基于岭回归地域差异下工资分配影响因素分析

宋丽娜 张文杰 王金良

（指导教师：张艳萍）

河北工程大学

一、研究背景

工资总额分配是与企业人力资源战略紧密联系的管理要素。企业的工资总额分配机制对企业的发展至关重要，它不仅影响员工的激励、调控、保障管理，而且有助于企业实现战略目标、改善经营绩效、提高市场竞争力和加强企业文化[1]。其中，国有企业工资总额管理是调节国家、企业、职工三者利益关系的重要方式，是深化国有企业工资决定机制改革的核心。2018 年 5 月，《关于改革国有企业工资决定机制的意见》发布，对国有企业工资总额分配提出了改革和完善意见。此次改革把国有企业工资分配管理权交给了企业，国有企业如何更好地进行工资总额分配是当前摆在国有企业面前突出的问题。对于集团化国有企业，其工资总额分配涉及集团本部和各个子分公司工资总额的核定，更具复杂性。集团化国有企业工资总额分配既要体现控制性，以符合国有企业工资总额预算管理的规定，又要体现激励性，以兼顾效率公平，实践中控制性和激励性往往存在一定的矛盾。如何建立一套科学、合理的工资总额分配方案，对国有企业来说是一个全新而重大的课题[2]。

二、研究目的

本文以某国有企业的 26 个省市分公司的工资总额分配情况为例，结合描述统计和推断统计方法，从公司间的地区差异、收入与成本规模、收益这三大方面的差异对工资总额分配所产生主要影响的因素进行分析；并根据确定的因素，通过建立统计模型对 2018 年初制定的省市分公司工资总额分配的合理性进行评价；最后，运用所建立的数学模型给出 2018 年各省市分公司工资总额合理分配方案。

三、数据来源

本文研究对象为企业的工资分配额，考虑到数据的代表性、准确性以及抽样过程的现实性，本文以某国有企业总公司 2018 年对 26 个省市分公司的工资总额分配为样本，选取地区差异、收入与成本规模、收益三个方面的指标对工资分配额主要影响因素进行分析，共 16 个指标，每个指标包含 26 个数据。

四、指标选取

选取合适的指标对接下来的分析研究具有十分重要的意义。由于影响工资分配额的因素有很多，具有随机性、模糊性和未知性的特点。针对某国有企业与不同省市分公司分配工资问题，通过与各省市地区构建差异相关联的收入与成本规模、收益等相关指标，建立一套科学合理的工资总额分配方案，并将其应用于大型连锁企业中，初步确定的解释变量如下：

（一）地区差异方面

地域人口（万人）X_1：分公司所在地域的人口数量；

城乡居民人均收入（元）X_2：城镇与农村居民年平均收入金额；

城乡人均消费支出（元）X_3：城镇与农村居民用于满足家庭日常生活消费的全部支出，包括购买实物支出和服务性消费支出；

城镇居民数（万人）X_4：分公司所在地区城镇居民数量；

城镇居民收入（万元）X_5：城镇居民在一定时期内创造的价值总和；

城镇居民人均可支配收入（元）X_6：城镇居民可用于最终消费支出和储蓄的总和，即居民可用于自由支配的收入。

GDP（万元）X_7：分公司所在地区所有常驻单位在一定时期内生产活动的最终成果；

其他国有企业在岗职工平均工资（元）X_8：分公司所在地区的其他国有企业在岗职工得到的劳动报酬的平均数；

商品房平均销售价格（元）X_9：分公司所在地区用于市场出售出租的房屋销售价格的平均数。

这些指标可以充分反映出各地区的差异情况，由于各解释变量间可能会有相关关系，在建模的过程中可能会产生多重共线性，因此会在后面对各变量进行详细分析。

（二）收入与成本规模方面

业务总收入（万元）X_{10}：分公司主营业务收入、其他业务收入、营业外收入、投资收益等收入总和；

业务总成本（万元）X_{11}：分公司为获取收入而提供产品或劳务所发生的生产费用的总和；

生产用固定资产（万元）X_{12}：用于物质生产或满足物质生产需要的固定资产。

（三）企业收益方面

净资产收益率 X_{13}：公司税后利润除以净资产得到的百分比；

人事成本费用率 X_{14}：人工成本总量与营业收入的比率；

劳动生产率 X_{15}：劳动者在一定时期内创造的劳动成果与其相适应的劳动消耗量的比值；

成本费用率 X_{16}：成本费用总额占营业收入的百分比。

五、问题分析

题中包含地区差异、收入与成本规模、收益这三大方面的数据，共有 16 个变量，现要分析这 16 个变量对因变量各地区工资总额分配的影响。初步考虑利用多元线性回归进行建模，首先需要了解变量间的相关性与其之间是否有多重共线性。选取与因变量高度相关的变量建模可以排除一些相关度较低的变量，从而达到降维的效果，同时也会使得建模的效果更好且精度更高。

