

南京理工大学经济管理学院

课程考核论文

课程名称： 统计分析软件及应用

论文题目： 关于汽车参数及价格情况分析报告

姓 名： 乔喜韬

学 号： 918107820131

成 绩：

任课教师评语：

签名：

年 月 日

注：请将该封面与论文装订成册。

英国汽车参数及价格情况分析报告

一、案例研究背景.....	1
二、数据来源及数据介绍.....	1
1、 变量.....	1
2、 数据集的补充信息.....	1
三、 数据分析及结果评价.....	2
1、 变量分布信息统计分析.....	2
(1) 目的：观察目前汽车市场的驱动轮分布情况.....	2
(2) 目的：分析汽车的马力、高速油耗是否存在不均衡情况.....	3
(3) 目的：观察车辆价格的总体分布.....	5
2、 大众认知的参数检验.....	6
(1) 目的：检验转速峰值在 5500 转之内的汽车是否达到了 90%.....	6
(2) 目的：检验引擎大小的均值是否为 2L.....	7
(3) 目的：检验不同门数是否会对整车重量造成影响.....	7
3、 品牌的价格策略.....	9
(1) 目的：观察不同厂商的整体的价格策略是否相同.....	9
(2) 目的：从分析金额是否有大量异常值入手，分析不同品牌的产品线分布情况.....	11
(3) 目的：以马力为核心指标分析不同品牌的性价比.....	13
4、 价格的影响因素：.....	14
(1) 目的：研究价格与高速油耗、马力之间是否存在关系.....	14
(2) 目的：进一步探究价格与高速油耗、马力之间的关系.....	16
(3) 目的：模型拓展（Python）.....	24
四、 结论.....	26
附录.....	28

英国汽车参数及价格情况分析报告

一、案例研究背景

随着汽车市场的逐渐壮大、国家政策对于新能源汽车的大力推动以及消费者的购买能力的提高，汽车企业成为了讨论热点和资本市场的宠儿。与此同时，汽车的种类、型号等参数变得复杂和专业化，如何选择性能良好且价格实惠的汽车成为了困惑汽车消费者的一大难题。

为了减少汽车消费者对于汽车参数的学习成本，本案例致力于呈现出一个直观且便捷的判断方法，给消费者对于汽车的核心参数和不同品牌的产品战略获得基本认知，从而帮助做出更加合理且满意的购车决策。

基于以上的目的，运用管理统计学的相关知识及 SPSS 软件对 205 个不同车型的参数和价格情况进行了分析，以了解汽车的整体情况、不同厂商的产品策略以及价格的主要影响因素，并分析个别变量的分布特点及相互关系。

二、数据来源及数据介绍

此数据来源于

<https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data>

1、变量

本次分析的数据集为“Auto Imports Database”，其中针对该研究目标进行了变量删减，共包含 16 个变量，分别为 make, fuel-type, aspiration, num-of-doors, drive-wheels, engine-location, curb-weight, engine-type, num-of-cylinders, engine-size, compression-ratio, horsepower, peak-rpm, city-mpg, highway-mpg, price（具体数据类型及取值见附录 1）；共计 205 个有效样本。

2、数据集的补充信息

源数据为 csv（comma separated value）格式（具体的数据集见附录 2），经过预处理并以此建立数据文件 auto mobile pricing.sav（见附录 3）。

此案例分析采用的是英国汽车市场的数据集进行分析，对于国内市场存在一定的时滞性和应用偏差。因此本文主要侧重于应用 SPSS 分析技能进行研究演示，同时建立起对汽车配置与价格之间的基本认知，读者可以根据事实情况适当进行应用参考。

三、 数据分析及结果评价

1、变量分布信息统计分析

(1) 目的：观察目前汽车市场的驱动轮分布情况

I. 采用方法：频数分析

II. 操作步骤：

a) 选择菜单：【分析】→【描述统计】→【频率】；

b) 将“驱动轮”变量选入【变量】框；

c) 点击【图表】按钮，选择【饼图】和【百分比】；

d) 点击【格式】按钮，选择【按计数的降序排序】

e) 单击【确认】完成。

分析结果如图 1 所示

III. 结果分析：

首先，该数据集的有效样本为 205 份，就驱动轮分布来看，fwd（前驱）有 120 辆，占比为 58.5%；rwd（后驱）有 76 辆，占比为 37.1%；4wd（四驱）有 9 辆，占比为 4.4%。而在家用车市场中前驱车确实也是最多的一种驱动方式，由于其经济性和效率上的优势被大众所接受，数据样本的驱动轮分布也基本符合这一实际情况。

