基于因子分析的二元 Logistic 回归对移动电话普及的研究

周世龄

黑河学院 理学院，四川 宜宾

摘　要

本研究采用统计分析方法，探讨影响中国居民家庭是否拥有移动电话的主要因素。通过因子分析简化变量，利用逻辑回归模型分析居住地、家庭规模、住房占用状态、全国人均支出五等分和总年度家庭支出等因素对家庭是否拥有移动电话的影响。研究发现，经济水平因子对拥有手机的概率影响最大，家庭规模因子次之，居住条件因子影响较小。模型预测准确率达到 74.5%，表明全国人均支出五等分和总年度家庭支出是影响家庭是否拥有移动电话的关键因素。

关键词

移动电话普及；因子分析；二元 Logistic 回归；经济水平；家庭规模。

Analysis of Mobile Phone Popularity Based on Factor Analysis and Binary Logistic Regression

Shiling Zhou

School of Science, Heihe University, Yibin, Sichuan, China

Abstract

This study employs statistical analysis to explore the primary factors influencing the prevalence of mobile phones in Chinese households. Utilizing factor analysis to reduce variable dimensions, a logistic regression model was applied to examine the impact of factors such as place of residence, family size, housing occupancy status, quintile of national per capita expenditure, and total annual household expenditure on the likelihood of household mobile phone ownership. The findings indicate that the economic status factor has the greatest influence on the probability of owning a mobile phone, followed by the family size factor, with the residential conditions factor having a relatively minor impact. The model achieved a predictive accuracy of 74.5%, suggesting that the quintile of national per capita expenditure and total annual household expenditure are key determinants of mobile phone ownership.

Keywords

Phone Popularity; Factor Analysis; Binary Logistic Regression; Economic Status; Family Size.

1、绪论

随着移动通信技术的迅猛发展，移动电话在日常生活中的普及程度不断提高。移动电话的拥有率不仅反映了居民的通讯需求，还在一定程度上反映了家庭经济水平和生活质量。研究居民家庭拥有移动电话的决策因素，对于理解移动通信技术的普及程度及其背后的社会经济影响具有重要意义。

本研究旨在通过统计分析方法，探讨影响中国居民家庭是否拥有移动电话的主要因素。具体来说，我们将从居住地（城市/农村）、家庭规模、住房占用状态（自有/租赁）、全国人均支出五等分、总年度家庭支出等五个自变量出发，通过因子分析简化变量，并进一步利用逻辑回归分析确定这些因素对家庭是否拥有移动电话这一因变量的影响程度。

收集和整理中国居民家庭的相关数据，对数据进行预处理和初步描述性统计分析，以了解数据的基本特征。通过因子分析方法，将多个相关变量简化成少数几个因子，减少变量间的冗余信息，为后续的回归分析做准备。利用逻辑回归分析方法，探讨居住地、家庭规模、住房占用状态、全国人均支出五等分和总年度家庭支出等因素对家庭是否拥有移动电话的影响。对逻辑回归分析的结果进行解释，指出各因素的显著性和影响方向，并与现有文献进行对比和讨论。

通过因子分析简化指标，提取主要成分，减少数据维度的同时提高分析的准确性。利用逻辑回归分析模型，探究各独立变量对因变量的影响，量化各因素的作用大小和显著性水平。

通过本研究，我们希望能够明确影响家庭是否拥有移动电话的主要社会经济因素，并量化这些因素的影响程度，为政策制定者和研究人员提供有价值的参考。

本研究的创新之处在于综合运用因子分析和逻辑回归分析方法，从多个角度探讨影响移动电话普及的决定因素，为未来相关领域的研究提供新的视角和方法参考。

2、基本理论

2.1 因子分析

因子分析是一种降维、简化数据的技术。它通过研究众多变量之间的内部依赖关系，探求观测数据中的基本结构，并用少数几个“抽象”的变量来表示其基本的数据结构。这几个抽象的变量被称作“因子”，能反映原来众多变量的主要信息。原始的变量是可观测的显在变量，而因子一般是不可观测的潜在变量。因子分析中的公共因子是不可直接观测但又客观存在的共同影响因素；每一个变量都可以表示成公共因子的线性函数与特殊因子之和，它的数学模型可表示为：

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ a_{21} & \cdots & a_{2p} \\ \vdots & & \vdots \\ a_{p1} & \cdots & a_{pm} \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix}$$

