基于 PCA 和多元线性回归的居民消费水平影响因素分析

陈鹏丞 聂再冉 白玉景

（指导教师：张艳萍）

河北工程大学

一、研究背景概述

早在 1936 年，凯恩斯在《通论》中就提出了绝对收入理论假说，随后国内外便开始了对消费水平进行相关研究。迄今为止，学术界主要存在两种不同的理论观点：一种是凯恩斯主义的消费函数，其强调现期消费主要取决于现期收入，随着可支配收入增加，消费也增加。另一种是面向未来的消费函数，强调消费对一生总财富的依赖，以及储蓄在稳定消费中的作用。王吉恒等(2012)在《论我国居民消费水平的影响因素》一文中指出国民生产总值、居民收入和储蓄、通货膨胀和社会保障等都是影响我国居民消费水平的重要因素，并采用层次分析法进行分析[1]。而刘慧敏(2014)在《我国居民消费水平影响因素的实证分析》中认为人均国内生产总值、居民可支配收入、物价水平等是影响居民消费水平的主要原因，并分别运用线性回归模型和非线性模型进行实证研究[2]。王韵如(2016)在《影响我国居民消费水平的因素分析》中通过对国内生产总值、人均可支配收入、城镇人口数建立多元线性回归模型，分析研究影响我国居民消费水平的主要因素[3]。李俊华(2017)在《农村居民消费水平影响因素的回归分析——以湖北省为例》从农村人均纯收入、人均 GDP、城市化率、城乡居民消费差额等 4 个方面构建对数变量，同时引入了城镇化与农村居民家庭人均纯收入的交互项，构建数学模型，对农村居民人均消费的影响进行了分析[4]。郎颖臻(2018)在《多元线性回归应用——居民消费影响因素分析》以若干影响消费水平的因素居民人均可支配收入、人均地区生产总值、居民消费水平（消费价格指数）、地区的失业率等为自变量，建立多元线性回归模型，分析了解影响居民消费的关键因素[5]。林开思(2018)在《我国农村居民消费影响因素分析》中国运用最小二乘法对农村居民消费水平、农村居民人均收入、人均国内生产总值以及农村居民消费者价格指数进行分析，探究影响我国农村居民消费的主要因素[6]。王一睿(2019)在《陕西省居民消费水平影响因素分析》选取了居民人均可支配收入、居民储蓄率、产业增加值等。先

通过相关性分析，判断潜在的影响因素与被解释变量之间的相关程度，对于相关性高的变量则进一步建立回归模型并进行实证研究[7]。

通过相关文献阅读可以发现，国内学者对居民生活水平的研究中选取的指标综合考虑多方面因素。在已有研究的基础上，本文选取基于居民消费总额、物价指数、进口总额、职工工资、储蓄增额以及人均 GDP 的指标体系进行研究。

二、研究方法

（一）研究思路

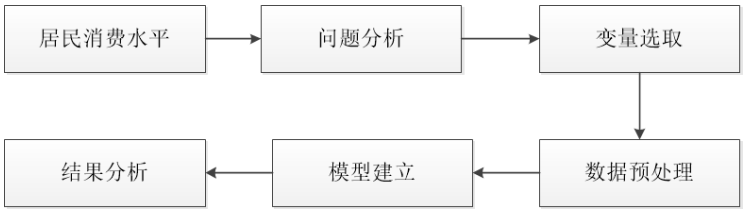

图 1　研究思路框架图

（二）指标体系的建立

表 1　变量解释

变量	变量名称	变量解释
Y	消费总额	居民消费水平绝对数，反映我国居民每年平均用在消费上所花费的绝对数额，对消费水平的反应较为明显。
X_1	物价指数	反映城乡商品价格变动趋势的一种经济指数。物价的调整变动直接影响到城乡居民的生活支出和国家的财政收入，影响居民购买力和市场供需平衡，影响消费与积累的比例。
X_2	进口总额	进口总额指实际进口我国国境的货物总金额，居民的消费不仅只考虑国内生产的商品，同时也会考虑国外商品，因此进口会在一定程度上影响我国居民的消费水平。
X_3	职工工资	职工平均工资，居民收入的多少影响着人民手中可支配收入，进而影响着居民在消费上所花费的具体数额。
X_4	储蓄增额	居民年均储蓄增加额，人均收入的两部分应用分别在储蓄和消费。
X_5	人均 GDP	人均国内生产总值，是反映某一国家全部生产活动最终成果的重要指标。

