

2015 年全国大学生统计建模大赛

题 目	电子商务之男士钱包销售统计研究 ¹
学 院	数统学院
专 业	数学金融
班 级	数学金融
学生姓名	冉燕 2013101135
	赵丽 2013102113
	蒋扬宇 2013102108
指导教师	张天永
职 称	讲师

¹ 注:该论文获得由中国统计教育学会举办的“2015 年 (第四届) 全国大学生统计建模大赛”大数据统计建模类本科生组三等奖。

电子商务之男士钱包销售统计研究

一 . 摘要

在如今互联网飞速发展的时代,人们已经习惯了网络购物的生活方式,因此越来越多的人在自己做淘宝。可现如今淘宝市场的竞争越来越激烈,想要在淘宝行业中以及同行业中做得更好,我们一定要有专业准确的行业数据分析才能让你更加了解淘宝的整个市场。因此我们就淘宝平台上的男士钱包的一些数据进行统计建模分析。我们从不同的平台、区域、店铺、品牌、属性、价格、上架、淘宝指数等进行统计分析,从而更好地让各个商家为不同的顾客选择适合他们的品牌、款式、定价的产品。我们针对需要投资开店以及寻求店铺突破的客户,我们从店铺运营情况和产品的定位这两个大方向来大体上进行分析。

针对问题一,我们利用 Eviews 来分析影响店铺销量或销售额的多重性因素分析,从而得出哪些因素对销售量的影响有显著性,从而为店铺以后发展的方向 and 方式提供可行性建议。

针对问题二,产品的定位包括很多因素,我们从产品的质地、款式、硬度、闭合方式、信誉这几个作为属性等进行分析,用 KNN 针对这几个属性来进行分类,了解产品适合的人群即了解消费者消费行为倾向,然后为不同的顾客推荐不同的产品,这样让店铺对自己的产品进行更好的分类。

针对上面的两个模型,我们用 Eviews 和 KNN 来进行求解。通过对消费者行为倾向分析以及影响销售量的多重性因素分析,我们可以为店家提供更多的可行性建议,为店铺经营寻求突破,让店铺的走势更好。

最后,本文讨论了该模型的优缺点,并作了进一步的推广与改进,通过分析其它条件和考虑更多因素可以使模型的应用范围更广,功能更强。

关键词 多重性因素分析 K 近邻算法 多元回归分析

二、问题重述

现如今已经进入网络信息化的时代，电子商务已经成为时代的主流，越来越多的人在自己创业，自己当老板做淘宝，然而如何才能在众多同类的淘宝店里脱颖而出呢，因此我们需要对该行业进行一个整体的了解，我们应该注重产品、渠道、服务、营销，对这些进行专业的分析，让我们更加的了解淘宝市场，有效的选择好淘宝上的热销的品牌、产品、款式、定价。

三、问题分析

我们要掌握淘宝上男士钱包各个店铺、品牌的情况，因此我们针对上面两个问题进行详细分析。

针对问题一，淘宝搜索的时候根据他搜索的包的款式的类型，推测他是哪个年龄段的人群，然后为他推荐适用这个年龄段的钱包，这样就能让顾客在短时间内找到适合自己的钱包，如此一来既能让销售量增加也能让顾客对此店铺的产品有一个好的评价，对此店铺的产品定位有一个更清晰的认识。通过消费者行为倾向分析，我们对消费者的行为更加了解，这样能达到双赢的效果。

综合以上分析，我们用 K 近邻算法来实现上面这一过程，我们对店铺的各种款式的包进行分类，分别从不同的角度来划分，综合分析，我们将从包的以下几个属性来进行分类——材质、款式、硬度、闭合方式、信誉，分类标签——儿童、少年、青年、中年。分析消费者行为倾向来为他们推荐合适的包是我们所要解决的问题，利用 KNN，在 python 软件上来实现这一算法。

针对问题二，我们旨在解决如何提高销售量销售额的问题，当然我们需要分析影响店铺销售量的多重性因素分析，我们选取了店铺商品做活动时的优惠价格浏览人数、收藏数、评论数来进行多元线性回归分析，通过 evIEWS 软件用最小二乘法做出回归结果，通过对回归方程结果的分析，我们做出以下检验和判断。