即 $X = AF + \varepsilon$。其中 $X = (x_1, x_2, ..., x_p)^T$ 是可观测的 p 维随机向量，每个分量代表一个指标或者向量。 $F = (F_1, F_2, ..., F_m)^T$ 中的 $F_1, F_2, \cdots, F_m$ 为 $m(m \le p)$ 个公因子变量，是各个原观测变量的表达式中都出现的因子，是相互独立的不可观测的理论变量。矩阵 A 称为因子载荷矩阵，a_{ij} 称为因子载

荷，表示第 i 个原有变量和第 j 个公共因子变量的相关系数，a_{ij} 越大说明公共因子 F_j 和原有变量 X 的相关性越强。ε 为特殊因子，表示原有变量不能被公共因子变量所解释的部分，相当于多元线性回归分析中的残差部分。

因子分析利用了降维的思想，由研究原始变量相关矩阵内部的依赖关系出发，根据相关性的大小把原始变量分组，使得同组内的变量之间相关性高，而不同组的变量间的相关性较低。每组变量代表一个基本结构，并用一个不可观测的综合变量表示，这个基本结构就称为公共因子。抓住这些主要的因子就可以帮助我们对复杂的问题进行分析和解释。

2.2 Logistic回归模型

2.2.1. 因变量为定性变量的回归模型

1) 定性变量：因变量只取两结果，当 $y=0$ 时表示事件未发生， $y=1$ 时表示事件发生。考虑简单的线性回归模型

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$E\left(y_i\right) = \beta_0 + \beta_1 x_i$$

由于 y_i 是 0~1 型伯努利随机变量，得到如下概率

$$P(y_i = 1) = \pi_i$$

$$P(y_i = 0) = 1 - \pi_i$$

根据离散型随机变量期望定义，得

$$E\left(y_i\right) = 1\left(\pi_i\right) + 0\left(1 - \pi_i\right) = \pi_i$$

所以 $E\left(y_i\right) = \pi_i = \beta_0 + \beta_1 x_i$。

2) 误差项

对取值为 0 或 1 的因变量，误差项 $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$ 只能取两值

$$y_i = 1, \varepsilon_i = 1 - (\beta_0 + \beta_1 x_i) = 1 - \pi_i$$

$$y_i = 0, \varepsilon_i = -(\beta_0 + \beta_1 x_i) = -\pi_i$$

误差项是两点型离散分布，所以不能假设其是正态误差回归模型。

零均值异方差：误差项为零均值，其方差不相等

$$D(\varepsilon_i) = D(y_i) = \pi_i(1 - \pi_i) = (\beta_0 + \beta_1 x_i)(1 - \beta_0 - \beta_1 x_i)$$

若用多元线性回归方程分析因变量与自变量之间的定量关系

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

3) 等式左边 y 取 0 或 1，等式右边可取任意实数，左右两边取值范围不对应。因此不能采用多元线 性回归进行因变量为定性变量的拟合。

2.2.2. Logistic 回归模型

Logistic 函数的形式为：

$$f\left(x\right)=\frac{e^{x}}{1+e^{x}}=\frac{1}{1+e^{-x}}$$

其自变量的取值范围是$\left(-\infty,+\infty\right)$，函数值的取值范围为$\left(0,1\right)$。

因变量y本身只取 0，1 两离散值，不适于作为回归模型中的因变量，令

$$\pi_i=f(x_i)=\frac{1}{1+\exp(-(\beta_0+\beta_1 x_i))}$$

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right)=\beta_0+\beta_1 x_i$$

其中π_i是随机变量y取 1 的概率，其值在$[0,1]$区间内连续变化，因此可用π_i代替y作为因变量。设y是 0~1 型变量，n组观测数据为$\left(x_{i1},\cdots,x_{ip},y_i\right)$，其中$y_1,y_2,\cdots,y_n$是取值 0 或 1 的随机变量，

$$E(y_i)=\pi_i=f(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})$$

Logistic 回归模型

$$\pi_i=f\left(x_i\right)=\frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}$$

于是y_i是均值为$\pi_i=f(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})$的 0~1 型随机变量，概率函数为

$$P(y_i=1)=\pi_i$$

$$P(y_i=0)=1-\pi_i$$

可以把y_i的随机概率定义为

$$P(y_i)=\pi_i^{y_i}(1-\pi_i)^{1-y_i},y_i=0,1;i=1,\cdots,n$$

于是$y_1,y_2,\cdots,y_n$的似然函数为

$$L=\prod_{i=1}^{n}P(y_i)=\prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

