**基于2017年全国各地区个人所得税税收收入的研究**

姓名

专业班级 学号

**摘要：**本文主要研究个人所得税税收收入的影响因素，以及不同地区的个人所得税税收收入水平。查阅资料得知可能影响个人所得税税收收入的因素有八个，即居民可支配收入、财政一般预算支出、地区生产总值、居民消费水平、就业人员工资总额、地方财政税收收入、社会消费品零售总额和商品零售价格指数。

基于2017年全国各地区的这些指标的具体数据和个人所得税税收收入的数据，本文首先进行了相关分析，判断得到除居民可支配收入外，其他因素与个人所得税税收收入的相关性是显著的；其次，本文以个税收入为因变量，其他八个可能影响因素为自变量，建立线性回归模型，得到居民可支配收入、居民消费水平、地方财政税收收入与个人所得税税收收入之间建立的线性关系回归模型具有显著的统计学意义；最后，本文对八个可能影响因素进行主成分分析，得到了贡献率占90.942%的两个主成分，并据此对各地区进行综合评分，得到广东省的各项指标的综合评分位列第一。

**关键词：**个人所得税税收收入；相关分析；回归分析；主成分分析

**1.问题提出与分析**

税收收入是国家经济收入的最主要来源，而其中个人所得税收入占据着税收收入的很大比重，于是需要对影响个人所得税税收收入的因素进行相关分析和回归分析，并通过主成分分析大致了解各地区的综合水平。

**2.数学模型、原理及算法**

**2.1 相关分析**

相关分析是研究现象之间是否存在某种依存关系，并对具体有依存关系的现象探讨其相关方向以及相关程度，是研究随机变量之间的相关关系的一种统计方法。根据各项指标的观测数据，由相关系数的计算公式



可计算得到各变量间的相关系数，从而根据相关系数来判断两因素之间的相关性是否显著。

**2.2 回归分析**

回归分析是应用极其广泛的数据分析方法之一，它提供了一套描述和分析变量间相关关系，揭示变量间的内在规律，并用于预测、控制等问题的行之有效的方法。在很多实际问题中，某个变量*Y*往往相关于另外一些变量*X*1, *X*2, …, *Xp*-1。一般情况下，我们认为*Y*的值由两个部分构成，一部分是由*X*1, *X*2, …, *Xp*-1能够决定的部分，它是*X*1, *X*2, …, *Xp*-1的某个函数，记为*f* (*X*1, *X*2, …, *Xp*-1)；另一部分是众多未加考虑的因素（包括随机因素）所产生的影响，被看作是随机误差，记为。于是*Y*与*X*1, *X*2, …, *Xp*-1的关系可表示为



特别，当*f* (*X*1, *X*2, …, *Xp*-1)是*X*1, *X*2, …, *Xp*-1的线性函数时，我们有



回归分析即是利用*Y*与*X*1, *X*2, …, *Xp*-1的观测数据，并在误差项的某些假定下确定。利用统计推断方法对所确定的函数的合理性以及由此关系所揭示的*Y*与*X*1, *X*2, …, *Xp*-1的关系作分析，进一步应用于预测、控制等问题。

对于回归参数*β*进行最小二乘估计，即选择*β*使误差项的平方和达到最小。

**2.3 主成分分析**

主成分分析（Principal Component Analysis, PCA）最早由*Pearson*于1901年发明，主要通过对协方差矩阵进行特征分析，达到在减少数据维数的同时保持数据集对方差贡献最大的目的（*Pearson*，1901）。也就是在数据信息损失最小的情况下，通过降维把原来的多个指标转化为一个或几个综合指标，即主成分，各主成分间互不相关。

具体计算过程为：

**step 1** 原始指标数据的标准化采集*p*维随机向量



构造样本阵，对样本阵元进行如下标准化变换：



其中，。从而得到标准化矩阵*Z*。

**step 2**对标准化阵*Z*求相关系数矩阵



其中。

**step 3**解样本相关矩阵*R*的特征方程



得*p*个特征根,确定主成分，按方差贡献率大于0.85确定主成分数量，使信息的利用率达85%以上，对每个特征根解特征方程得单位特征向量。

**step 4**将标准化后的指标变量转换为主成分



其中，*U*1称为第一主成分, *U*2称为第二主成分，…，*Up*称为第*p*主成分。

**step 5**对*m*个主成分进行综合评价,对*m*个主成分进行加权求和，即得最终评价值，权数为每个主成分的方差贡献率。

**3.实验数据收集及预处理**

从国家统计局[1]搜集2017年31个地区的居民人均可支配收入，财政一般预算支出，地区生产总值，居民消费水平，就业人员工资总额，地方财政税收收入、社会消费品零售总额以及商品零售价格指数的相关数据，并将它们统一整理到EXCEL表格中，利用SPSS软件分析它们与个人所得税税收收入之间的关系以及不同地区的个人所得税税收收入水平。

