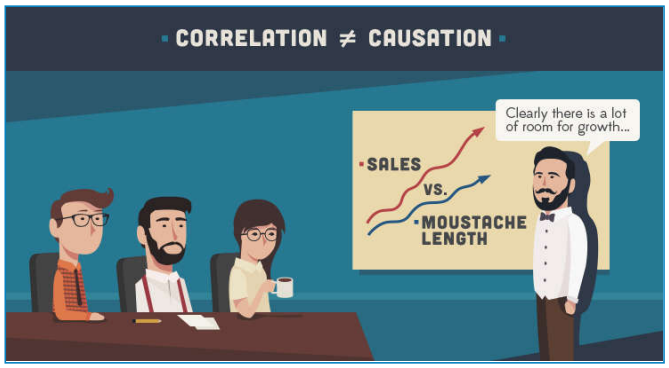
转 5种常用的相关分析方法

2017年05月16日 14:24:25 buptygz 阅读数：23620 标签： 相关性

相关分析（Analysis of Correlation）是网站分析中经常使用的分析方法之一。通过对不同特征或数据间的关系进行分析，发现业务运营中的关键影响及驱动[illegible]针对业务测。本篇文章将介绍5种常用的分析方法。在开始介绍相关分析之前，需要特别说明的是相关关系不等于因果关系。



相关分析的方法很多，初级的方法可以快速发现数据之间的关系，如正相关，负相关或不相关。中级的方法可以对数据间关系的强弱进行度量，如完全相关，不完全相关等以将数据间的关系转化为模型，并通过模型对未来的业务发展进行预测。下面我们以一组广告的成本数据和曝光量数据对每一种相关分析方法进行介绍。

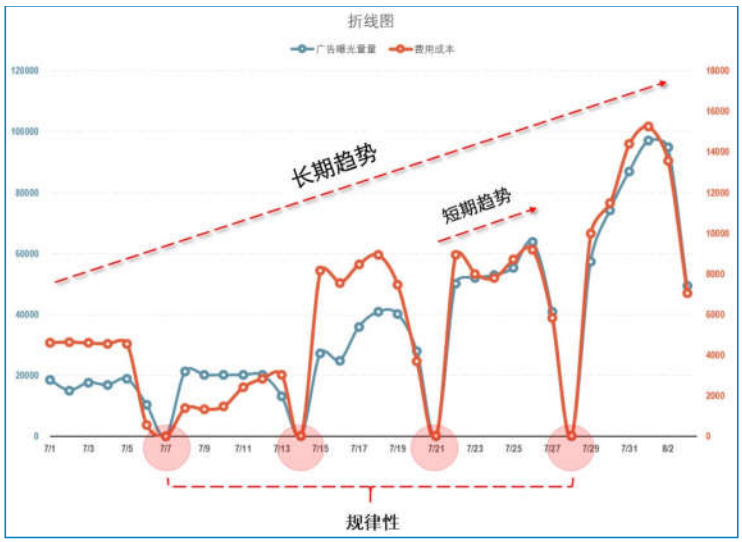
以下是每日广告曝光量和费用成本的数据，每一行代表一天中的花费和获得的广告曝光数量。凭经验判断，这两组数据间应该存在联系，但仅通过这两组数据我们无法证明[illegible]在，也无法对这种关系的强度进行度量。因此我们希望通过相关分析来找出这两组数据之间的关系，并对这种关系进度度量。

投放时间	广告曝光量(y)	费用成本(x)
2016/7/1	18,481	4,616
2016/7/2	15,094	4,649
2016/7/3	17,619	4,600
2016/7/4	16,825	4,557
2016/7/5	18,811	4,541
2016/7/6	10,430	568
2016/7/7	18	-
2016/7/8	...	...
2016/7/9	...	...

1，图表相关分析（折线图及散点图）

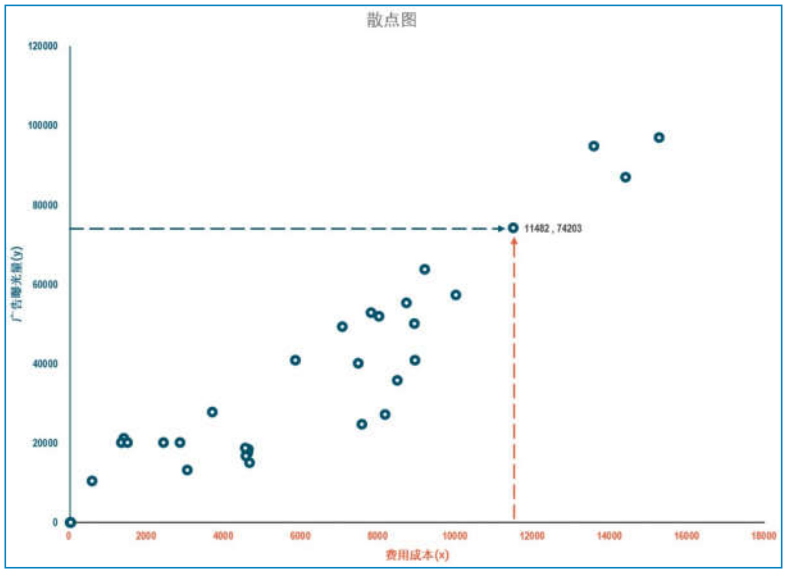
第一种相关分析方法是将数据进行可视化处理，简单的说就是绘制图表。单纯从数据的角度很难发现其中的趋势和联系，而将数据点绘制成图表后趋势和联系就会变的清晰[illegible]显时间维度的数据，我们选择使用折线图。