（一）相关系数

相关系数是用来衡量两个数据集合是否在一条直线上，也是衡量定距变量间的线性关系，通常用 R 表示。具体计算公式如下：

$$R_{XY}=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\left(Y_i-\bar{Y}\right)^2}}$$

其中，$\bar{X}$ 和 $\bar{Y}$ 分别为 X_i 和 Y_i（i=1，2，…，n）的算数平均值，$R_{XY}\le 1$。

当 $0<R_{XY}<1$，称变量 X 和 Y 正相关；$R_{XY}>0.8$ 时称为高度相关，当 $R_{XY}<0.3$ 时称为低度相关，其它时候为中度相关；$-1<R_{XY}<0$，称变量 X 和 Y 负相关；且 $\left|R_{XY}\right|$ 越接近于 1，则说明变量 X 和变量 Y 之间的线性关系越显著。如果 $R_{XY}=0$，则称变量 X 和 Y 不（线性）相关。当 $R_{XY}=1$，则称变量 X 和 Y 完全（线性）相关。用 R 软件计算得到的相关系数见下表：

表 1　相关系数

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Y	1.000	0.841	0.493	0.435	0.936	0.943	0.608	0.942	0.384
X_1	0.841	1.000	0.078	-0.005	0.953	0.723	0.233	0.870	-0.035
X_2	0.493	0.078	1.000	0.983	0.292	0.658	0.944	0.433	0.890
X_3	0.435	-0.005	0.983	1.000	0.223	0.608	0.916	0.369	0.910
X_4	0.936	0.953	0.292	0.223	1.000	0.869	0.433	0.960	0.154
X_5	0.943	0.723	0.658	0.608	0.869	1.000	0.752	0.911	0.571
X_6	0.608	0.233	0.944	0.916	0.433	0.752	1.000	0.572	0.895
X_7	0.942	0.870	0.433	0.369	0.960	0.911	0.572	1.000	0.313
X_8	0.384	-0.035	0.890	0.910	0.154	0.571	0.895	0.313	1.000
X_9	0.357	-0.040	0.905	0.891	0.142	0.544	0.853	0.235	0.855
X_{10}	0.986	0.868	0.482	0.413	0.947	0.937	0.581	0.959	0.349
X_{11}	0.989	0.859	0.505	0.437	0.943	0.947	0.603	0.955	0.374
X_{12}	0.960	0.764	0.508	0.468	0.893	0.932	0.610	0.898	0.374
X_{13}	0.355	0.389	-0.026	-0.121	0.362	0.265	0.031	0.280	-0.140
X_{14}	-0.384	-0.481	-0.134	-0.019	-0.442	-0.314	-0.179	-0.390	0.086
X_{15}	0.639	0.443	0.603	0.535	0.543	0.654	0.653	0.631	0.520
X_{16}	-0.539	-0.527	-0.347	-0.241	-0.542	-0.473	-0.391	-0.502	-0.093

（续上表）

	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
Y	0.357	0.986	0.989	0.960	0.355	-0.384	0.639	-0.539
X_1	-0.040	0.868	0.859	0.764	0.389	-0.481	0.443	-0.527
X_2	0.905	0.482	0.505	0.508	-0.026	-0.134	0.603	-0.347
X_3	0.891	0.413	0.437	0.468	-0.121	-0.019	0.535	-0.241
X_4	0.142	0.947	0.943	0.893	0.362	-0.442	0.543	-0.542
X_5	0.544	0.937	0.947	0.932	0.265	-0.314	0.654	-0.473
X_6	0.853	0.581	0.603	0.610	0.031	-0.179	0.653	-0.391
X_7	0.235	0.959	0.955	0.898	0.280	-0.390	0.631	-0.502
X_8	0.855	0.349	0.374	0.374	-0.140	0.086	0.520	-0.093
X_9	1.000	0.336	0.364	0.395	0.069	0.113	0.524	0.287
X_{10}	0.336	1.000	0.999	0.936	0.383	0.449	0.682	0.576
X_{11}	0.364	0.999	1.000	0.942	0.363	0.432	0.673	0.565
X_{12}	0.395	0.936	0.942	1.00	0.347	0.339	0.563	0.499
X_{13}	0.069	0.383	0.363	0.347	1.000	0.698	0.521	0.632
X_{14}	-0.113	-0.449	-0.432	0.339	0.698	1.000	0.641	0.924
X_{15}	0.524	0.682	0.673	0.563	0.521	0.641	1.000	0.740
X_{16}	-0.287	-0.576	-0.565	0.499	0.632	0.924	0.740	1.000