**4.基于统计方法的数据分析**

**4.1 相关分析**

用*x*1表示居民人均可支配收入，*x*2表示财政一般预算支出，*x*3表示地区生产总值，*x*4表示居民消费水平，*x*5表示就业人员工资总额，*x*6表示地方财政税收收入，*x*7表示社会消费品零售总额，*x*8表示商品零售价格指数，*y*表示个人所得税收入，使用SPSS软件对这八个因素之间的相关性进行分析[2]，得到如下结果：

**表1 各因素之间的相关性**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 相关性 | *x*1 | *x*2 | *x*3 | *x*4 | *x*5 | *x*6 | *x*7 | *x*8 | *y* |
| *x*1 | 1 | 0.383\* | 0.383\* | 0.407\* | 0.655\*\* | 0.702\*\* | 0.352 | -0.431\* | 0.384\* |
| *x*2 | 0.383\* | 1 | 1.000\*\* | 0.998\*\* | 0.909\*\* | 0.877\*\* | 0.921\*\* | -0.086 | 0.996\*\* |
| *x*3 | 0.383\* | 1.000\*\* | 1 | 0.998\*\* | 0.909\*\* | 0.877\*\* | 0.921\*\* | -0.086 | 0.996\*\* |
| *x*4 | 0.407\* | 0.998\*\* | 0.998\*\* | 1 | 0.916\*\* | 0.887\*\* | 0.930\*\* | -0.091 | 0.998\*\* |
| *x*5 | 0.655\*\* | 0.909\*\* | 0.909\*\* | 0.916\*\* | 1 | 0.977\*\* | 0.874\*\* | -0.255 | 0.913\*\* |
| *x*6 | 0.702\*\* | 0.877\*\* | 0.877\*\* | 0.877\*\* | 0.977\*\* | 1 | 0.848\*\* | -0.163 | 0.884\*\* |
| *x*7 | 0.352 | 0.921\*\* | 0.921\*\* | 0.930\*\* | 0.874\*\* | 0.848\*\* | 1 | 0.067 | 0.937\*\* |
| *x*8 | -0.431\* | -0.086 | -0.086 | -0.091 | -0.255 | -0.163 | -0.067 | 1 | -0.077 |
| *y* | 0.384\* | 0.996\*\* | 0.996\*\* | 0.998\*\* | 0.913\*\* | 0.884\*\* | 0.937\*\* | -0.077 | 1 |

其中\*表示在0.05级别（双尾），两因素之间的相关性显著；\*\*表示在0.01级别（双尾），两因素之间的相关性显著。

从表1可以看出，选取的八个因素中，只有商品零售价格和个人所得税税收收入的相关性是不显著的，其他七个因素和个人所得税税收收入的相关性都是显著的。

**4.2 回归分析**

**4.2.1 回归模型的建立**

以居民可支配收入、财政一般预算支出、地区生产总值、居民消费水平、就业人员工资总额、地方财政税收收入、社会消费品零售总额和商品零售价格指数作为自变量，个人所得税税收收入作为因变量，设置显著性水平为0.05，用SPSS软件进行线性回归分析[3]，得到相关系数R=0.999，接近1，因此可以初步判断该模型的拟合效果良好。

具体的回归结果如下所示：

**表2 系数表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 未标准化系数 | | 标准化系数 | t | 显著性 |
| B | 标准误差 | Beta |
| 常量 | 10137.255 | 6706.041 |  | 1.512 | 0.144 |
| 居民人均可支配收入 | -0.017 | 0.005 | -0.071 | -3.380 | 0.003 |
| 地区生产总值 | 0.031 | 0.138 | 0.036 | 0.223 | 0.825 |
| 居民消费水平 | 0.821 | 0.160 | 0.862 | 5.129 | 0.000 |
| 就业人员工资总额 | -0.051 | 0.059 | -0.067 | -0.877 | 0.389 |
| 地方财政税收收入 | 0.187 | 0.083 | 0.145 | 2.243 | 0.035 |
| 社会消费品零售总额 | 0.017 | 0.009 | 0.065 | 1.782 | 0.088 |
| 商品零售价格指数 | -95.872 | 66.013 | -0.024 | -1.452 | 0.160 |