(1) 对方程的各个变量进行 t 检验，看在 $\alpha=5\%$ 显著性水平下是否显著。若有变量的 t 统计值不大于 t 临界值，则可说明该模型可能存在多重共线性；

(2) 通过 F 检验来判断方程整体上的显著性；

(3) 可决系数能检验方程回归的拟合程度，可决系数说明影响销售量有可决系数能由着几个变量来解释；

(4) 通过 evIEWS 软件计算出各变量之间的相关系数，若相关系数值高，则说明存在多重共线性，利用逐步回归法，以 Y 为被解释变量，逐个引入解释变量构成回归模型，进行模型检验，根据拟合优度的变化决定新引入的变量是否可以用其他变量线性组合代替，而不是作为独立的解释变量。如果拟合优度变化显著，则说明新引入的变量是一个独立解释变量，如果拟合优度变化很不显著，则说明新引入的变量不是一个独立解释变量，他可以用其他变量的线性组合代替，也就是说与其他变量之间存在共线性的关系。

(5) 根据回归结果可以看出 D.W. 值, 检验是否存在自相关, 如果有用广义最小二乘法或广义差分法修正。

(6) 根据 G-Q 检验或怀特检验看异方差是否存在, 用加权最小二乘法修正。综上, 我们会得出这个模型的最终回归方程, 从而从模型中可以看出销售量的影响因素, 从而店家可以清晰的了解销售量好坏的原因, 分析各变量是如何影响销售量, 从而为店铺的运营有了更清晰的方向。

综合以上分析, 我们建立了消费者行为倾向以及影响销售量以及销售额的多重性因素分析两个模型, 为解决产品定位以及产品运营量大问题。

四 . 符号说明

x_i 表示钱包的材质 (PU、 PVC、 涤纶、 鳄鱼皮、 帆布、 棉纶、 牛津纺、 牛皮、 牛仔布、 蛇皮、 羊皮、 猪皮、 无纺布 、 其他)

y_i 表示钱包的款式 (短款、 长款、 硬币包、 其他)

z_i 表示钱包的硬度 (软、 硬)

d_i 表示钱包的闭合方式 (包盖式、 敞口、 搭扣、 挂钩、 拉链、 拉链搭扣、 魔术贴、 其他)

e_i 表示钱包的信誉 (皇冠 1、 钻石 1、 皇冠 2、 皇冠 3、 钻石 3、 钻石 4、 钻石 5)

f_i 表示样本分类 (儿童、 少年、 青年、 中年)

五 . 指标选择

1. K 近邻算法的原理是存在一个样本数据集合, 也称作训练样本集, 并且样本集中每个数据都存在标签, 即我们知道样本集中每一数据与所属分类的对应关系, 输入没有标签的新数据后, 将新数据的每个特征与样本集中数据对应的特征进行比较, 然后苏纳法提取样本集中最相似数据的分类标签, 一般来说我们只选择样本数据集中前 K 个最相似的数据, 这就是 K 近邻算法中 K 的出处, 选择 K 个最相似数据中出现次数最多的分类, 作为新数据的分类。

2.最小二乘法要求样本回归函数尽可能好的拟合这组值,即样本回归线上的点 \hat{Y}_i 与真实观测点 Y_i 的总体误差尽可能小,也就是说被解释变量的估计值与实际观测值只差的平方和最小,即在样本观测值下选择 $\hat{\beta}_0$ $\hat{\beta}_1$ 使 Y_i 与 \hat{Y}_i 之差的平方和最小。

六 . 数据描述

我们在数据堂上下载了淘宝男士钱包的数据,然后根据我们所建立的模型对需要的数据进行抓取,并且有些数据是用 excel 表格计算统计出来的。

七 . 模型假设

- 1.所选取的指标符合模型的要求
- 2.数据都是独立的

八 . 模型建立与求解

一.消费者行为倾向的模型建立与求解

1.1 准备数据

收集淘宝、天猫等平台的男士钱包销售情况原始数据,存放于 excel1.xlsx 中,从中提取部分所需数据放在 excel2.xlsx 中,使用 Python 将字符串数据转换成数值型存放于文本文件 wallet.txt 中,每个样本数据占一行,总共 3078 行。样本包含以下 5 中特征 :信誉、质地、硬度、款式、闭合方式。创建名为 file2matrix 的函数,该函数输入为文件名字字符串,输出为训练样本矩阵和类标签向量。