		驱动轮			
		频数	百分比	有效百分比	累计百分比
有效	fwd	120	58.5	58.5	58.5
	rwd	76	37.1	37.1	95.6
	4wd	9	4.4	4.4	100.0
总计		205	100.0	100.0	

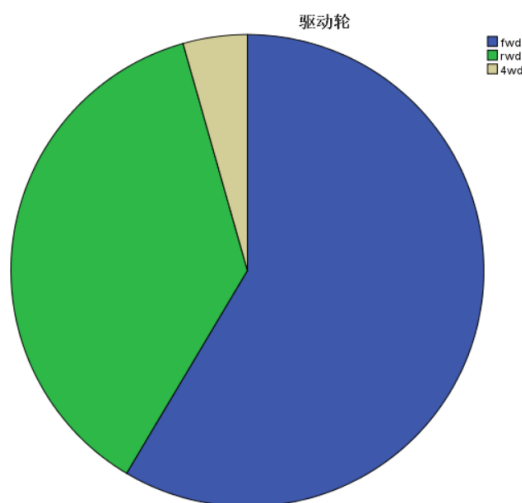


图 1 驱动轮分布情况饼图

(2) 目的：分析汽车的马力、高速油耗是否存在不均衡情况

I. 采用方法：描述性统计分析——描述

II. 操作步骤：

- a) 选择菜单：【分析】→【描述统计】→【描述】；
- b) 将“高速油耗”和“马力”变量选入【变量】框；
- c) 点击【选项】按钮，选择【标准误差平均值】、【偏度】和【峰度】；
- d) 单击【确认】完成。

分析结果如表 1、2 与图 2、3 所示

III. 结果分析：

高速油耗和马力均呈右偏尖峰分布，由于分布右偏，因此均值作为集中趋势的代表存在高估，具体见表 1。

	描述统计									
	个案数	最小值	最大值	平均值		标准差	偏度		峰度	
	统计	统计	统计	统计	标准误差	统计	统计	标准误差	统计	标准误差
高速油耗	205	16	54	30.75	.481	6.886	.540	.170	.440	.338
马力	203	48	288	104.26	2.787	39.714	1.391	.171	2.623	.340
有效个案数 (成列)	203									

表 1 高速油耗和马力的对比情况

并且表 2 表明，除了两缸汽车出现异常，不同缸数导致在马力和高速油耗上均存在较大差异。主要表现为随着缸数越多，高速油耗在逐渐下降，而马力在逐渐上升（图 2、3 的箱型图同样可以证明此结论）；其中，四缸和八缸都呈现不同程度的右偏尖峰分布，以均值作为集中趋势，都存在一定程度的高估；六缸呈现出左偏平峰分布，以均值作为集中趋势，存在一定程度的低估。另外，标准差表明，八缸的标准差最大即离散程度最大，而四缸的标准差最小即离散程度最小，相对较为集中稳定。

		描述统计									
缸数		个案数	最小值	最大值	平均值	标准差	标准差	偏度	峰度	统计	标准误差
		统计	统计	统计	统计	统计	统计	统计	统计		
2	高速油耗	4	23	23	23.00	.000	.000
	马力	4	101	135	109.50	8.500	17.000	2.000	1.014	4.000	2.619
	有效个案数 (成列)	4									
3	高速油耗	1	53	53	53.00
	马力	1	48	48	48.00
	有效个案数 (成列)	1									
4	高速油耗	159	22	54	32.77	.468	5.901	.626	.192	.606	.383
	马力	157	52	175	90.55	2.062	25.835	1.073	.194	.897	.385
	有效个案数 (成列)	157									
5	高速油耗	11	20	25	23.91	.530	1.758	-1.452	.661	1.014	1.279
	马力	11	110	160	122.45	4.666	15.475	1.655	.661	2.807	1.279
	有效个案数 (成列)	11									
6	高速油耗	24	19	28	23.83	.495	2.426	-.366	.472	-.023	.918
	马力	24	106	207	161.92	5.792	28.376	-.095	.472	-.484	.918
	有效个案数 (成列)	24									
8	高速油耗	5	16	28	19.20	2.245	5.020	2.017	.913	4.225	2.000
	马力	5	155	288	193.20	24.571	54.943	1.856	.913	3.665	2.000
	有效个案数 (成列)	5									
12	高速油耗	1	17	17	17.00
	马力	1	262	262	262.00
	有效个案数 (成列)	1									