由表 1 可知，Y 与 X_8、X_9、X_{13}、X_{14} 的相关系数均在 0.3 和 0.4 之间，相关性较小，说明 X_8、X_9、X_{13}、X_{14} 对分配的工资总额的影响很小。此外，分配的工资总额（Y）与地域人口（X_1）、城镇居民数（X_4）、城镇居民收入（X_5）、GDP（X_7）、业务总收入（X_{10}）、业务总成本（X_{11}）、生产用固定资产（X_{12}）的相关系数均在 0.8 以上，这几个自变量与 Y 高度线性相关，Y 与高度相关的自变量作多元线性回归分析较为合适。其余变量与分配的工资总额（Y）为中度相关。因此，根据相关性分析，就初步选取与分配的工资总额高度相关的变量作为自变量。

（二）VIF 值检验多重共线

解释变量之间完全不相关得情形是非常少见的，尤其是研究某个经济问题时，

涉及的自变量较多，我们很难找到一组自变量，他们之间不相关，而且又都对因变量有显著影响[2]。客观来说当某一经济现象涉及多个影响因素时，这些影响因素之间大多有一定关联性，当他们之间的相关性较弱时，我们一般就认为符合多元线性回归模型设计矩阵的要求，当这一组变量间的相关性较弱时，我们一般就认为符合多元线性回归模型基本假设的情形。

相关系数可以一方面检验说明多重共线性，但是较高的简单相关系数只是多重共线性存在的充分条件，而不是必要条件，特别是在多于两个解释变量的回归模型中，有时较低的简单相关系数也可能存在多重共线性。因此不能简单的依据相关系数进行多重共线性的准确判断。于是，我们接下来用方差膨胀因子来检验多重共线性。

对于多元线性回归模型来说，如果分别以每个解释变量为被解释变量，作与其他解释变量的回归，这称为辅助回归。以 X_j 为解释变量作对其他解释变量辅助线性回归的可决系数用 R_j^2 表示，则可以说明解释变量 X_j 参数估计值 $\hat{\beta}_j$ 的方差可以表示为

$$Var\left(\hat{\beta}_j\right)=\frac{\sigma^2}{\sum x_j^2}*\frac{1}{1-R_j^2}=\frac{\sigma^2}{\sum x_j^2}*VIF_j$$

式中 VIF_j 是变量 X_j 的方差膨胀因子，即 $VIF_j=\frac{1}{1-R_j^2}$.

由于 R_j^2 度量了 X_j 与其他解释变量的线性相关程度，这种程度越强，说明变量间的多重共线性越强，VIF_j 也就越大。反之，X_j 与其他解释变量的线性相关程度越弱，说明变量间的共线性越弱，VIF_j 也就越接近 1。由此可见 VIF_j 的大小反映了解释变量之间是否存在多重共线性，可用来衡量多重共线性的严重程度。经验表明，$VIF_j\geq 10$ 时，说明解释变量与其余解释变量之间有严重的多重共线性，且这种多重共线性可能会过度的影响最小二乘估计。

根据计算 Y 与 X_1, X_2，X_3，X_4，X_5，X_6，X_7，X_8，X_9，X_{10}，X_{11}，X_{12}，X_{13}，X_{14}，X_{15}，X_{16} 之间的关系，剔除了对因变量 Y 影响不大的一 X_2，X_3，X_6，X_8，X_9，X_{13}，X_{14}，X_{15}，X_{16} 这一部分变量，接下来利用方差膨胀因子法检验余下 X_1，X_4，