质地	款式	硬度	闭合方式	信誉	样本分类
PU101	短款钱包 201	软 301	包盖式 401	皇冠 1 501	儿童 1
PVC102	长款钱包 202	硬 302	敞口 402	钻石 1 502	青年 2

涤纶 103	其他 203		抽带 403	皇冠 2 503	少年 3
鳄鱼皮 104			搭扣 404	皇冠 3 504	中年 4
帆布 105			挂钩 405	钻石 3 505	
锦纶 106			拉链 406	钻石 4 506	
牛津纺 107			拉链搭扣 407	钻石 5 507	
牛皮 108			魔术贴 408	淘宝 508	
牛仔布 109			其他 409		
羊皮 110					
猪皮 111					
无纺布 112					
其他 113					

1.2 分析数据：

使用 Matplotlib 制作原始数据的散点图,在 Python 命令环境中,输入下列命令：

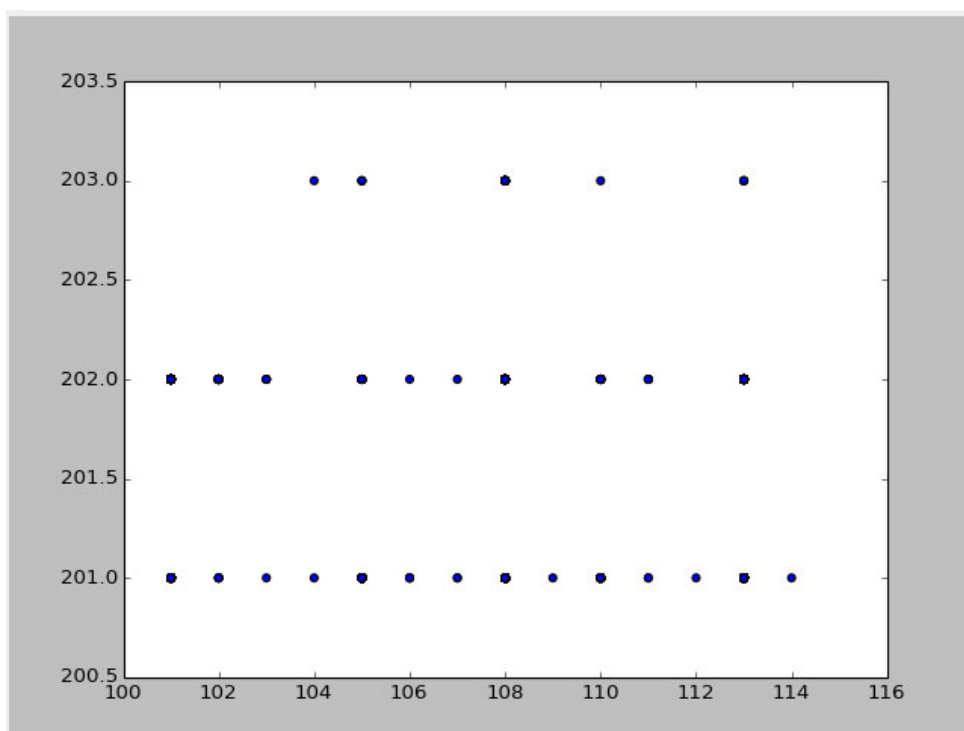
```
import kNN
import matplotlib
import matplotlib.pyplot as plt

group, labels=kNN.createDataSet()
print group, labels
datingDataMat, datingLabels=kNN.file2matrix('wallet.txt')
print datingDataMat, datingLabels[0:10]

fig=plt.figure()
ax=fig.add_subplot(111)
ax.scatter(datingDataMat[:,1], datingDataMat[:,3])
plt.show()
```