为了更清晰的对比这两组数据的变化和趋势，我们使用双坐标轴折线图，其中主坐标轴用来绘制广告曝光量数据，次坐标轴用来绘制费用成本的数据。通过折线图可以发现告曝光量两组数据的变化和趋势大致相同，从整体的大趋势来看，费用成本和广告曝光量两组数据都呈现增长趋势。从规律性来看费用成本和广告曝光量数据每次的最低点[illegible]天。从细节来看，两组数据的短期趋势的变化也基本一致。



经过以上这些对比，我们可以说广告曝光量和费用成本之间有一些相关关系，但这种方法在整个分析过程和解释上过于复杂，如果换成复杂一点的数据或者相关度较低的数
问题。

比折线图更直观的是散点图。散点图去除了时间维度的影响，只关注广告曝光量和费用成本这里两组数据间的关系。在绘制散点图之前，我们将费用成本标识为X，也就是自
曝光量标识为y，也就是因变量。下面是一张根据每一天中广告曝光量和费用成本数据绘制的散点图，X轴是自变量费用成本数据，Y轴是因变量广告曝光量数据。从数据点的
发现，自变量x和因变量y有着相同的变化趋势，当费用成本的增加后，广告曝光量也随之增加。



折线图和散点图都清晰的表示了广告曝光量和费用成本两组数据间的相关关系，优点是对相关关系的展现清晰，缺点是无法对相关关系进行准确的度量，缺乏说服力。并且
时也无法完成各组数据间的相关分析。若要通过具体数字来度量两组或两组以上数据间的相关关系，需要使用第二种方法：协方差。

2，协方差及协方差矩阵

第二种相关分析方法是计算协方差。协方差用来衡量两个变量的总体误差，如果两个变量的变化趋势一致，协方差就是正值，说明两个变量正相关。如果两个变量的变化趋
就是负值，说明两个变量负相关。如果两个变量相互独立，那么协方差就是0，说明两个变量不相关。以下是协方差的计算公式：

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

下面是广告曝光量和费用成本间协方差的计算过程和结果，经过计算，我们得到了一个很大的正值，因此可以说明两组数据间是正相关的。广告曝光量随着费用成本的增长
工作中不需要按下面的方法来计算，可以通过Excel中COVAR()函数直接获得两组数据的协方差值。

投放时间	广告曝光量(y)	费用成本(x)	$y_i-\bar{y}$	$x_i-\bar{x}$	$(x_i-\bar{x})(y_i-\bar{y})$
2016/7/1	18,481	4,616	-16,344	-1,283	20,966,307
2016/7/2	15,094	4,649	-19,731	-1,250	24,663,380
2016/7/3	17,619	4,600	-17,206	-1,299	22,350,167
2016/7/4	16,825	4,557	-18,000	-1,342	24,154,482
2016/7/5	18,811	4,541	-16,014	-1,358	21,741,416
2016/7/6	10,430	568	-24,395	-5,331	130,058,373
2016/7/7	18	-	-34,807	-5,899	205,327,475
2016/7/8	…	…	…	…	…
2016/7/9	…	…	…	…	…
均值:	**34,825**	**5,899**		**求和:**	**3,508,979,770**
				n=34	106,332,720

协方差只能对两组数据进行相关性分析，当有两组以上数据时就需要使用协方差矩阵。下面是三组数据x，y，z，的协方差矩阵计算公式。

$$C=\begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

协方差通过数字衡量变量间的相关性，正值表示正相关，负值表示负相关。但无法对相关的密切程度进行度量。当我们面对多个变量时，无法通过协方差来说明那两组数据要衡量和对比相关性的密切程度，就需要使用下一个方法：相关系数。，