似然函数取对数，得

$$\ln L=\sum_{i=1}^{n}\left[y_i\ln\pi_i+\left(1-y_i\right)\ln\left(1-\pi_i\right)\right]=\sum_{i=1}^{n}\left[y_i\ln\frac{\pi_i}{1-\pi_i}+\ln\left(1-\pi_i\right)\right]$$

将式带入得

$$\ln L=\sum_{i=1}^{n}\left[y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)-\ln\left(1+\exp\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\right)\right]$$

最大似然估计得到 $\beta_0,\beta_1,\cdots,\beta_p$ 的估计值 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$

3、预测模型

3.1 数据处理

本文数据来自于“世界银行” *https://data.worldbank.org.cn/* 中的数据集

WLD_2023_SYNTH-SVY-EN_v01_M

数据集中的指标说明如下表 1 所示，该数据集主要用于分析居住地（城市/农村）、家庭规模、住房占用状态（自有/租赁）、全国人均支出五等分和总年度家庭支出等定量和定性指标如何影响个体拥有手机的可能性。

表 1 指标说明

符号	指标	指标类别
	家庭 ID	定性指标
X_1	居住地，城市/农村	定性指标
X_2	家庭规模	定量指标
X_3	住房占用状态，自有/租赁	定性指标
X_4	全国人均支出五等分	定性指标
X_5	总年度家庭支出	定量指标
Y	是否拥有移动电话	定性指标

数据清理的过程中，主要经过以下步骤：

1) 数据审查；

2) 缺失值处理；

3) 异常值检测；

4) 编码定性变量。

经过以上步骤，发现原始数据集，不存在缺失、异常值，数据完整。但数据集中有一字段为“家庭 ID”，经过与其他数据集对比，发现，该字段的分析效果，分析意义不大，所以将“家庭 ID”字段删除。

3.2 因子分析

1) 因子分析适用性检验

利用 Spss 软件对糖尿病数据进行 KMO 和 Bartlett 球度适用性检验，结果如表 2 所示，一般认为 KMO 度量值若大于 0.5，则可以进行因子分析。且显著性 p=0，说明原有变量之间存在一定的关联性，具备进行因子分析的条件。

2) 提取公因子

对数据进行因子分析，通过主成分分析法进行主成分的提取。在特征值为 1 的原则下，保留三个主因子，即 5 个变量归为 3 类。

表 2 KMO 和巴特利特检验

KMO 和巴特利特检验		
KMO 取样适切性量数		0.566
巴特利特球形度检验	近似卡方	13395.314
	自由度	10
	显著性	0.000

表 3 总方差解释

总方差解释						
成分	初始特征值			提取载荷平方和		
	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	1.999	39.990	39.990	1.999	39.990	39.990
2	1.314	26.290	66.280	1.314	26.290	66.280
3	0.962	19.233	85.513	0.962	19.233	85.513
4	0.601	12.022	97.535			
5	0.123	2.4650	100.00			

3) 公共因子命名

通过提取出来的 3 个公共因子，进行最大方差正交旋转，对原始因子载荷矩阵进行旋转，得到方差最大正交旋转矩阵，如表 4 所示。

表 4 旋转后的因子载荷矩阵

旋转后的因子载荷矩阵			
	Factor1	Factor2	Factor3
居住地，城市/农村	0.47294	-0.12173	**0.34263**
家庭规模	0.06622	**0.88178**	-0.17240
住房占用状态，自有/租赁	-0.00254	-0.06588	**0.38572**
全国人均支出五等分	**0.71768**	-0.64423	0.11168
总年度家庭支出	**0.89738**	0.17530	-0.13594

根据旋转后的成分矩阵，可将 3 个公共因子进行命名。

第一个因子 Z1 在全国人均支出五等分和总年度家庭支出指标上具有较大载荷；

第二个因子 Z2 在家庭规模指标上具有较大载荷；

第三个因子 Z3 在居住地，城市/农村、住房占用状态，自有/租赁上具有较大载荷；

可以得到，Z1 所对应的评价指标是"经济水平因子"，Z2 所对应的评价指标是"家庭规模因子"，Z3 所对应的是"居住条件因子"。分别命名为经济水平因子、家庭规模因子、居住条件因子。

3.1 二元Logistic 回归

1) ROC 曲线和 AUC 值

在本研究中，为了评估二元 Logistic 回归模型的性能，采用了 ROC 曲线和 AUC 值作为主要的评估工具。ROC 曲线是通过在不同阈值下绘制真正例率（TPR）与假正例率（FPR）来构建的，而 AUC 值则量化了模型区分正负样本的能力。模型在训练集上进行了拟合，并在独立的测试集上进行了评估。