输出效果如图所示：

散点图使用 datingDataMat 矩阵的第二、第四列数据，分别表示“质地”和“款式”。



上图的 X 轴表示“质地”，Y 轴表示“款式”。根据图示可以知道，男士更倾向于短款钱包，并且更喜欢帆布、牛皮、羊皮材质的钱包。

1.3 归一化数值：

函数 `autoNorm()` 自动将数字特征化转化为 0 到 1 的区间，并创建新的矩阵

原始数据改进后的样本数据部分数据

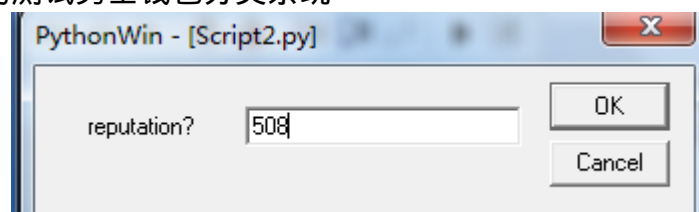
信誉	质地	硬度	款式	闭合方式	适用对象
508	108	301	202	408	4
508	108	301	202	406	4
508	108	301	202	406	4
508	101	302	202	406	2
508	110	301	202	406	2
508	108	301	203	402	2
508	108	301	202	406	4
508	108	301	202	406	2
508	108	301	201	409	2

1.4 测试算法：

函数 `datingClassTest()`，计算分类器的错误率并输出结果。可以改变函数 `datingClassTest` 内变量 `hoRatio` 和变量 `k` 的值，赖于分类算法、数据集和程序设置，分类器的输出结果可能有很大的不同。

1.5 使用算法：

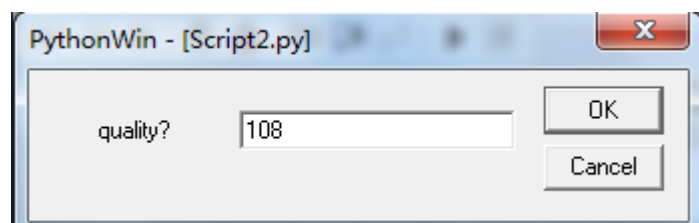
构建完整的测试男士钱包分类系统



PythonWin - [Script2.py]

reputation?

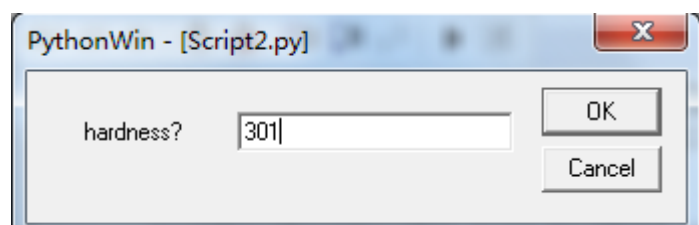
OK Cancel



PythonWin - [Script2.py]

quality?

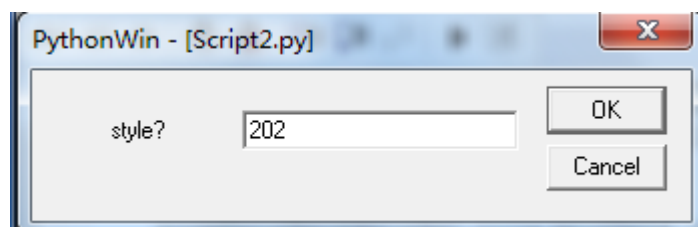
OK Cancel



PythonWin - [Script2.py]

hardness?

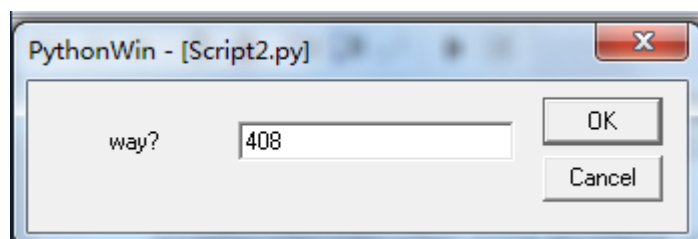
OK Cancel



PythonWin - [Script2.py]

style?

OK Cancel



PythonWin - [Script2.py]

way?

