概率论与数理统计课程小论文

Perason 系数与相关性检测

姓　名	psp
日　期	**2023.11.30**

摘　要

本研究首先分析了Perason系数相关的定义、性质及其在数据相关性的计算中的应用；然后利用第六届“泰迪杯”数据分析技能赛中包含2019年近11000家企业的财务数据，通过对企业财务数据的分析，使用Pearson相关系数来衡量各个指标与利润总额之间的相关性。研究发现，YYSR、YWFY、YYCB等指标与利润总额之间存在较强的线性相关性，而其他指标与利润总额的相关性较弱。这一研究结果对于理解企业财务数据的关联关系具有重要意义，也展现了Perason相关系数在数据分析领域具有的重要作用和重大价值。

关键词：Pearson相关系数，企业财务数据，利润总额，相关性分析

研究背景

笔者近期参加了第六届“泰迪杯”数据分析技能赛，选择的题目是 B 题（企业财务数据分析与造假识别），其中任务 3.1 要求是读取企业财务数据样本集文件“financial_data.csv”，计算各个指标与利润总额的相关性，并选出相关度最高的 5 个指标。作为最常用的一种相关系数，Perason 系数在这个任务中有很好的应用，此外，在各种现实问题中，Perason 系数也有很多应用。因此，作者在此继续深入分析 Perason 系数相关的定义、性质及其在数据相关性的计算中的应用，以探寻其深层含义。

研究的目的和意义

使用 Pearson 相关系数进行相关性检测具有以下现实意义：

量化关系：它提供了两个变量之间线性关系强度的定量度量。在经济学、金融学和社会科学等领域，Pearson 相关系数对了解变量之间关系的强度非常重要。

预测能力：它有助于理解一个变量的变化如何预测另一个变量的变化。例如，在金融领域，它被用于理解一个股票的变动如何预测另一个股票的变动。

数据探索：在探索性数据分析中，通常使用它快速识别变量之间的潜在关系。这可以指导进一步的、更详细的分析。

变量选择：在机器学习和统计学等领域，相关性分析用于识别和选择预测建模中的变量。它有助于理解哪些变量最有可能对预测结果产生影响。

质量控制：在制造业和工程领域等领域，相关性分析可用于理解不同过程变量与产品质量或工艺效率的关系。

风险管理：在金融和保险领域，了解不同资产和负债之间的相关性对于风险管理和分散投资非常重要。

研究和政策制定：在社会科学和公共政策领域，了解不同因素之间的相关性可以指导研究和决策过程。

验证检查：它可用于检查统计分析的基本假设的有效性。例如，在回归分析中，检查多重共线性非常重要，相关性分析可以帮助识别潜在问题。

总之，Pearson 相关系数是理解和量化变量之间关系的强大工具，在各个领域的研究、决策和实际问题解决中具有广泛应用。

具体研究

一、理论基础

1、协方差及协方差矩阵

变量 X 的方差的定义为：

$$\mathrm{Var}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})}{n-1}$$

两个变量的协方差的定义为：

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

协方差有如下的意义：

当 X 与 Y 两个变量变化趋势相同，协方差为正值，说明两变量正相关；

当 X 与 Y 两个变量变化趋势相反，协方差为负值，说明两变量负相关；

当 X 与 Y 两个变量相互独立，协方差为 0，说明两变量不相关；

三个变量的协方差：三维协方差矩阵

$$C=\begin{bmatrix}\mathrm{Cov}(X,X) & \mathrm{Cov}(X,Y) & \mathrm{Cov}(X,Z)\\ \mathrm{Cov}(Y,X) & \mathrm{Cov}(Y,Y) & \mathrm{Cov}(Y,Z)\\ \mathrm{Cov}(Z,X) & \mathrm{Cov}(Z,Y) & \mathrm{Cov}(Z,Z)\end{bmatrix}$$

由于协方差满足对称性 $\mathrm{Cov}(X,Y)=\mathrm{Cov}(Y,X)$，因此上述矩阵是实对称矩阵。

在实际应用中，三维协方差矩阵可以帮助我们分析和理解变量之间的相关性。通过观察协方差矩阵的各个元素，我们可以判断变量之间的线性关系、强度和方向。

协方差矩阵的对角线元素表示各个变量的方差，非对角线元素表示两个变量之间的协方差。如果协方差为正值，表示两个变量正相关；如果协方差为负值，表示两个变量负相关；如果协方差接近于零，则表示两个变量之间没有线性关系。