X_5，X_7，X_{10}，X_{11}，X_{12}这七个变量之间是否依然具有多重共线性，X_1，X_4，X_5，X_7，X_{10}，X_{11}，X_{12}多重共线性分析结果为：

表 2　多重共线性结果

	X_1	X_4	X_5	X_7	X_{10}	X_{11}	X_{12}
VIF	35.48502	87.44664	27.79626	27.34445	1055.05033	1200.91073	15.32144

由上述计算结果可知，$VIF_{10}=1055.05033$，$VIF_{10}=1200.91073$远远超过了10，说明X_{10}和X_{11}的*VIF*值非常大，出现了严重的多重共线性。X_{10}是业务总收入，X_{11}是业务总成本，两者的相关系数r_{12}=0.994，相关系数非常高，说明出现了严重的多重共线性，还可以说明，这两个自变量与其余自变量之间也可能存在多重共线性。另一方面，由于X_{11}和X_{12}之间存在严重的多重共线性，说明这两个变量我们可以选择其中一个作为对因变量产生影响的分析，这样也就消除了多重共线性对模型的影响。由此我们选择剔除*VIF*值最大的X_{11}，接下来再对剩余变量X_1，X_4，X_5，X_7，X_{10}，X_{12}这六个变量分析*VIF*值。

变量X_1，X_4，X_5，X_7，X_{10}，X_{12}的多重共线性结果：

表 3　删除 X_{11} 后的 VIF 值

	X_1	X_4	X_5	X_7	X_{10}	X_{12}
VIF	33.89171	85.25814	15.88755	25.59258	32.35977	14.22554

由以上分析可知 *VIF* 由明显降低，但是即使剔除了 X_{11}，VIF_j 的值依然还是大于 10，说明模型还是存在严重的多重共线性，无法直接对模型进行分析，如果直接进行分析，将有可能造成以下结果[3]：

（1）参数估计值得方差与协方差增大；

（2）对参数进行区间估计时，置信区间趋于变大；

（3）严重多重共线性时，假设检验容易做出错误的判断；

（4）当出现严重多重共线性时，可能造成可决系数较高，经 *F* 检验的参数联合显著性也很高，但对各个参数单独的 *t* 检验却可能不显著，甚至可能使估计的回归系数符号相反，甚至得出错误的结论。

基于以上分析，我们采用岭回归的方法消除多重共线性的影响建立合适的模型。

（三）岭回归估计

当设计矩阵 $\boldsymbol{X}$ 呈病态时其列向量之间有较强的线性相关性，即解释变量间出现严重的多重共线性。这种情况下，用普通的最小二乘法估计模型参数，往往参数估计方差太大，使普通最小二乘法的效果变得很不理想[3]。为解决这一问题，采用回归诊断和自变量选择来克服多重共线性的影响。

针对出现多重共线性时，普通最小二乘法明显变坏的问题，霍尔在 1962 年首次提出岭回归方法，用以控制与最小二乘估计相关的方差膨胀性和产生的不稳定性[4]。岭回归是在普通最小二乘法的参数估计 $\hat{\boldsymbol{\beta}}=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$ 中引入一个正常数矩阵 $k\boldsymbol{I}\left(k>0,\boldsymbol{I}\text{为单位矩阵}\right)$，得到 $\hat{\boldsymbol{\beta}}\left(k\right)=\left(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$，这里 $\hat{\boldsymbol{\beta}}\left(k\right)$ 为岭回归估计，k 为岭参数。因为岭参数 k 不是唯一确定的，所以得到的岭回归估计 $\hat{\boldsymbol{\beta}}\left(k\right)$ 实际是回归参数 $\boldsymbol{\beta}$ 的一个估计族。

理论岭回归模型：

$$Y=\beta_0+\beta_1\left(k\right)X_1+\cdots+\beta_n\left(k\right)X_n+\varepsilon$$

式中，Y 为因变量，β_0 为回归常数，$\beta_j\left(k\right)\left(j=1,2,\cdots n\right)$ 为岭回归系数，k 为岭参数，ε 表示其他随机因素的影响。

1. 岭迹法选择 k 值

在岭回归分析中，当岭参数 k 在 $\left(0,\infty\right)$ 内变化时，$\hat{\beta}_j\left(k\right)$ 是 k 的函数，在平面坐标系上把函数绘制出来，画出来的曲线称为岭迹。在实际应用中，可根据岭迹曲线的变化形状来确定适当的 k 值和进行自变量的选择[5]。