因为每个国家的国情不一样，因此在进行指标选取的时候通常参考本国研究现状，本文在查阅大量文献后遵循可获得性、可比性、动态性原则，选取物价指数、进口总额、职工工资、储蓄增额以及人均 GDP 作为自变量，选取居民消费总额作为因变量。具体变量解释见表 1。

三、数据预处理

本文研究居民消费水平中的数据主要来自于《中国统计年鉴》、《中国经济统计年鉴》。

（一）缺失值填补

收集数据后发现 2015 年的数据中缺少储蓄增额的数据，考虑到本文中自变量的相关性较高，所以采用线性回归法对缺失值进行填补，填补结果如表 2 所示。

表 2　缺失值填补

消费总额	物价指数	进口总额	职工工资	储蓄增额	人均 GDP
19308	615.2	104336.1	63241	51897	49992

（二）异常值检验

考虑到异常值对于整体分析过程的影响，因此采用箱线图进行检验。箱线图是利用数据中的五个统计量：最小值、第一四分位数、中位数、第三四分位数与最大值来描述数据的一种方法，它也可以简明扼要的看出数据的分布以及异常值点。

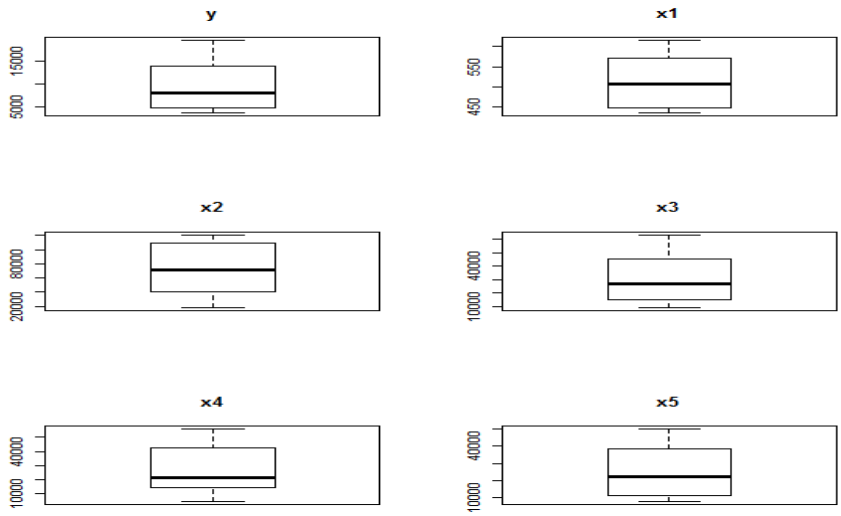

图 2　各变量的箱线图

从图 2 中可以看出数据中不存在异常值，无需进行异常值的处理。

（三）数据标准化处理

数据标准化的处理方法有 Z-Score 值标准化、最小最大值标准化、归一化标准化等方法，针对经济类数据，本文采用常用的 Z-Score 标准化方法，对原始数据进行标准化处理。

Z-Score 值标准化方法计算公式：

$$X_i^{'} = \frac{(X_i - \bar{X})}{\sigma}$$

（注：$\bar{x}$ 表示均值，σ 表示标准差）

四、相关性分析

在进行回归分析之前，首先对各变量进行相关性分析，相关性分析目的在于先检验一下众多的自变量和因变量之间是否存在相关性，如果相关分析时各自变量跟因变量之间没有相关性，就没有必要再做回归分析。

表 3　变量间相关系数矩阵

	Y	X_1	X_2	X_3	X_4	X_5
Y	1.00000	0.98890	0.92549	0.99809	0.84695	0.997
X_1	0.98890	1.00000	0.95981	0.99249	0.88522	0.996
X_2	0.92549	0.95981	1.00000	0.93554	0.85971	0.948
X_3	0.99809	0.99249	0.93554	1.00000	0.86528	0.998
X_4	0.84695	0.88522	0.85971	0.86528	1.00000	0.868
X_5	0.99700	0.99629	0.94829	0.99842	0.86838	1.000