2、Pearson 系数

皮尔逊相关系数也称皮尔森积矩相关系数(Pearson product-moment correlation coefficient) ，是一种线性相关系数，是最常用的一种相关系数，记为ρ_{XY}，ρ_{XY}值介于-1 到 1 之间，绝对值越大表明 X 与 Y 的相关性越强。

皮尔逊相关系数定义为 X 与 Y 两个变量之间的协方差和标准差之积的商（或者说，归一化的协方差）定义式如下：

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\sigma_X^2\sigma_Y^2}}$$

样本的皮尔逊相关系数用 r 代表，计算公式如下：

$$r=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}$$

其中，$\overline{X}$ 和 $\overline{Y}$ 分别表示两者的样本均值。

R=1 表示两者完美的正向线性相关，即满足 $Y=aX+b(a>0)$的关系; $R=-1$ 表示两者完美的负向线性相关，即满足 $Y=aX+b(a<0)$的关系. 在 X-Y 散点图上看的话，散点图完全处于一条直线上。R = 0 则表示两者没有线性相关性，但不代表 X 与 Y 没有任何相关性。

3、Pearson 系数的数学性质

3.1 对称性

由 Pearson 系数的定义可知，Pearson 系数是对称的，即 $\rho_{XY} = \rho_{YX}$

3.2 位移不变性

若 $X' = X + x_0$，$Y' = Y + y_0$，其中x_0和y_0是常数，则 $\rho_{X'Y'} = \rho_{XY}$

直观地看，由于在皮尔逊相关计算中，无论是总体的还是样本的，分子和分母都通过减去均值将均值的影响消除掉了，因此 X 和 Y 的均值的变化不会影响两者之间的皮尔逊相关系数。

证明：

由于

$$E[(X' - \mu_{X'})] = E[(X + x_0) - (\mu_X + x_0)] = E[X - \mu_X]$$
$$E[(Y' - \mu_{Y'})] = E[(Y + y_0) - (\mu_Y + y_0)] = E[Y - \mu_Y]$$

从而

$$\rho_{X'Y'} = \frac{Cov(X', Y')}{\sigma_{X'}\sigma_{Y'}} = \frac{E[(X' - \mu_{X'})(Y' - \mu_{Y'})]}{\sqrt{E[(X' - \mu_{X'})^2]E[(Y' - \mu_{Y'})^2]}} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{E[(X - \mu_X)^2]E[(Y - \mu_Y)^2]}} = \rho_{XY}$$

3.3 尺度不变性

若 $X' = k_1X$，$Y' = k_2Y$，其中k_1和k_2是常数，则 $\rho_{X'Y'} = \rho_{XY}$

证明：

$$\rho_{X'Y'} = \frac{Cov(X', Y')}{\sigma_{X'}\sigma_{Y'}} = \frac{E[(X' - \mu_{X'})(Y' - \mu_{Y'})]}{\sqrt{E[(X' - \mu_{X'})^2]E[(Y' - \mu_{Y'})^2]}} = \frac{E[k_1k_2(X - \mu_X)(Y - \mu_Y)]}{\sqrt{k_1^2k_2^2E[(X - \mu_X)^2]E[(Y - \mu_Y)^2]}} = \rho_{XY}$$

综合以上两个性质，能得出 Pearson 系数的线性性质：

若 $X' = k_1X + x_0$，$Y' = k_2Y + y_0$，其中k_1、k_2、x_0和y_0是常数，则$\rho_{X'Y'} = \rho_{XY}$