岭迹法的直观考虑是，如果最小二乘法看起来有不合理之处，如估计值以及正负号不符合经济意义[6]，则希望能通过采用适当的 $\hat{\boldsymbol{\beta}}\left(k\right)$ 岭估计来获得一定程度的改善，岭参数 k 值的选择就显得尤为重要。选择 k 值的一般原则是：

（1）各回归系数的岭估计基本稳定；

（2）用最小二乘估计的符号不合理的回归系数，其岭估计的符号变得合理；

（3）回归系数没有不合乎经济意义的绝对值；

（4）残差平方和增加不太多；

2. 岭回归选择变量

岭回归的一个重要应用是选择变量，通常的原则是：

在岭回归的计算中，假定设计矩阵 X 已经中心化和标准化，这样可以直接比较标准化岭回归系数的大小，我们可以剔除掉标准化岭回归系数比较稳定且绝对值很小的自变量。

当 k 值较小时，标准化岭回归系数的绝对值并不很小，但是不稳定。随着 k 的增大而迅速趋于零，像这样的岭回归系数不稳定、振动趋于零的自变量我们也可以予以剔除。

剔除标准化岭回归系数很不稳定的自变量，如果有若干个岭回归系数不稳定，究竟剔除几个变量，剔除哪几个变量，并无一般原则可循，需根据剔除某个变量后重新进行岭回归分析的效果来确定[7]。

（四）通过岭回归分析建立线性模型

1. 变量选择与模型建立

前面已经通过相关分析与方差扩大因子法剔除了一些变量，接下来结合回归系数的岭迹图、岭参数 k 值的选择，运用 R 软件确定恰当的岭参数，以及对剩下的自变量地域人口 X_1、城镇居民数 X_4、城镇居民收入 X_5、GDPX_7、业务总收入 X_{10}、生产用固定资产 X_{12} 进行选择与建模。

初始岭回归模型：

$$Y=\hat{\beta}_0+\hat{\beta}_1(k)X_1+\hat{\beta}_4(k)X_4+\hat{\beta}_5(k)X_5+\hat{\beta}_7(k)X_7+\hat{\beta}_{10}(k)X_{10}+\hat{\beta}_{12}(k)X_{12}$$

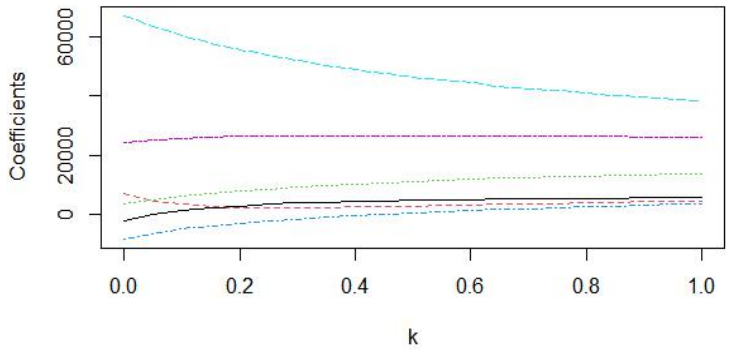

图 1　岭迹图

选择岭参数k从0到1，步长为0.05进行变化，结果如图1所示。从图中可以看到，当k取0时，变量地域人口X_1回归系数估计值为负，但随着k取值的增大，岭

回归系数$\hat{\beta}_1(k)$从负值变为正值，且在k=0.4时，两个变量岭回归系数估计值逐渐稳定。在k值逐渐增大的过程中，GDPX_7、业务总收入X_{10}岭回归系数估计值变化幅度较大，其中岭回归系数估计值变化幅度最小的为生产用固定资产X_{12}。整体来看，在k取0.2时，六个变量的岭回归系数估计值逐渐变得平缓。