从表2可以看出，进入回归方程的只有七个变量，财政一般预算支出这一变量由于对个人所得税税收收入没有显著影响，没有进入回归方程。除此之外，地区生产总值、就业人员工资总额以及商品零售价格指数这三个变量的显著性均大于0.05，表明系数并未通过显著性检验。

于是删去未通过显著性检验的四个变量，保留居民人均可支配收入、居民消费水平、地方财政税收收入、社会消费品零售总额这四个自变量，与个人所得税税收收入这一因变量建立回归模型，进行回归分析，发现社会消费品零售总额的显著性仍然未通过检验。所以删去社会消费品零售总额这一变量，对剩下的三个自变量与因变量建立回归模型，即



使用SPSS软件进行回归分析，得到如下结果：

**表3 系数及置信区间表**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 未标准化系数 | | 标准化系数 | t | 置信区间 | | 显著性 |
| B | 标准误差 | Beta | 下限 | 下限 |
| 常量 | 393.000 | 129.449 |  | 3.036 | 127.393 | 658.608 | 0.005 |
| 居民人均可支配收入 | -0.016 | 0.004 | -0.069 | -4.096 | -0.024 | -0.008 | 0.000 |
| 居民消费水平 | 0.889 | 0.025 | 0.933 | 35.762 | 0.838 | 0.939 | 0.000 |
| 地方财政税收收入 | 0.135 | 0.043 | 0.105 | 3.148 | 0.047 | 0.224 | 0.004 |

从表3可以看出，三个变量及常量都通过了显著性检验，且它们的置信区间都不包含零点，表明系数估计合理且通过了检验。于是得到回归方程：



**4.2.2 显著性检验**

对上述回归模型中的线性关系进行显著性检验，得到如下结果：

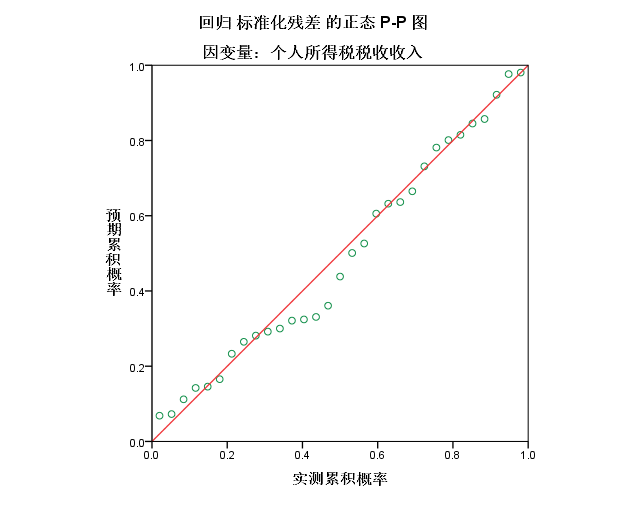
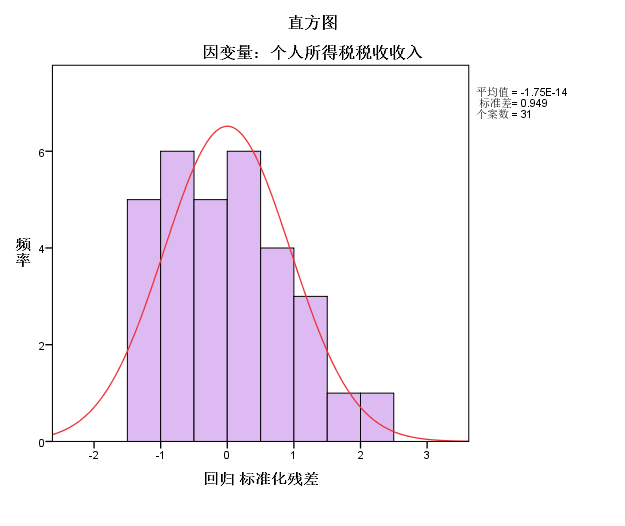
**表4 线性关系显著性检验表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 平方和 | 自由度 | 均方 | F | 显著性 |
| 回归 | 184849008.781 | 3 | 61616336.260 | 4015.472 | .000 |
| 残差 | 414307.679 | 27 | 15344.729 |  |  |
| 总计 | 185263316.460 | 30 |  |  |  |