OK Cancel

输出结果为：

```
PythonWin 2.7.8 (default, Jun 30 2014, 16:03:49) [MSC v.1500 32 bit (Intel)] on win32.  
Portions Copyright 1994-2008 Mark Hammond - see 'Help/About PythonWin' for further copyright  
information.  
>>> what kind of people: zhongnian
```

根据以上测试可以根据钱包的五个属性将其分为对应的适用对象。

二．研究影响销售量的多重性因素分析的模型建立与求解

一．建立模型

依据经济学原理，销售量与商品的价格，评论数，收藏数，浏览人数有一定关联，在本文中我们选定商品的价格，评论数，收藏数，浏览人数作为解释变量，选取商品销售量为被解释变量，因此可以建立如下的四元线性回归模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \mu_i$$

其中，Y 为销售量、商品的价格 X1、评论数 X2、收藏数 X3、浏览人数 X4，由于存在其他不确定的因素影响故增添了 μ 随机误差项。

利用 eviews 通过最小二乘法估计得出：

$$\begin{aligned} Y &= 87.19484 + 10298985X_1 - 0.072899X_2 + 0.005793X_3 - 0.002517X_4 \\ t & (3.909202) (55.43207) (-1.531622) (3.161865) (-0.866765) \\ R^2 &= 0.917402 \quad D.W. = 1.996030 \\ F &= 2440.721 (P=0.000000) \end{aligned}$$

从经济角度上来看，X2 和 X4 系数为负不符合经济学意义，但它应该是一定程度上的决定变量，所以暂不排除。

从统计学角度上来看， R^2 \bar{R}^2 F 值都比较高，X4 的 t 统计量不显著可能存在多重共线性。

二．多重共线性检验

	X1	X2	X3	X4
X1	1.000000	-0.161900	0.800896	0.666761
X2	-0.161900	1.000000	-0.084331	-0.129299
X3	0.800896	-0.084331	1.000000	0.834107
X4	0.666761	-0.129299	0.834107	1.000000

从上图可以看出，变量 X3 与 X4 的相关系数最大，达到 0.834107，可见变量 X4 和 X3 有较强烈的相关性，模型存在多重共线性。

修正多重共线性

1.以模型中的 Y 作为被解释变量，对各个解释变量分别进行回归得到如下结果：

(1) 对 X1 进行检验

$$\begin{aligned} Y &= 49.41015 + 1.368178X_1 \\ t &(3.454545) \quad (105.0443) \\ F &= 11034.30 \quad R^2 = 0.912884 \end{aligned}$$

(2) 对 X2 进行检验

$$\begin{aligned} Y &= 837.4936 - 0.798215X_2 \\ t &(12.84608) \quad (-5.011603) \\ F &= 25.11616 \quad R^2 = 0.027627 \end{aligned}$$

(3) 对 X3 进行检验

$$\begin{aligned} Y &= 358.0580 + 0.065576X_3 \\ t &(12.47487) \quad (40.97232) \\ F &= 1678.731 \quad R^2 = 0.614530 \end{aligned}$$

(4) 对 X4 进行检验

$$\begin{aligned} Y &= 228.8975 + 0.103835X_4 \\ t &(6.178947) \quad (27.23764) \\ F &= 741.8893 \quad R^2 = 0.413795 \end{aligned}$$

则由上的数据得知在四个辅助模型中，辅助回归中 X1 的 R^2 的最大，因此以辅助回归 1 作为基本回归模型。

在基本回归模型上，按变量自然顺序，逐个引入解释变量 1，引入变量 X2

$$Y=66.60443+1.365038X_1-0.055641X_2$$

$$t \quad (3.211135) \quad (96.75761) \quad (-1.173586)$$

$$F=4826.558 \quad R^2=0.916193$$

较原来回归模型 $R^2=0.912884$ 有明显提高,满足逐步回归法的第一个条件。原来的解释变量 X_1 的系数为1.37, t 值对应的相伴概率为0,因此原来的解释变量 X_1 的经济意义合理且在统计上显著,满足逐步回归法的第二个条件。再来看新增加的解释变量 X_2 ,其系数为-0.556,经济意义不合理, t 值对应的相伴概率为0.24,在统计上也不显著,不满足逐步回归法的第三个条件。因此新增加的解释变量 X_2 应当从模型中剔除。这时,模型仍然为辅助回归1模型。