为了更加具体的观察各变量岭回归系数估计值随k取值变化而变化情况，下面是选取0到1不同岭回归系数变化的各岭回归系数估计值矩阵。

表 4　选取不同岭回归系数的变化

k	X_1	X_4	X_5	X_7	X_{10}	X_{12}
0	-0.7492887	4.422369	0.000143279	-5.08E-05	0.2838503	0.1344025
0.05	0.0295712	2.967518	0.000202329	-3.91E-05	0.2683781	0.138854
0.1	0.5301605	2.203118	0.000248964	-3.07E-05	0.2557953	0.1415904
0.15	0.8770028	1.798685	0.00028759	-2.40E-05	0.2451292	0.1433613
0.2	1.1302012	1.598483	0.000320526	-1.84E-05	0.2358585	0.1445324
0.25	1.3223187	1.520685	0.000349161	-1.37E-05	0.2276655	0.1453045
0.3	1.4725218	1.518901	0.000374403	-9.58E-06	0.220338	0.1457979
0.35	1.5928172	1.565325	0.000396887	-5.90E-06	0.2137249	0.1460892
0.4	1.691102	1.642581	0.00041708	-2.57E-06	0.2077135	0.1462298
0.45	1.7727764	1.739457	0.000435336	4.66E-07	0.2022167	0.1462556
0.5	1.8416529	1.848542	0.000451931	3.26E-06	0.1971652	0.1461924
0.55	1.9004954	1.964849	0.000467089	5.84E-06	0.1925031	0.1460593
0.6	1.9513533	2.084975	0.000480989	8.25E-06	0.188184	0.1458709
0.65	1.9957752	2.206582	0.000493782	1.05E-05	0.1841692	0.1456382
0.7	2.0349511	2.328055	0.000505593	1.26E-05	0.1804261	0.1453701
0.75	2.0698087	2.448285	0.000516528	1.46E-05	0.1769267	0.1450734
0.8	2.1010802	2.566515	0.000526677	1.65E-05	0.1736469	0.1447537
0.85	2.1293502	2.682235	0.000536118	1.83E-05	0.1705659	0.1444154
0.9	2.1550903	2.795118	0.000544919	2.00E-05	0.1676656	0.1440624
0.95	2.178684	2.904962	0.000553139	2.16E-05	0.1649299	0.1436976
1	2.2004458	3.011659	0.00056083	2.31E-05	0.1623447	0.1433234

由上表可知，在k=0，即用最小二乘法进行多元线性回归时，变量地域人口X_1与GDP X_7的回归系数估计值符号为负，即X_1与X_7分别与Y呈负相关关系，亦为地域人口与GDP越大，地区工资总分配额度越小，这与实际是不相符合的[8]。从

岭回归的角度来看，两个变量只保留一个就可以了。

通过分析，我们决定剔除X_1，用Y与其余五个变量做岭回归，把岭参数步长改为0.02，范围减小到0.2。

剔除X_1后的岭回归模型：

$$Y=\hat{\beta}_0+\hat{\beta}_4(k)X_4+\hat{\beta}_5(k)X_5+\hat{\beta}_7(k)X_7+\hat{\beta}_{10}(k)X_{10}+\hat{\beta}_{12}(k)X_{12}$$

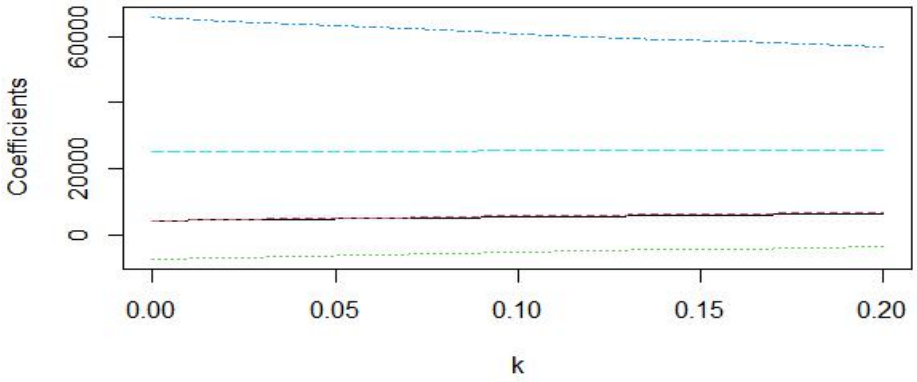

图 2　岭迹图

表 5　选取不同岭回归系数的变化

k	X_4	X_5	X_7	X_{10}	X_{12}
0	2.638773	0.000170246	-4.60E-05	0.2795795	0.1378703
0.02	2.803006	0.000183036	-4.32E-05	0.2750155	0.1382378
0.04	2.959897	0.000195299	-4.06E-05	0.2706521	0.1385704
0.06	3.109933	0.000207068	-3.81E-05	0.2664762	0.1388707
0.08	3.253561	0.000218372	-3.56E-05	0.2624762	0.1391413
0.1	3.391192	0.000229239	-3.32E-05	0.258641	0.1393843
0.12	3.523199	0.000239694	-3.10E-05	0.2549608	0.1396017
0.14	3.649927	0.000249759	-2.88E-05	0.2514263	0.1397955
0.16	3.771694	0.000259457	-2.67E-05	0.2480291	0.1399674
0.18	3.888791	0.000268808	-2.46E-05	0.2447612	0.1401188
0.2	4.001487	0.000277829	-2.26E-05	0.2416156	0.1402513

