岭回归分析的SPSS操作，在此与大家一起复习一下。

岭回归分析(Ridge Regression)

是一种改良的最小二乘估计方法，它是用于解决在线性回归分析中自变量存在共线性的问题。什么?共线性是什么?共线性就是指自变量之间存在一种完全或良好的线性关系，进而导致自变量相关矩阵之行列式近似为0，导致最小二乘估计失效。此时统计学家就引入了 k 个单位阵(I)，使得回归系数可估计。设么?没看懂，那就算了。

知道岭回归分析就是用来解决多重共线性的问题，就够了。在医学科研的实际工作中，往往不需要创造算法，会用算法就行。当然如果你有心研究其原理，那更是极好的。

下面我们还是通过实例来学习岭回归分析的应用条件和SPSS实习方法吧。用SPSS自带的例子(来自SPSS 20.0版的示例数据库，其他版本的就别找了)，某研究者想了解B超下胎儿的身长、头围、体重与胎儿受精周数之间的关系，即B超测得上述参数之后，用它们来推测胎儿的受精时长(周数)。我们很容易想到用多重线性回归来解决，以胎儿周数为因变量，以身长、头围和体重为因变量，做回归之后我们发现，结果如下：

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	966.023	3	322.008	238.130	.000 ^b
	Residual	24.340	18	1.352		
	Total	990.364	21			

a. Dependent Variable: 胎儿受精周龄

b. Predictors: (Constant), 体重(g), 身长(cm), 头围(cm)

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	11.012	1.718		6.408	.000		
	身长(cm)	1.693	.370	2.393	4.580	.000	.005	200.013
	头围(cm)	-2.159	.536	-2.156	-4.031	.001	.005	209.430
	体重(g)	.007	.001	.751	6.531	.000	.103	9.688

a. Dependent Variable: 胎儿受精周龄

不会吧?!头围竟然与周龄成负相关，开玩笑啊。这个方程肯定是有问题，细心的读者也已经发现方差膨胀因子(VIF)大到200多了(VIF是用来判断自变量共线性的一种方法，如果大于10即认为存在较为严重的共线性)。现在该怎么办?岭回归该发挥作用了。

岭回归分析在SPSS中没有可供点击的对话框，我们需要写一段超级简单的语法来调用SPSS的宏。SPSS公司可能也觉得羞愧，没有提供人机交互的对话框，于是他们提供了一段宏程序，存储路径为“你的SPSS安装目录SPSSStatistics22SamplesSimplified ChineseRidge regression.sps”。

```
*使用include调用SPSS的宏.
INCLUDE 'C:\Program Files\IBM\SPSS\Statistics\22\Samples\Simplified Chinese\Ridge regression.sps'.
RIDGEREG ENTER = long touwei weight *ENTER后面是自变量，用空格隔开.
/ dep=y *DEP后面是因变量.
/ start=0 *start是k的起始值，默认0.
/ inc=0.01 *inc是搜索步长，默认0.05.
/ stop=1 *stop是k的终止值，默认为1.
/ k=999. *k为个数，默认999.
EXECUTE.
```

我们在做岭回归分析时，只需要调用它就行，调用语法如下(*后面是注释)：

什么?你调用不了?哦，那是因为你没有SPSS目录下的修改权限。什么是修改权限?这是电脑问题，Windows为了保护其程序文件，一般默认不给用户修改权限，需要用户自己去改，至于怎么改自己问百度吧。