从表2可知，显著性值为0.000，远小于0.05，即表明居民可支配收入、居民消费水平、地方财政税收收入与个人所得税税收收入之间建立的线性关系回归模型具有显著的统计学意义。

**4.2.3 残差分析**

使用SPSS软件对残差图进行正态性检验，结果如下图所示：



**图1 残差直方图 图2 残差正态P-P图**

首先，从直方图来看，残差大致服从正态分布；其次，从P-P图来看，点基本上都围绕 线的周围，于是可以认为残差是服从正态分布的。

**4.3 主成分分析**

**4.3.1 样本相关系数矩阵**

根据SPSS的输出结果可得到样本相关系数矩阵：

由于各变量的相关系数差距不大，所以直接从R出发做主成分分析。

**4.3.2 各主成分的贡献率**

根据SPSS的输出结果可得到各主成分的特征根和贡献率，如下表所示：

**表6各主成分的特征根和贡献率**

|  |  |  |  |
| --- | --- | --- | --- |
| 成分 | 初始特征值 | | |
| 总计 | 方差百分比 | 累积% |
| 1 | 5.923 | 74.037 | 74.037 |
| 2 | 1.352 | 16.905 | 90.942 |
| 3 | .559 | 6.983 | 97.925 |
| 4 | .094 | 1.171 | 99.096 |
| 5 | .060 | .756 | 99.852 |
| 6 | .010 | .123 | 99.976 |
| 7 | .002 | .024 | 100.000 |
| 8 | .000 | .000 | 100.000 |

由上表可知：第一主成分对月平均消费的贡献率为74.037%，第二主成分的贡献率为16.905%，第三主成分的贡献率为6.983%，第四主成分的贡献率为1.171%，第五主成分的贡献率为0.756%，第六主成分的贡献率为0.123%，第七主成分的贡献率为0.024%，第八主成分的贡献率为0.000%。，所以前两个主成分的累计贡献率为90.942%。

**4.3.3 第一和第二主成分的求解**

根据SPSS的输出结果可以得到第一和第二主成分的主成分载荷矩阵，如下表所示：

**表7 主成分载荷矩阵**

|  |  |  |
| --- | --- | --- |
|  | 成分 | |
| 1 | 2 |
| 居民人均可支配收入 | 0.580 | -0.650 |
| 财政一般预算支出 | 0.967 | 0.173 |
| 地区生产总值 | 0.967 | 0.173 |
| 居民消费水平 | 0.973 | 0.158 |
| 就业人员工资总额 | 0.980 | -0.124 |
| 地方财政税收收入 | 0.962 | -0.105 |
| 社会消费品零售总额 | 0.924 | 0.279 |
| 商品零售价格指数 | -0.173 | 0.860 |

将主成分的载荷向量除以各自主成分特征值的算术平方根，即得到主成分的特征向量 于是得到第一和第二主成分分别为

*F*1=0.098*x*1+0.163*x*2+0.163*x*3+0.164*x*4+0.165*x*5+0.162*x*6+0.156*x*7-0.029*x*8

*F*2=-0.481*x*1+0.128*x*2+0.128*x*3+0.117*x*4-0.091*x*5-0.077*x*6+0.206*x*7+0.636*x*8

当一个地区*F*1较大时，可以推断该地区的财政一般预算支出、地区生产总值、居民消费水平、就业人员工资总额、地方财政税收收入和社会消费品零售总额都较大或者其中一个较大，商品零售价格指数相对较小，所以第一主成分可以看作反映一个地区工资水平、消费水平、财政支出和收入水平的综合指标； 而在第二主成分*F*2的表达式中，*x*1和 *x*6对应的系数都为负数，当一个地区的*F*2较大时，可以推断该地区的商品零售价格指数较大，而居民人均可支配收入较小，所以第二主成分可以看作反映一个地区商品零售价格指数的指标。