2, 引入变量 X_3

$$Y=58.44093+1.306237X_1+0.004522X_3$$

$$t \quad (4.045923) \quad (60.49585) \quad (3.585389)$$

$$F=5585.693 \quad R^2=0.913935$$

较原来回归模型 $R^2=0.912884$ 有明显提高,满足逐步回归法的第一个条件。原来的解释变量 X_1 的系数为1.306, t 值对应的相伴概率为0,因此原来的解释变量 X_1 的经济意义合理且在统计上显著,满足逐步回归法的第二个条件。再来看新增加的解释变量 X_3 ,其系数为0.004522,经济意义不合理, t 值对应的相伴概率为0,因此原来的解释变量 X_1 的经济意义合理且在统计上显著,满足逐步回归法的第三个条件。因此新增加的解释变量 X_2 应当保留在模型中。

3, 引入变量4

在前模型基础上,再引入变量 X_4 ,其估计参数为:

$$Y=61.89097+1.306183X_1+0.005311X_3-0.001834X_4$$

$$t \quad (4.046550) \quad (60.41918) \quad (3.151340) \quad (-0.706201)$$

$$F=3714.757 \quad R^2=0.913969$$

较原来回归模型 $R^2=0.916193$ 明显下降,不满足逐步回归法的第一个条件。应当从模型中剔除 x_4 。

则经过逐步回归,最终得到的模型为

$$Y=58.44093+1.306237X_1+0.004522X_3$$

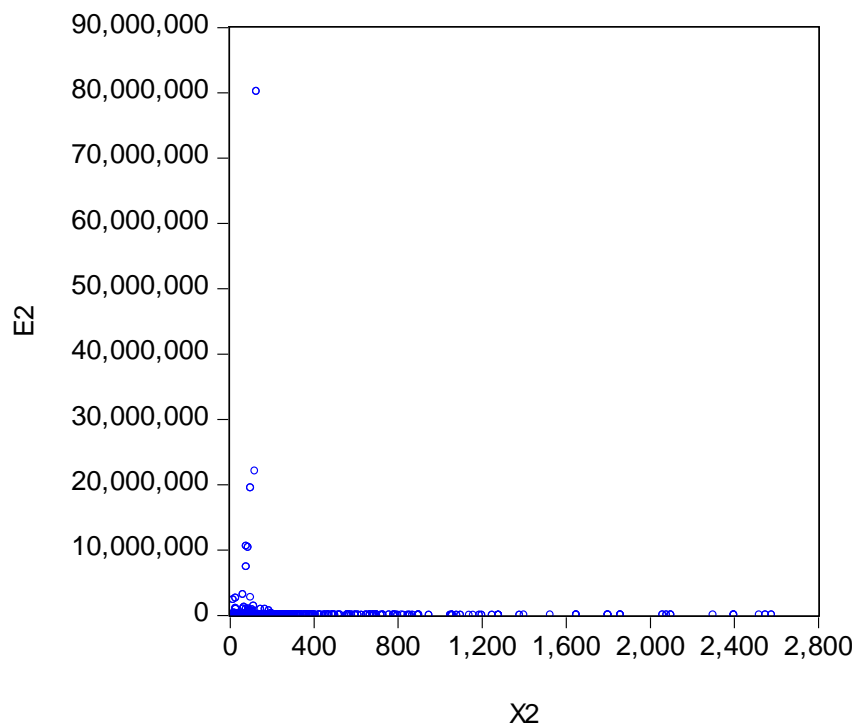
$$t \quad (4.045923) \quad (60.49585) \quad (3.585389)$$