根据表 3 可以得到，消费总额（Y）与物价指数（ X_1 ）、进口总额（ X_2 ）、职工工资（ X_3 ）、储蓄增额（ X_4 ）以及人均 GDP（ X_5 ）的相关系数分别为 0.98890、0.92549、0.99809、0.84695、0.997，关系都非常密切（ $r>0.8, P<0.001$ ），其中消费总额与职工工资之间的关系最为密切，其相关系数达到 0.998，说明相关性最高，即相关程度最大。

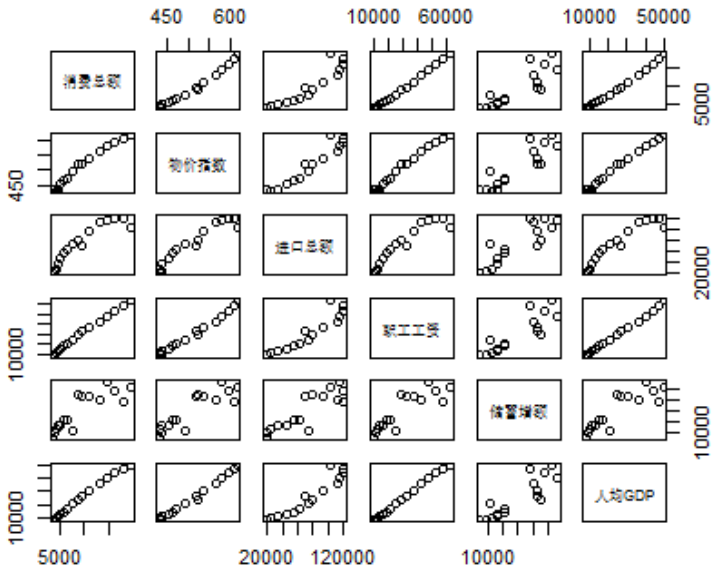

图 3　矩阵散点图

通过图 3 可以发现，除了储蓄增额这个变量外，其余变量之间都呈现出较为明显的线性关系，所以适合选取多元线性回归模型。但各变量之间都存在线性相

关关系，这可能会导致多重共线性问题。另外，储蓄增额与消费总额之间虽然呈现并不明显的线性趋势，但在实际情况下，两者存在着一定的负相关性，所以不能忽略储蓄增额对消费总额的影响。

五、模型建立

（一）主成分分析

考虑到选取的自变量相对较多，可能存在一定的共线性，而主成分分析可以对变量进行降维，而且得到的新变量之间彼此不相关，因此采用主成分分析法对数据初步建模。

以 2002-2017 年我国居民消费水平影响因素的相关数据为基础，对数据进行标准化处理，然后采用主成分分析法，对影响居民消费水平的因素进行分析。主成分载荷表如表 4 所示，结合图 4 的碎石图可以发现应该选择的主成分个数为 1，但结合实际情况，储蓄增额与其它变量之间的差别较明显，所以确定主成分个数为 2，计算主成分得分结果如表 5 所示。

表 4 主成分载荷表

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
物价指数	0.458	-0.176	-0.116	0.836	-0.219
进口总额	0.445	-0.149	0.866	-0.166	
职工工资	0.454	-0.258	-0.401	-0.505	-0.558
储蓄增额	0.423	0.902			
人均 GDP	0.456	-0.257	-0.266	-0.131	0.799