4、Pearson 系数应用于数据

Pearson 系数对数据质量要求较高：

实验数据是连续型变量；数据之间的差距不能太大，不能有离散点、异常值。

二、数据分析

“financial_data.csv”这一文件中包含 2019 年近 11000 家企业的财务数据，对每家企业，有 YYSR、YWFY、YYCB、YYSJJFJ、ZCJZSS 等十几个会影响企业利润总额（LRZE）的指标，因此，需要单独分析其余的各个指标与利润总额这一指标的相关性。以下是这个文件的部分截图。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	TICKER_SYM	END_DATE	LDBL	LXBZBS	ZCFZL	CHZZL	ZCBCL	CHZLDZCB	CQZBSYL	YYMLL	YSL	YSZKZZL	YYSR	YYCB	YYSJJFJ	YWFY	ZCJZSS	LRZE
2	4019	1	8.24911	1.83324	0.10018	3.13214	0.09723	0.07257	0.10554	0.97278	0.1561	0.26727	479500571	293550024	1505596.1	70399459	23119755	124977633
3	8166	1	1.3001	15.58644	0.66081	3.33273	0.07795	0.22293	0.21593	0.92089	0.12659	0.06131	5.325E+09	3.499E+09	28213514	1.374E+09	29930865	415222703
4	11737	1	4.62003	3.31066	0.17311	1.19244	0.08734	0.18267	0.10118	1.00495	0.09472	0.26297	1.684E+09	788367345	8298888.5	586469435	1766300.2	384325616
5	16479	1	2.77734	3.59707	0.35293	5.12706	0.04856	0.13312	0.06808	0.98878	0.14381	0.11057	2.23E+09	1.743E+09	11585500	394234510	4007795.8	121917741
6	16842	1	2.48068	68.08648	0.33309	4.27485	0.19819	0.29959	0.28469	0.90921	0.14099	0.25481	394554259	298639210	1621817.6	37244011	970096.95	60426742
7	16916	1	5.55654	0.70293	0.10189	16.86014	0.05577	0.05245	0.06023	1.02605	0.16882	0.09614	522718156	437769964	3748918.6	66226872	636576.22	27679388
8	17119	1	2.52615	3.07397	0.4289	1.32887	0.25811	0.34766	0.3979	1.09646	0.13797	0.36537	165523828	97225559	1058506.4	28423480	2511441.8	46178212
9	18545	1	1.78455	3.36717	0.39853	4.78355	0.26812	0.20125	0.4446	0.93506	0.14505	0.24986	447942915	307186691	2930419.5	51948515	730273.51	93125921
10	18909	1	2.79048	4.71719	0.29622	2.14924	0.08801	0.14306	0.12506	0.91017	0.15905	0.53731	975650646	654511333	7485360.8	97431050	45979463	187011454
11	22933	1	1.43097	1.93961	0.50852	10.69298	0.15955	0.16068	0.24645	0.70161	0.12839	0.42592	4.552E+09	3.652E+09	20112934	502237043	5431486	443489756
12	28956	1	1.2236	3.52316	0.51529	2.51839	0.05556	0.44573	0.10295	0.96295	0.2639	0.13572	3.603E+09	2.345E+09	28616315	1.107E+09	3709374.6	160498203
13	29690	1	3.06543	4.92482	0.23971	2.82624	0.23012	0.20818	0.28219	0.87298	0.1433	0.20557	282258940	140432103	3867579.4	58935187	-183082.1	80712335
14	31035	1	0.66595	-2.29372	0.83502	1.42114	-0.04239	0.58283	-0.1483	0.93169	0.0077	0.05102	3.722E+09	3.531E+09	59641833	776515208	59408938	-6.74E+08
15	38567	1	1.37863	34.49098	0.58725	0.95516	0.09358	0.41219	0.20472	0.90644	0.14818	0.1806	287815873	185292842	1423190.5	45757909	1600099.6	57187840
16	40307	1	3.6766	2.60944	0.24258	18.14536	0.266	0.09623	0.34584	0.6574	0.16497	0.08659	1.658E+10	1.378E+10	80752079	1.04E+09	30364124	1.788E+09
17	43064	1	1.25504	1.06737	0.61917	11.20846	0.11712	0.05316	0.25988	0.44114	0.10204	0.49105	5.544E+09	3.99E+09	23765085	546534522	245557079	587317134
18	46234	1	2.82411	5.94044	0.22301	3.23675	0.1976	0.24088	0.25049	0.72644	0.18659	0.24761	4.801E+09	3.118E+09	28855397	512893386	16791472	1.134E+09
19	48653	1	1.29846	2.87017	0.3556	6.31723	0.22201	0.34952	0.32427	0.9984	0.14813	0.23304	457858280	337154448	4027632.2	61951689	2339036.1	61396584
20	50357	1	0.83664	-2.54405	0.62224	20.74272	-0.03196	0.22126	-0.05839	0.86578	-0.00546	0.08282	1.657E+09	1.547E+09	2155539.3	143384206	67028919	-46878671
21	51412	1	3.03238	2.37227	0.2209	5.34601	0.10419	0.18107	0.1311	0.63414	0.21527	0.1765	2.606E+09	1.991E+09	11423872	372573134	4037795.1	242203780
22	59849	1	0.63475	2.87656	0.69581	5.79819	0.09579	0.30983	0.28235	0.89495	0.1673	0.39315	700275600	537160400	2713000	118076800	4381500	50672600
23	60127	1	4.74778	-1.29311	0.15903	3.95605	-0.0037	0.05024	-0.00437	1.193	0	0.07573	139940815	121822931	537987.74	29345521	-306735	-13670475
24	61598	1	1.9277	2.72781	0.41715	4.21954	0.0684	0.24395	0.10476	0.97885	0.10996	0.32904	2.72E+09	2.22E+09	14165254	398653123	12064551	161368006
25	73338	1	2.2896	1.03664	0.42414	0.81041	0.03395	0.48887	0.05374	0.70548	0.2707	0.28834	1.844E+09	957284269	15737736	780397351	42742346	49508886
26	75332	1	0.8787	2.0439	0.4504	11.05191	0.08422	0.17952	0.15324	0.24968	0.32643	0.12366	660151464	579714820	1445197	102475565	1053339.8	41752786
27	76021	1	0.78253	-0.47732	0.78777	3.20293	0.01562	0.19679	0.04519	0.84354	0.16395	0.08183	5.715E+09	4.89E+09	24219839	919378918	98555416	-2.16E+08
28	79499	1	1.34782	1.77418	0.47211	4.59097	0.07093	0.23032	0.12078	0.84769	0.17085	0.4927	1.823E+09	1.409E+09	8372107.9	299376952	-175055.8	108646920
29	80746	1	6.0974	7.43147	0.10596	1.93915	0.1424	0.18839	0.15652	1.02263	0.14195	0.16682	266418565	146676456	2260025.6	38958775	1672127.9	91591283