从表中可以看到，剔除 X_1 后岭回归系数变化幅度减小。虽然 X_7 仍为负值，但与剔除 X_1 前相比，X_7 负的程度已经降低。

从岭迹图看，岭参数 k 在 0.08-0.16 之间时，岭参数已经基本稳定，当 k=0.1 时，R^2=0.97568 仍然很大，因而在此区间取一个 k 值作岭回归能得到较好的结果，选取岭参数 k=0.1。给定 k=0.1，重新做岭回归。

2. 模型求解

剔除变量 X_1 且选定岭参数 k=0.1 后，软件输出结果显示，负相关系数 R=0.98777，决定系数 R^2=0.97568，由决定系数来看，岭回归方程高度显著。

表 6　模型结果

Name	Value
Mult R	0.98777
RSquare	0.97568
Adj RSqu	0.96960
SE	16262.24927

方程 F 检验 F=160.47477，显著性水平 $Sig.F$=0.00 远小于显著性水平 0.05，该岭回归在检验水平α=0.05 下的回归效果是高度显著，说明保留下来的变量城镇居民数 X_4、城镇居民收入 X_5、GDP X_7、业务总收入 X_{10}、生产用固定资产 X_{12} 整体上对各地区分配的工资总额 Y 有高度显著的线性关系。

表 7　方差分析表

	df	SS	MS	F value	Sig F
Regress	5	2.12E+011	4.24E+010	160.47	0.000
Residual	20	5.29E+009	264460751		

表 8　岭回归参数估计与检验

	B	SE(B)	Beta	B/SE(B)	sig
X4	8.203559	2.641073	.140943	3.106146	.005567
X5	.000610	.000176	.165252	3.472071	.002406
X7	.000058	.000025	.101530	2.290647	.032980
X10	.127113	.016795	.327285	7.568610	.000000
X12	.128966	.024352	.256865	5.295867	.000035
Constant	5903.275410	5816.314892	.000000	1.014951	.322251

回归系数的显著性检验。自变量 X_4、X_5、X_7、X_{10}、X_{12} 对 Y 均有显著影响，其中 X_7 GDP 的 P 值=0.03 最大，但仍在 0.05 的显著性水平上对 Y 高度显著。得到岭回归方程为：

$$Y=\hat{\beta}_0+\hat{\beta}_4(0.1)X_4+\hat{\beta}_5(0.1)X_5+\hat{\beta}_7(0.1)X_7+\hat{\beta}_{10}(0.1)X_{10}+\hat{\beta}_{12}(0.1)X_{12}$$

估计参数代入整理得：

$$\hat{Y}=5903.2757+8.2036X_4+0.00061X_5+0.000058X_7+0.1271X_{10}+0.1290X_{12}$$

从岭回归结果看来，模型拟合理想，结果符合实际。从回归系数看来，城镇居民数 X_4、城镇居民收入 X_5、GDP X_7、业务总收入 X_{10} 以及生产用固定资产 X_{12} 与分配的工资总额 Y 均呈正相关关系，说明城镇居民越多、城镇居民收入越高、GDP 越高、业务总成本越高、生产用固定资产越高，则分配的工资总额就越大。