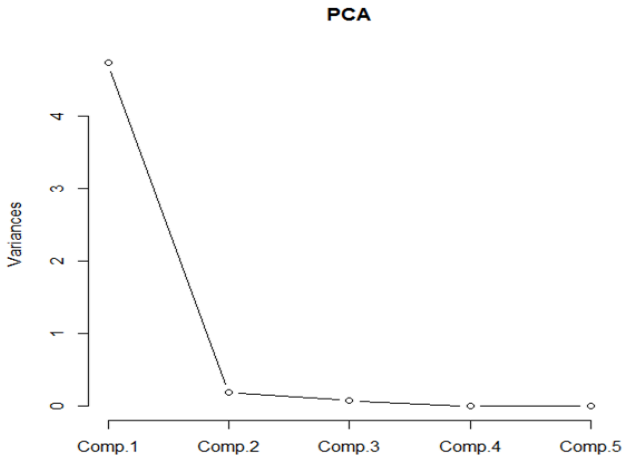

图 4 碎石图

表 5 主成分得分

Comp.1	Comp.2
-2.9637851	-0.29602314
-2.7388092	-0.09199686
-2.5494521	0.06315056
-2.2192557	0.15483486
-1.8516167	-0.04845729
-1.4310607	0.12726884
-1.1447648	-0.06462654
-0.8765394	-0.81573043
0.5129670	0.80911931
0.4477888	0.64077322
1.1548923	0.31365899
1.8560544	-0.14297238
2.6294501	0.50979724
2.8671652	-0.12583183
2.9200341	-0.85398581
3.3869320	-0.17897874

则主成分模型的表达式为：

$$F_1 = 0.458X_1 + 0.445X_2 + 0.454X_3 + 0.423X_4 + 0.456X_5$$

$$F_2 = -0.176X_1 - 0.149X_2 - 0.258X_3 + 0.902X_4 - 0.257X_5$$

F_1为主成分 1，F_2为主成分 2，经过以上的主成分提取，可以尽可能多地保留原始变量的信息，且变量间彼此间互不相关，为进一步的线性回归分析做铺垫。

（二）多元线性回归分析

通过主成分分析得到两个可以概括原始变量信息的主成分，将 2002-2017 年的居民消费总额作为因变量，两个主成分值作为自变量，利用 R 软件进行多元线性回归分析，进而得出预测模型。

表 6 模型汇总

模型	R	R 方	调整 R 方	标准估计误差	更改统计量				
					R 方更改	F 更改	df1	df2	Sig.F 更改
1	.875[a]	.766	.749	.500866	.766	45.793	1	14	0.00
2	.99[b]	.981	.977	.150996	.981	322.452	2	13	0.00

表 6 是逐步筛选的过程，相关系数 R 和判定系数的R^2值逐渐增加，自变量的影响也在不断增加。R^2值越大，因变量与自变量的共同变量比例较高，拟合度也越好。表 6 中R^2的值最终达到 0.99，说明模型与数据的拟合程度较高。

表 7　方差分析表（F 检验）

模型		平方和	df	均方	F	Sig.
1	回归	11.488	1	11.488	45.793	0.000[a]
	残差	3.512	14	0.251		
	总计	15	15			
2	回归	14.704	2	7.352	322.452	0.000[b]
	残差	0.296	13	0.023		
	总计	15	15			

a.预测变量：(常量),F1

b.预测变量：(常量),F2

因变量:Zscore(居民消费总额)

表 7 是对变量所建立的线性方程的 F 检验，从表中可以看出模型 1 和模型 2 中回归方差都明显大于残差，F 值分别是 45.793、322.452，这时的显著水平均为 0 小于 0.05。F 值检验证明被解释变量和解释变量之间的线性关系显著，证明建立线性回归模型合理。

表 8　系数输出表（t 检验）

		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准误差				容差	VIF
1	常量	-4.424E-17	.125		.000	1.000		
	主成分 1	.875	.129	.875	6.767	.000	1.000	1.000
2	常量	-6.994E-17	.038		.000	1.000		
	主成分 1	.875	.039	.875	22.447	.000	1.000	1.000
	主成分 2	.463	.039	.463	11.876	.000	1.000	1.000