$$F=5585.693 \quad R^2=0.913935$$

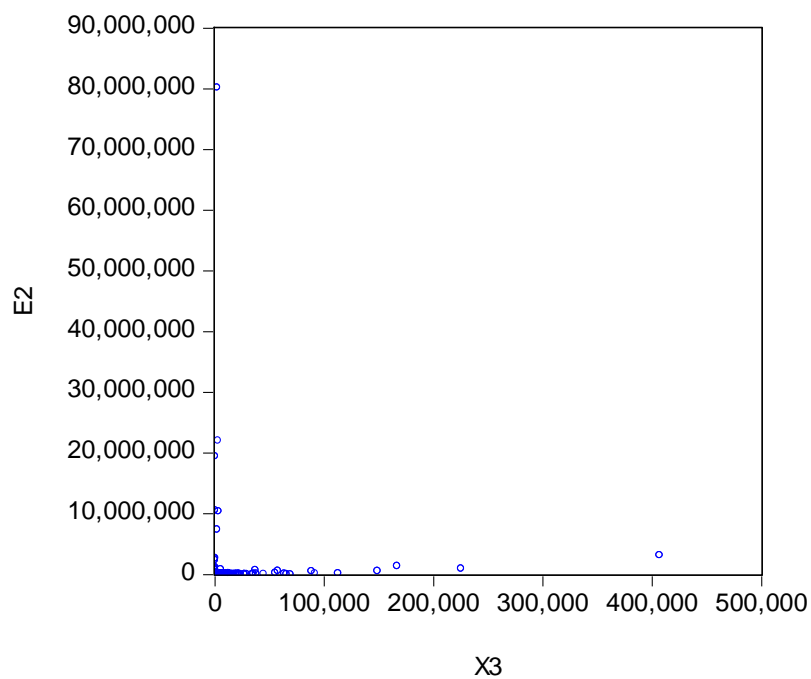
综合上面的回归结果,可以知道此时 X_1 X_3 变量的 t 检验都是显著的,并且 F 值很高,说明方程在整体上是显著的,并且 $R^2=0.913935$ 说明整个模型的拟合优度很好,在销售量的变动中,有91.3935%可由浏览人数和评论数的变动来得到解释。

三. 异方差

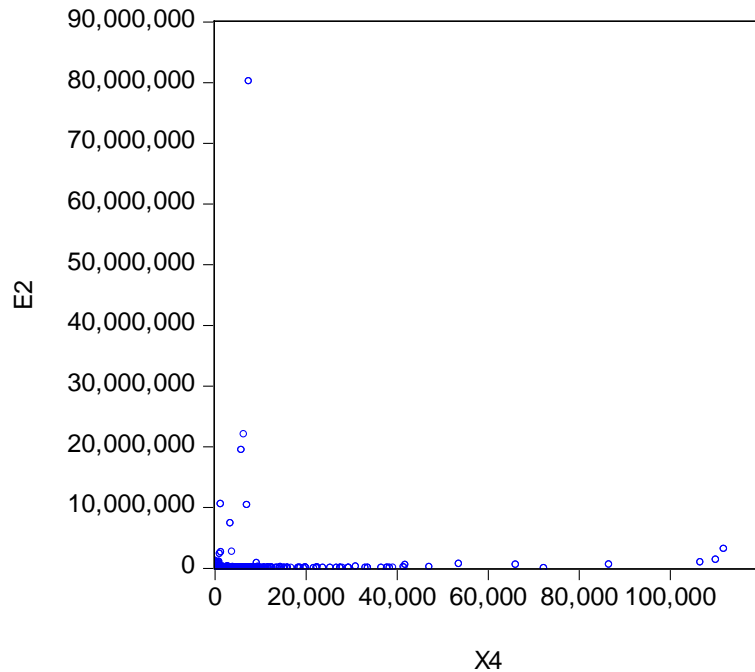
散点图可以看出随着 X_1 的增加而减少,因此怀疑模型存在递减型异方差。



散点图可以看出随着 X_2 的增加而减少，因此怀疑模型存在递减型异方差。



散点图可以看出随着 X_3 的增加而减少，因此怀疑模型存在递减型异方差。



散点图可以看出随着 X_4 的增加而减少，因此怀疑模型存在递减型异方差。

G-Q 检验

由图示检验可大致知道模型存在递减型异方差，且四个变量均可能是异方差。

Variable	Coefficient		t-Statistic	Prob.
	Std.	Error		
C	187.4731	58.49888	3.204730	0.0015
X1	-0.729003	0.731809	-0.996166	0.3199
X2	-0.140642	0.046895	-2.999059	0.0029
X3	-0.024610	0.044695	-0.550633	0.5822
X4	0.121401	0.020853	5.821872	0.0000
R-squared	0.143370	Mean dependent var	163.5155	
Adjusted				
R-squared	0.133580	S.D. dependent var	380.4351	
S.E. of		Akaike info		
regression	354.1153	criterion	14.59111	
Sum squared resid	43889179	Schwarz criterion	14.64564	