六、总结与建议

（一）模型总结

根据选取的 16 个影响因素进行分析，由于数据存在严重的多重共线性，利用岭回归消除了多重共线性对模型的影响，最后分析可知，影响该国有企业 26 个分公司工资分配因素主要有城镇居民数、城镇居民收入、GDP、业务总收入和生产用固定资产对各省市工资分配产生的影响较大，其中以业务总收入影响最大。城镇居民数是城镇人口的数量，是指居住于城市、集镇的人口，主要依据人群的居住地和所从事的产业进行归类[9]。一个地区城镇居民数越多，在企业参与工作的员工人数相对就越多，企业的员工质量也会提高，相对的企业会在该地区投入更多且发展更好[10]。在模型中，自变量城镇居民数每增加一个单位，各地区工资的分配总额平均增加 8.2036 个单位。城镇居民收入对工资总额的影响为正向影响，当城镇居民收入每增加 1 个单位时，工资总额平均增加 0.1565252 个单位。城镇居民收入越多，对工资总额的影响越大。业务总收入对工资总额的影响为正向影响。当业务总收入增加 1 单位时，各分公司的工资总额平均增加 0.032980 个单位。说明该地区分公司业务总收入越大，工资总额越大。生产用固定资产对工资总额的影响为正向影响。当业务总收入增加 1 单位时，各分公司的工资总额平均增加 5.295867 个单位。GDP 每增加一个单位，各地区工资的分配总额平均增加 0.000058 个单位。

（二）研究建议

总的来说，城镇居民数、城镇居民收入、GDP、业务总收入和生产用固定资产对各省市工资分配产生的影响较大，其中以业务总收入影响最大，该企业在制定分配方案时，应优先考虑该地区的业务总收入，考虑该分公司的业务水平，参考相关的财务报表进行工资分配。再分别考虑生产用固定资产、城镇居民数、城镇居民收入和 GDP 等因素，以此保证分配的合理性，促进公司的综合发展。另一方面，由于其他因素对各分公司的工资总额分配影响较小，所以其他影响因素可以选择考虑或不予考虑。

参考文献

[1] 李康康,梁锦钰,吴蒙,等.地域差异下工资分配额度定量表征模型研究[J].华北理工大学学报,2020,42(1):93-97.

[2] 周利,高栓喜,白思俊.股价主要影响因素的统计分析[J].河南大学学报(自然科学版),2001,31(4):41-46.

[3] 马利芸.我国寿险需求影响因素的岭回归分析[J].现代商贸工业,2019,40(05):117-119.

[4] 何晓群，应用回归分析[D]，中国人民出版社,2015,173-205.

[5] 胡良平.岭回归分析[J].四川精神卫生,2018,31(03):193-196.

[6] 王飞,孙嘉聪,沈丹.多重共线性问题的岭回归实例[J].数学学习与研究,2019(20):132-134.

[7] 王锐.岭回归分析在解决经济数据共线性问题中的应用[J].经济研究导刊,2018(22):144-147.

[8] 庞浩.计量经济学[D].科学出版社,2014,95-109.

[9] 刘悦.我国行业收入差距的影响因素分析[D].北京:北京交通大学，2014.

[10] 吴万勤,黄陈津.云南省服务业工资收入的主要影响因素分析[J].云南民族大学学报,2020,29(3):225-231.

附录：本案例所使用的 R 软件程序命令

相关系数检验

```
> csvpath<-file.choose()
> csvpath
> ad<-read.csv(csvpath,header = T)
> View(ad)
> cor(ad）
```

VIF 检验

```
> library(car)
> csvpath<-file.choose()
> ex1<-read.csv(csvpath,header=T)
>    vif.dia<-vif(lm(Y~.,data=ex1))
> vif.dia
> csvpath<-file.choose()
> ex2<-read.csv(csvpath,header=T)
> vif.dia<-vif(lm(Y~.,data=ex2))
> vif.dia
```

岭回归模型

第一步

```
csvpath1<-file.choose()
csvpath1
ad1<-read.csv(csvpath1,header=T)
library(MASS)
ridge.dia<-lm.ridge(Y~.,ad1,lambda=seq(0,1,0.05))
beta<-coef(ridge.dia)
beta
k<-ridge.dia$lambda
 matplot(ridge.dia$lambda,t(ridge.dia$coef),xlab = expression(k),ylab =
```

"Coefficients",type = "l",lty = 1:20)

第二步

```
csvpath2<-file.choose()
csvpath2
ad2<-read.csv(csvpath,header=T)
library(MASS)
ridge.dia2<-lm.ridge(Y~.,ad,lambda=seq(0,0.2,0.02))
beta2<-coef(ridge.dia)
beta2
k2<-ridge.dia2$lambda
  matplot(ridge.dia2$lambda,t(ridge.dia2$coef),xlab = expression(k),ylab =
"Coefficients",type = "l",lty = 1:20)
```