三、相关性分析

本文先读取“financial_data.csv”，再利用皮尔逊相关系数 scipy.stats.pearsonr 计算各个指标与利润总额（LRZE）的相关性，代码实现如下

```
# 计算各个指标与利润总额的相关性，并存储在字典中
correlation_dict = {}
profit_column = 'LRZE'

for indicator in selected_columns:
    correlation,  = pearsonr(df[indicator], df[profit_column])
    correlation_dict[indicator] = correlation
```

挑选相关度最高的 5 个指标，并展示出来

```
# 挑选相关度最高的 5 个指标
top_5_indicators = sorted(correlation_dict, key=correlation_dict.get, reverse=True)[:5]

# 输出相关度最高的 5 个指标及其相关性
print("相关度最高的 5 个指标：")
for indicator in top_5_indicators:
    correlation = correlation_dict[indicator]
    print(f"{indicator}: {correlation:.4f}")
```

相关度最高的五个指标及对应的相关性如下表所示：

YYSR	0.7827
YWFY	0.7728
YYCB	0.7377
YYSJJFJ	0.5654
ZCJZSS	0.2385

这一结果说明 YYSR、YWFY、YYCB 三个指标与利润总额（LRZE）有较强的线性相关性，其余指标和利润总额（LRZE）的相关性不强。

这一分析展现了 Perason 系数在数据分析领域的重要作用。

总结

本研究旨在探究 Pearson 相关系数在企业财务数据分析中的应用。通过对 2019 年近 11000 家企业的财务数据进行分析，计算各个指标与利润总额的相关性，并挑选出与利润

总额相关性最高的五个指标。研究结果表明，YYSR、YWFY、YYCB 等指标与利润总额之间存在较强的线性关系，这为企业经营决策提供了重要参考。此外，本研究还对 Pearson 相关系数的定义、性质进行了深入探讨，并阐述了其在数据相关性分析中的应用。这一研究对于理解和量化企业财务数据之间的关系具有重要意义，并为进一步的研究和实践提供了指导和参考。

参考文献

[1]张宇航. 面向空分装备的时滞皮尔逊相关性分析及关键变量预测方法研究[D].杭州电子科技大学,2023.DOI:10.27075/d.cnki.ghzdc.2023.000752.

[2]赵国杰,刘成浩.基于数据挖掘与相关性分析的电网一次设备缺陷预测方法[J].微型电脑应用,2023,39(09):138-141+145.

[3]金涛.基于相关性分析的一次风量控制优化研究[J].电站系统工程,2023,39(06):52-53+56.

[4]刘璐.新时代财务管理风险的管控措施分析[J].中国市场,2023(29):135-138.DOI:10.13939/j.cnki.zgsc.2023.29.135.

[5]吕新亚.优化财务管理让企业项目投资决策更科学[J].中国商界,2023(10):135-137.