进行上述运算后，你会得到如下几个结果：1、不同K值下自变量的标准化回归系数;2、岭轨图，3、R方的变化图。

Report

C:\PROGRA~1\IBM\SPSS\STATIS~1\22\rr_tmp2.sav

R-SQUARE AND BETA COEFFICIENTS FOR ESTIMATED VALUES OF K

K	RSQ	long	touwei	weight
-----	-----	-----	-----	-----
.00000	.97542	2.393471	-2.15574	.751090
.01000	.95864	.607980	-.283630	.657896
.02000	.95430	.426266	-.064358	.616119
.03000	.95209	.363391	.026211	.584518
.04000	.95055	.333852	.077547	.558977
.05000	.94932	.317746	.111300	.537699
.06000	.94827	.308130	.135457	.519612
.07000	.94734	.302014	.153711	.503999
.08000	.94649	.297930	.168027	.490351
.09000	.94569	.295087	.179562	.478294
.10000	.94494	.293032	.189047	.467544
.11000	.94423	.291492	.196971	.457883
.12000	.94355	.290295	.203673	.449138
.13000	.94289	.289329	.209399	.441171
.14000	.94225	.288521	.214331	.433874
.15000	.94162	.287820	.218608	.427154
.16000	.94100	.287193	.222338	.420938
.17000	.94039	.286614	.225605	.415162
.18000	.93979	.286068	.228478	.409776
.19000	.93919	.285542	.231011	.404734
.20000	.93859	.285030	.233250	.400000

临床流行病学和循证医学

如何选择结果呢?我们需要选择一定K值下的标准化回归系数，选择的原则是各个自变量的标准化回归系数趋于稳定时的最小K值。因为K值越小我们引入的单位矩阵就少，偏差就小。有同学说上图看不起，那么我们把岭迹图放大如下，就可以看出k大约在0.05时，各个自变量的标准化回归系数就趋于稳定了。

```

*-----
* Calculate raw coefficients from standardized ones, compute standard errors
* of coefficients, and an intercept term with standard error. Then print
* out similar to REGRESSION output.
*-----

. compute beta={b;0}.
. compute b= ( b &/ std ) * sy.
. compute intercpt=ybar-t(b)*t(xmean).
. compute b={b;intercpt}.
. compute xpx=(sse/(sst*(n-nv-1)))*inv(xpx+(k &* ident(nv,nv)))*xpx*
               inv(xpx+(k &* ident(nv,nv))).
. compute xpx=(sy*sy)*(mdia(1 &/ std)*xpx*mdia(1 &/ std)).
. compute seb=sqrt(diag(xpx)).
. compute seb0=sqrt( (sse)/(n*(n-nv-1)) + xmean*xpx*t(xmean)).
. compute seb={seb;seb0}.
. compute rrms={varname,'Constant'}.
. compute ratio=b &/ seb.
. compute ppp=2*(1-tcdf(abs(ratio),n-nv-1)).
. compute bvec={b,seb,beta,ratio,ppp}.
. print bvec
  /title='-----Variables in the Equation-----'
  /names=rrms /clabels='B' 'SE(B)' 'Beta.std' 't' 'sig'.
. print /space=newpage.
end if.

```

 临床流行病学和循证医学

有了上述的结果，我们就获得了岭回归的各个自变量的标准化回归系数，也算是做完了。但是有人又问了，我们能不能获得非标准的偏回归系数、t值和p值呢？当然是可以的，但是SPSS原始的宏不提供p值的计算，所以我们需要在SPSS的宏中加入这一句话“`. compute ppp=2*(1-tcdf(abs(ratio),n-nv-1)).`”，这句话就是计算p值的。同时我们对print结果略作修改。这句话加在下图的位置上：

Matrix

Run MATRIX procedure:

***** Ridge Regression with k = 0.05 *****

Mult R .974332767
RSquare .949324342
Adj RSqu .940878399
SE 1.669785903

ANOVA table

	df	SS	MS
Regress	3.000	940.176	313.392
Residual	18.000	50.187	2.788

F value	Sig F
112.4000404	.0000000

-----Variables in the Equation-----

	B	SE(B)	Beta.std	t	sig
long	.224719852	.044363019	.317746284	5.065476981	.000080592
touwei	.111459526	.060579319	.111300009	1.839893994	.082334817
weight	.005349162	.000969711	.537698642	5.516242721	.000030804
Constant	8.696670483	1.674598274	.000000000	5.193287618	.000061219

----- END MATRIX -----

 临床流行病学和循证医学

修改完宏之后，再修改上述调用语句，将其中的k改为等于0.05，SPSS就会做k=0.05时的岭回归分析，并给出各个自变量的检验结果，结果如下：

至此完整的岭回归分析就算做完了，各个自变量的标准化回归系数合理多了吧。什么？你还是没学会。哎，复习一下SPSS的语法运算吧，我只能帮你到这儿了。

更多 统计方法 请访问 <https://www.iikx.com/news/statistics/>

本文版权归作者所有，请勿用于商业用途，[爱科学iikx.com](http://iikx.com)转发