表 8 输出的是各自变量的系数检验结果，表中回归系数的显著性检验值都为 0，小于显著水平 0.05。常量值的显著水平检验为 1，表明其无限趋近于 0，故可将其舍去。表中自变量 F1、F2 的容差与膨胀系数都为 1，且都小于 5，所以自变量之间没有出现共线性。通过检验可以证明自变量的选择对于居民消费水平的影响较为直接，模型建立合理。根据表 8 输出结果，可以得到多元线性回归方程为：

$$Y = 0.875X_{F_1} - 0.463X_{F_2}$$

最终得到的回归方程为：

$$Y = 0.482X_1 + 0.458X_2 + 0.517X_3 - 0.048X_4 + 0.518X_5$$

六、结论及政策建议

文章采用理论与实证相结合，定性与定量相结合的方法，选取 2002-2017 年的数据，对我国居民消费水平的变化进行时间上的动态分析。由 PCA 及多元线性回归模型的论证可得，职工工资与居民储蓄增额这两项指标对我国居民消费水平有重要影响。

其中居民消费水平与职工工资呈正相关关系，即随着职工工资的上调居民的消费水平会不断上升。居民消费水平与储蓄增额之间呈负相关关系，即随着储蓄增额的增加，居民消费水平会有下降的趋势。所以，要想提高居民消费水平不能仅仅依靠社会生产力的进步人民工资水平的提高，更进一步是要改变传统的思维模式，要把资产运作起来，从而刺激消费。

针对以上结论给出具体建议如下：

（一）提高职工工资水平

要坚持以人民为中心的发展思想，提高要素分配中劳动要素的分配比例，深化改革开放大力发展国民经济，提高居民整体收入水平，缩小城乡收入差距。由于我国还处于社会主义初级阶段，农村人口基数庞大，只有调整和优化农业结构、振兴乡镇企业、加大对农业的投入、改良和完善农村的市场环境才能提高农民的收入，增强农民的购买力[8]。

（二）降低储蓄额刺激消费

存钱是中国人的传统思想，所以要想提高居民消费水平必须降低储蓄增额，改变居民的消费思想倡导合理、科学的消费结构和模式。与此同时政府要鼓励居民改变传统的消费习惯向精神文化消费转变，提高居民的消费档次，从而促进居民消费水平的提高。

国家发展的前提必然是经济的发展，然而经济发展最大的动力之一就是消费，只有消费水平的提高才能拉动经济的发展，不能走只生产不消费的道路，只有二者协同发展才能使国家变的富强[9-10]。

参考文献

[1] 王吉恒，李敏，孟菲.论我国居民消费水平的影响因素[J].开放导报, 2012(2).

[2] 王华丽.多元线性回归分析实例分析[J].科技资讯, 2014(29).

[3] 王韵如.影响我国居民消费水平的因素分析[J].商业研究, 2016(5).

[4] 李俊华，袁力.农村居民消费水平影响因素的回归分析——以湖北省为例[J].汉江师范学院学报, 2017.12(6).

[5] 郎颖臻.多元线性回归应用——居民消费影响因素分析[J].社会发展, 2018.04(100).

[6] 林开思.我国农村居民消费影响因素分析[J].现代商贸工业, 2018(24).

[7] 王一睿.陕西省居民消费水平影响因素分析[J].现代商贸工业, 2019(6).

[8] 宋少青.中国农村居民消费水平影响因素分析[J].河北企业, 2017(12):42-43.

[9] 俞琴.居民消费水平影响因素分析[J].当代经济, 2017(19):151-153.

[10] 冯垚.影响居民消费水平的主要因素分析——以湖南省为例[J].统计与管理, 2017(03):57-58.