		Hannan-Quinn	
Log likelihood	-2584.921	criter.	14.61280
F-statistic	14.64450	Durbin-Watson stat	1.780386
Prob(F-statistic)	0.000000		

上图是以 x1 进行 G-Q 检验，将范围由 1-1055 减为 1-500 进行回归。得到的残次平方和为=43889179
再对后500个样本进行回归得到

Variable	Coefficient	t-Statistic	Std. Error	Prob.
C	54.36098	40.52603	1.341384	0.1804
X1	1.310941	0.028419	46.12925	0.0000
X2	0.037840	0.207711	0.182176	0.8555
X3	0.005723	0.002199	2.602087	0.0096
X4	-0.003264	0.003515	-0.928432	0.3537
R-squared	0.933470	Mean dependent var	1049.854	
Adjusted R-squared	0.932911	S.D. dependent var	2053.416	
S.E. of regression	531.8667	Akaike info criterion	15.40100	
Sum squared resid	1.35E+08	Schwarz criterion	15.44441	
Log likelihood	-3698.941	Hannan-Quinn criter.	15.41806	
F-statistic	1669.666	Durbin-Watson stat	1.385607	
Prob(F-statistic)	0.000000			

上图是以 x1 进行 G-Q 检验，将范围由 1-1055 减为 556-1055 进行回归。得到的残次平方和为=1.35E+08
F 的统计量为 $F = 1.35E+08 / 43889179 = 3.0759290348083$
则当取 5% ，对应的 F 分布的临界值分别为 $F(, 8) = 2.93$, $F(, 13) = 3.17$. 显然在 5% 的显著情况下不能拒绝同方差的原假设。

由此我们可以得出商品的销售量与商品的价格和收藏数有很大的关系 , 因此我们的商家可以根据上面的结论为自己店铺提供可行性建议。

九.结论与建议

综上所述,我们可以利用 Python 和 Eviews 软件求解出以上两个模型,根据模型建立的过程以及得出的结果我们可以根据顾客在店铺里的浏览记录以及所点击的包的类型我们可以分类出消费者所需要的包的类型;并且根据在网上抓取得数据我们可以建立一个多元线性模型得出销售量的影响因素。

由此我们可以得出商品的销售量与商品的价格和收藏数有很大的关系,因此我们的商家可以对商品的价格做出适当调整,并且可以在一些节假日里面做出一些优惠活动。

十.模型的优缺点

一. 优点

- 1.数据来源真实,可以得出正确的结论。
- 2.本论文中采用几个多样化的模型,灵活独特的对题中所涉及的问题进行详细准确的求解。
- 3、思路清晰,语言严谨,假设比较合理,所用模型具有一般性,有利于推广,并且能合理的解决问题
4. 建立影响销售额多重性因素分析的回归模型,选用 Eviews 进行回归分析,具有一定的实际价值。
- 5.结合实际情况对问题进行求解,KNN 算法能与实际生活紧密联系使模型具有很好的通用性和推广性。

二. 缺点

- 1.由于题中数据量很大,所建模型对数据的统计抓取存在一定的难度,从而导致求解模型的最终结果不是很容易。
- 2.由于影响销售量的因素众多,而我们只分析了其中的几种,模型存在不完整性,因此需要加入更多的因素进去模型进行分析。

十一 . 模型改进

本文所建模型为信息的反馈提供了依据 ,充分发挥了评价的导向和激发功能 ,从而促进学生综合素质的提高 ,但大量的数据计算浪费了大量的时间 ,故寻找合适的计算方法是本文所改进的具体方向。