**4.3.4 各地区个人所得税税收收入水平综合评分及排名**

根据公式



计算各地区个人所得税税收收入综合评分，可得到

**表8 各地区综合评分及排名**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 排名 | 地区 | 综合评分 | 排名 | 地区 | 综合评分 |
| 1 | 广东省 | 15417.4846 | 17 | 江西省 | 3829.2561 |
| 2 | 江苏省 | 11804.1734 | 18 | 黑龙江省 | 3780.5335 |
| 3 | 山东省 | 10949.5867 | 19 | 广西壮族自治区 | 3726.9281 |
| 4 | 浙江省 | 8498.4338 | 20 | 重庆市 | 3571.9547 |
| 5 | 河南省 | 7702.4312 | 21 | 内蒙古自治区 | 3429.4558 |
| 6 | 四川省 | 7518.3434 | 22 | 吉林省 | 3142.7258 |
| 7 | 湖北省 | 6536.1201 | 23 | 山西省 | 3057.4914 |
| 8 | 上海市 | 6389.2748 | 24 | 贵州省 | 3007.9379 |
| 9 | 北京市 | 6137.0575 | 25 | 新疆维吾尔族自治区 | 2774.1403 |
| 10 | 河北省 | 6090.3517 | 26 | 天津市 | 2685.0473 |
| 11 | 湖南省 | 6014.0889 | 27 | 甘肃省 | 2183.9265 |
| 12 | 安徽省 | 5020.6206 | 28 | 海南省 | 858.1217 |
| 13 | 辽宁省 | 4822.0407 | 29 | 西藏自治区 | 770.0343 |
| 14 | 福建省 | 4723.4276 | 30 | 青海省 | 751.4532 |
| 15 | 云南省 | 3892.9576 | 31 | 宁夏回族自治区 | 675.2934 |
| 16 | 陕西省 | 3866.8883 |  |  |  |

从表中可知，根据主成分分析得到的贡献率高达90.942%的两个主成分计算得到的综合得分，广东省位列第一，江苏省、山东省紧随其后。

**5.结果分析**

（1）相关分析中，在选取的可能与个人所得税税收收入相关的八个变量中，无论是在显著性水平为0.05还是显著性水平为0.01的情况下，财政一般预算支出、地区生产总值、居民消费水平、就业人员工资总额、地方财政税收收入、社会消费品零售总额和个人所得税税收收入的相关性都是显著的；居民人均可支配收入在显著性水平为0.05的情况下和个人所得税税收收入相关性显著；而商品零售价格和个人所得税税收收入之间的相关性是不显著的。

（2）回归分析中，居民可支配收入、居民消费水平、地方财政税收收入与个人所得税税收收入之间建立的线性关系回归模型具有显著的统计学意义。根据线性回归方程可知，居民可支配收入每增加1个单位，个人所得税税收收入减少0.016个单位；居民消费水平每增加1个单位，个人所得税税收收入增加0.889个单位；地方财政税收收入每增加1个单位，个人所得税税收收入增加0.135个单位。由此，我们可以知道，居民可支配收入的增加并不会引起个税收入的增加，而居民消费水平或地方财政税收收入如果得到了提高，个税收入会有相应的提升。

（3）主成分分析中，通过分析得到了两个主要指标，分别是反映一个地区工资水平、消费水平、财政支出和收入水平的综合指标和反映一个地区商品零售价格指数的指标。根据这两个指标及其各自的贡献率对各个地区的综合情况进行评分，可看出广东省、江苏省和山东省位列三甲，其他发展较快的地区，如上海市、北京市等排名也比较靠前；而发展比较落后的地区，如少数民族的自治区的排名则靠后，说明这些地区的工资水平、消费水平、收入水平都较低，商品零售价格指也较低。

**参考文献**

[1] 国家统计局. 2017年年度数据<http://data.stats.gov.cn/>.

[2] 刘龙青,张国庆.基于扎根理论-相关-主成分分析的农民发展权评价指标体系构[J/OL].

浙江农业科学,2018(09):1724-1728[2019-01-04]. <https://doi.org/10.16178/j.issn>.

0528-9017.20180964.

[3] 赵晓宇.基于回归分析的大连市公交客运量预测[J].智能城市,2017,3(09):27-28.

[4] 孙萍,陈诗怡.基于主成分分析法的营商政务环境评价研究——以辽宁省14市的调查数据为例[J/OL].东北大学学报(社会科学版),2019(01):1-5[2019-01-04]. <http://kns.cnki.net/kcms/detail/21.1413.G4.20190103.1535.008.html>.

**附录**

原始数据