AUC Score	0.8026344444444444

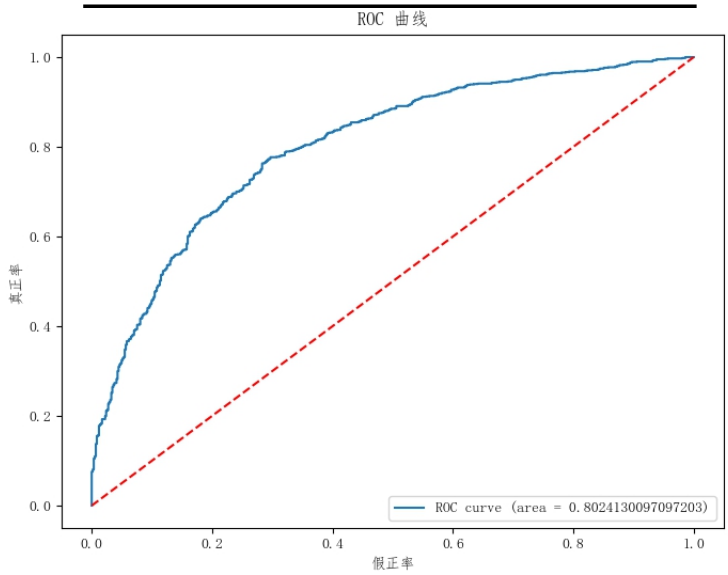

图 1 ROC 曲线

结果显示，模型的 ROC 曲线下面积（AUC）为 0.803，表明模型具有良好的分类性能。这一结果意味着我们的模型能够有效地区分两个类别的样本，且在测试集上具有较高的泛化能力。

2) 准确度如表 5 准确度为 74.5%，说明模型预测较为准确。

表 5 预测准确度

		是否拥有移动电话		正确百分比
		0	1	
是否拥有移动电话	0	527	373	71
	1	213	1287	78
总体百分比				74.5

3) 由表 6，显著性 P 值均很小，表示经济水平因子、家庭规模因子、居住条件因子对于是否拥有移动电话具有十分显著的影响。其影响程度由高到低排序如下：经济水平因子>家庭规模因子>居住条件因子。

表 6 Logistic 回归分析

	B	标准误差	瓦尔德	自由度	显著性	Exp(B)	EXP(B) 的 95% 置信区间	
							下限	上限
经济水平因子	1.4930	0.041	1329.132	1	0.000	4.501	4.152	4.880
家庭规模因子	0.0821	0.030	4.270	1	0.039	1.063	1.003	1.126
居住条件因子	-0.0048	0.051	0.044	1	0.048	1.011	0.914	1.118
常量	0.8171	0.030	733.863	1	0.000	2.229		

4) 由多因素的回归分析，建立二元 Logistic 回归方程

$$\mathrm{Logit}P = 0.8171 + 1.4930z_1 + 0.0821z_2 - 0.0048z_3$$

$$P = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

4、总结

经济水平因子对因变量的影响最大，具有正系数，表明随着经济水平的提升，拥有手机的概率增加；家庭规模因子也对因变量有正向影响，但其影响相对较小；居住条件因子的影响较小，且系数接近零，表明其对是否拥有手机的影响可能可以忽略不计。结合原始数据集，得到以下结论：全国人均支出五等分和总年度家庭支出指标对因变量是否拥有移动电话的影响最大。

本文的主要目的是利用 Logistic 回归模型对是否拥有移动电话进行分析，同时结合因子分析的思想，对数据进行降维处理。总体来说，用于初步诊断，本文得到的模型预测效果较好，准确率达到了 74.5%。其中全国人均支出五等分、总年度家庭支出影响较大。

参考文献

[1] 张初兵,高康,杨贵军.判别分析与 Logistic 回归的模拟比较[J].统计与信息论坛,2010,25(01):19-25.

[2] 徐娇. 基于因子分析的二元 Logistic 回归对糖尿病预测的研究[J]. 应用数学进展, 2022, 11(1): 108-115. DOI: 10.12677/aam.2022.111016

[3] 卢美婧,李玉娟.ESG 评级对中小企业信用风险评估的影响研究——基于因子分析法和 Logistic 回归法[J].中国物价,2024,(06):40-45.

[4] 郑苏.职业学校德育教育加强网络意识形态安全教育研究——基于智能手机普及的环境下[J].现代商贸工业,2024,45(12):91-93.DOI:10.19311/j.cnki.1672-3198.2024.12.032.