3，相关系数

第三个相关分析方法是相关系数。相关系数(Correlation coefficient)是反应变量之间关系密切程度的统计指标，相关系数的取值区间在1到-1之间。1表示两个变量完全线性相变量完全负相关，0表示两个变量不相关。数据越趋近于0表示相关关系越弱。以下是相关系数的计算公式。

$$r_{xy}=\frac{S_{xy}}{S_x S_y}$$

其中rxy表示样本相关系数，Sxy表示样本协方差，Sx表示X的样本标准差，Sy表示y的样本标准差。下面分别是Sxy协方差和Sx和Sy标准差的计算公式。由于是样本协方差和因此分母使用的是n-1。

Sxy样本协方差计算公式：

$$s_{xy}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

Sx样本标准差计算公式：

$$s_x=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$$

Sy样本标准差计算公式：

$$s_y=\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}$$

下面是计算相关系数的过程，在表中我们分别计算了x，y变量的协方差以及各自的标准差，并求得相关系数值为0.93。0.93大于0说明两个变量间正相关，同时0.93非常接近变量间高度相关。

投放时间	广告曝光量(y)	费用成本(x)	$y_i-\bar{y}$	$x_i-\bar{x}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})^2$
2016/7/1	18,481	4,616	-16,344	-1,283	20,966,307	267,109,992	1,645,712
2016/7/2	15,094	4,649	-19,731	-1,250	24,663,380	389,292,630	1,562,532
2016/7/3	17,619	4,600	-17,206	-1,299	22,350,167	296,029,230	1,687,435
2016/7/4	16,825	4,557	-18,000	-1,342	24,154,482	323,982,000	1,800,838
2016/7/5	18,811	4,541	-16,014	-1,358	21,741,416	256,432,182	1,843,330
2016/7/6	10,430	568	-24,395	-5,331	130,058,373	595,091,630	28,424,497
2016/7/7	18	-	-34,807	-5,899	205,327,475	1,211,492,442	34,799,533
2016/7/8	…	…	…	…	…	…	…
2016/7/9	…	…	…	…	…	…	…
$n=34$					S_{xy}	S_y	S_x
					106,332,720	26,615	4,266
					r_{xy}		
					0.936447666		

在实际工作中，不需要上面这么复杂的计算过程，在Excel的数据分析模块中选择相关系数功能，设置好x，y变量后可以自动求得相关系数的值。在下面的结 以看到，费用成本的相关系数与我们手动求的结果一致。

	广告曝光量(y)	费用成本(x)
广告曝光量(y)	1	
费用成本(x)	0.936447666	1

相关系数的优点是可以通过数字对变量的关系进行度量，并且带有方向性，1表示正相关，-1表示负相关，可以对变量关系的强弱进行度量，越靠近0相关性越弱。缺点是无 对数据进行预测，简单的说就是没有对变量间的关系进行提炼和固化，形成模型。要利用变量间的关系进行预测，需要使用到下一种相关分析方法，回归分析。，

4，一元回归及多元回归

第四种相关分析方法是回归分析。回归分析（regression analysis)是确定两组或两组以上变量间关系的统计方法。回归分析按照变量的数量分为一元回归和多元回归。两个 归，两个以上变量使用多元回归。进行回归分析之前有两个准备工作，第一确定变量的数量。第二确定自变量和因变量。我们的数据中只包含广告曝光量和费用成本两个变 元回归。根据经验广告曝光量是随着费用成本的变化而改变的，因此将费用成本设置为自变量x，广告曝光量设置为因变量y。

以下是一元回归方程，其中y表示广告曝光量，x表示费用成本。b0为方程的截距，b1为斜率，同时也表示了两个变量间的关系。我们的目标就是b0和b1的值，知道了这两个 变量间的关系。并且可以通过这个关系在已知成本费用的情况下预测广告曝光量。

$$y = b_0 + b_1 x$$

这是b1的计算公式，我们通过已知的费用成本x和广告曝光量y来计算b1的值。

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

以下是通过最小二乘法计算b1值的具体计算过程和结果，经计算，b1的值为5.84。同时我们也获得了自变量和因变量的均值。通过这三个值可以计算出b0的值。

投放时间	广告曝光量(y)	费用成本(x)	$y_i-\bar{y}$	$x_i-\bar{x}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$
2016/7/1	18,481	4,616	-16,344	-1,283	20,966,307	1,645,712
2016/7/2	15,094	4,649	-19,731	-1,250	24,663,380	1,562,532
2016/7/3	17,619	4,600	-17,206	-1,299	22,350,167	1,687,435
2016/7/4	16,825	4,557	-18,000	-1,342	24,154,482	1,800,838
2016/7/5	18,811	4,541	-16,014	-1,358	21,741,416	1,843,330
2016/7/6	10,430	568	-24,395	-5,331	130,058,373	28,424,497
2016/7/7	18	-	-34,807	-5,899	205,327,475	34,799,533
2016/7/8	…	…	…	…	…	…
2016/7/9	…	…	…	…	…	…
	$\bar{y}$	$\bar{x}$			$\sum(x_i-\bar{x})(y_i-\bar{y})$	$\sum(x_i-\bar{x})^2$
	34,825	5,899			3,508,979,770	600,651,674
					$b_1 =$	5.841954536