附录：本案例所使用的 R 软件程序命令(部分)

```
#缺失值填补前的数据
shuju0=read.csv("C:/Users/admin/Desktop/data.csv",header=T,na.strings = NA)#导
    入数据
shuju1=shuju0[,-1]
shuju1
sum(is.na(shuju1))#计算缺失值数目
sum(complete.cases(shuju1))#计算完整样本数量
md.pattern(shuju1)
sub=which(is.na(shuju1[,5]) == TRUE)#返回 shuju 数据集中第 5 列为 NA 的行
dataTR=shuju1[-sub,]#将第 5 列不为 NA 的数据存数 dataTR 中
dataTE=shuju1[sub,]#将第 5 列为 NA 的数据存入 dataTE
lm=lm(储蓄增额~消费总额+物价指数+进口总额+职工工资+人均 GDP,data=data
    TR)#利用 dataTR 中储蓄增额为因变量,其余为自变量，构建线性回归模型 l
    m
dataTE[,5]=round(predict(lm,dataTE))#按模型 lm 对 dataTE 中的缺失值进行预测
dataTE
#缺失值填补后的数据
shu=read.csv("C:/Users/admin/Desktop/data2.csv",header=T,na.strings = NA)#导入
    数据
shuju2=shu[,-1]
shuju2
y=shuju2$消费总额#变量替换
x1=shuju2$物价指数#变量替换
x2=shuju2$进口总额#变量替换
x3=shuju2$职工工资#变量替换
x4=shuju2$储蓄增额#变量替换
x5=shuju2$人均 GDP#变量替换
par(mfrow=c(3,2))
boxplot(shuju2$消费总额,main="y")
boxplot(shuju2$物价指数,main="x1")
boxplot(shuju2$进口总额,main="x2")
```

```
boxplot(shuju2$职工工资,main="x3")
boxplot(shuju2$储蓄增额,main="x4")
boxplot(shuju2$人均 GDP,main="x5")
par(mfrow=c(1,1))#异常值检验
summary(shuju2)#描述数据
Mean=sapply(shuju2,mean)#平均值
Min=sapply(shuju2,min)
Median=sapply(shuju2,median)
Max=sapply(shuju2,max)
SD=sapply(shuju2,sd)
cbind(Mean,Min,Median,Max,SD)
cor(shuju2)#相关分析
pairs(shuju2)#矩阵散点图
shuju3=scale(shuju2,center = T,scale = T)#数据标准化处理
shuju3
#标准化后的数据
shu=read.csv("C:/Users/admin/Desktop/多元统计分析（期末）/数据/附件 3.csv",he
    ader=T,na.strings = NA)#导入数据
shuju2=shu[,-1]
shuju2
Mean=sapply(shuju2,mean)#平均值
Min=sapply(shuju2,min)
Median=sapply(shuju2,median)
Max=sapply(shuju2,max)
SD=sapply(shuju2,sd)
cbind(Mean,Min,Median,Max,SD)
cor(shuju2)#相关分析
pairs(shuju2)#矩阵散点图
fm=lm(y~x1+x2+x3+x4+x5,data=shuju)#显示多元线性回归模型
summary(fm)#全模型分析
par(mfrow=c(2,2))
plot(fm,which = c(1:4))
anova(fm)#模型的方差分析
```

```
library(car)
vif(fm)#多重共线性检验
PCA=princomp(shuju[,-1],cor=T)#主成分分析
PCA#特征根开根号结果
summary(PCA)
PCA$loadings#主成份载荷
screeplot(PCA,type = "lines")#确定主成分
C=PCA$scores[,1:2]#主成分得分
C
CC=data.frame(C)#矩阵转换
CC
C1=CC$Comp.1
C2=CC$Comp.2
xinshuju=cbind(C1,C2,y)
xinshuju=data.frame(xinshuju)
xinshuju
fm1=lm(y~C1+C2,data=xinshuju)#显示新的多元线性回归模型
summary(fm1)#新模型分析
par(mfrow=c(2,2))
plot(fm1,which = c(1:4))
vif(fm1)#多重共线性检验
anova(fm)#模型的方差分析
library(leaps)
DW1=dwtest(fm1)#自相关性检验
DW1
fm2=lm(y~x2+x4+x5,data=shuju)
vif(fm2)
vif(fm1)
data.frame(result$outmat,BIC=result$bic)
shuju
fm=lm(y~x1+x2+x3+x4+x5,data=shuju)
fm1=lm(y~C1+C2,data=xinshuju)
fm2=lm(y~x2+x4+x5,data=shuju)
```

```
AIC(fm,fm1,fm2)
BIC(fm,fm1,fm2)
y1=predict(fm,shuju)
y2=predict(fm1,xinshuju)
y3=predict(fm2,shuju)
y0=shuju[,1]
yz=cbind(y0,y1,y2,y3)
yz
r1=y0-y1
r2=y0-y2
r3=y0-y3
resid=abs(as.data.frame(cbind(r1,r2,r3)))
sapply(resid,mean)
```