以下是b0的计算公式，在已知b1和自变量与因变量均值的情况下，b0的值很容易计算。

$$b_0 = \bar{y} - b_1 \bar{x}$$

将自变量和因变量的均值以及斜率b1代入到公式中，求出一元回归方程截距b0的值为374。这里b1我们保留两位小数，取值5.84。

$$b_0 = \bar{y} - b_1 \bar{x} = 34825 - 5.84 * 5899 = 374$$

在实际的工作中不需要进行如此繁琐的计算，Excel可以帮我们自动完成并给出结果。在Excel中使用数据分析中的回归功能，输入自变量和因变量的范围后可以自动获得b0 值362.15和b1的值5.84。这里的b0和之前手动计算获得的值有一些差异，因为前面用于计算的b1值只保留了两位小数。

这里还要单独说明下R Square的值0.87。这个值叫做判定系数，用来度量回归方程的拟合优度。这个值越大，说明回归方程越有意义，自变量对因变量的解释度越高。

SUMMARY OUTPUT

回归统计

Multiple R	0.936447666
R Square	0.87693423
Adjusted R Squar	0.873088425
标准误差	9481.556867
观测值	34

方差分析

	df	SS	MS	F	Significance F
回归分析	1	20499300283	20499300283	228.0235638	4.12012E-16
残差	32	2876797460	89899920.62		
总计	33	23376097743			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	362.1503948	2802.246201	0.129235752	0.897980008	-5345.838329	6070.139119	-5345.838329	6070.139119
费用成本(x)	5.841954536	0.386872899	15.10044913	4.12012E-16	5.053920227	6.629988845	5.053920227	6.629988845

将截距b0和斜率b1代入到一元回归方程中就获得了自变量与因变量的关系。费用成本每增加1元，广告曝光量会增加379.84次。通过这个关系我们可以根据[illegible]广告曝光量，以根据转化所需的广告曝光量来反推投入的费用成本。获得这个方程还有一个更简单的方法，就是在Excel中对自变量和因变量生成散点图，然后选择添加趋势线，在添加趋势线选中显示公式和显示R平方值即可。

$$y = 374 + 5.84x$$

以上介绍的是两个变量的一元回归方法，如果有两个以上的变量使用Excel中的回归分析，选中相应的自变量和因变量范围即可。下面是多元回归方程。

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

5，信息熵及互信息

最后一种相关分析方法是信息熵与互信息。前面我们一直在围绕消费成本和广告曝光量两组数据展开分析。实际工作中影响最终效果的因素可能有很多，并且不一定都是数值化的，我们站在更高的维度来看之前的数据。广告曝光量只是一个过程指标，最终要分析和关注的是用户是否购买的状态。而影响这个结果的因素也不仅仅是消费成本或其他数值化的一些特征值。例如用户所在的城市，用户的性别，年龄区间分布，以及是否第一次到访网站等等。这些都不能通过数字进行度量。

度量这些文本特征值之间相关关系的方法就是互信息。通过这种方法我们可以发现哪一类特征与最终的结果关系密切。下面是我们模拟的一些用户特征和数据。在这些数据中，除了消费成本和广告曝光量数据，只关注特征与状态的关系。

城市	消费成本	广告曝光量	性别	新用户	年龄分布	状态
杭州	13,588	78,844	男	是	岁25-34	未购买
杭州	20,738	120,473	男	否	岁25-34	未购买
北京	18,949	111,982	女	否	岁25-34	购买
上海	30,908	167,093	男	是	岁35-45	未购买
北京	27,822	167,897	男	否	岁<25	购买
北京	30,100	185,418	男	否	岁35-45	未购买
南京	23,317	129,550	女	是	岁25-34	未购买
广州	19,057	120,861	女	否	岁<25	未购买
北京	16,091	101,915	女	否	岁25-34	购买
…	…	…	…	…	…	…
…	…	…	…	…	…	…

对于信息熵和互信息具体的计算过程请参考我前面的文章《决策树分类和预测算法的原理及实现》，这里直接给出每个特征的互信息值以及排名结果。经过计算城市与购买率最高，所在城市为北京的用户购买率较高。

特征	互信息	排名
城市	0.557727779	1
性别	0.072780226	4
新用户	0.251629167	2
年龄分布	0.156656615	3

到此为止5种相关分析方法都已介绍完，每种方法各有特点。其中图表方法最为直观，相关系数方法可以看到变量间两两的相关性，回归方程可以对相关关系进行提炼，并生成预测，互信息可以对文本类特征间的相关关系进行度量。