表 2 不同缸数车辆马力与油耗的对比情况

高速油耗

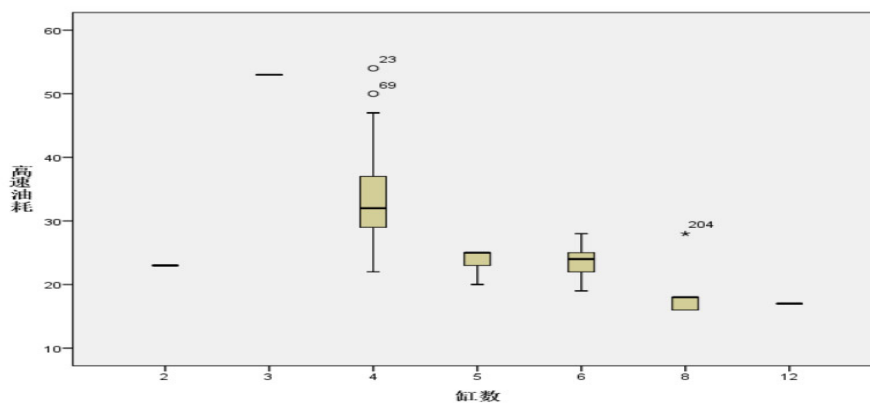


图 2 高速油耗箱型图

马力

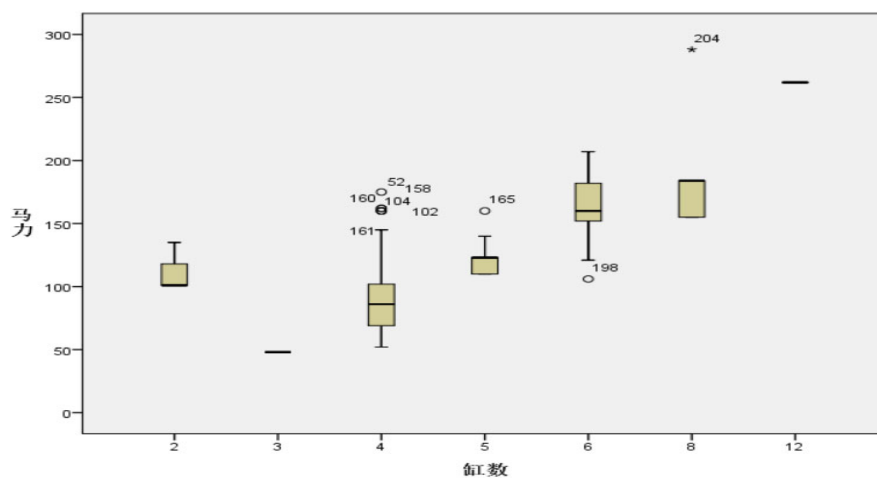


图 3 马力箱型图

(3) 目的：观察车辆价格的总体分布

I. 采用方法：描述统计——频率

II. 操作步骤：

- a) 选择菜单：【分析】→【描述统计】→【频率】；
- b) 将“价格”变量选入【变量】框；
- c) 点击【图表】按钮，选择【直方图】和【在直方图中显示正态曲线】；
- d) 点击【统计】按钮，选择【四分位数】、【偏度】和【峰度】
- e) 单击【确认】完成。

分析结果如图 4 所示

III. 结果分析：

就价格分布情况来看，表明有 25%的车辆价格低于£7775，有 25%的车辆价格高于£16501.5，有 50%的车辆价格在£7775~£16501.5 之间，四分位差为£8726.5，某种程度上也可以说明目前汽车市场上主流车型的价格主要集中在£10000 左右。通过与正态分布曲线的对比，容易发现车辆价格的分布是不对称的，具体见图 4 的直方图。

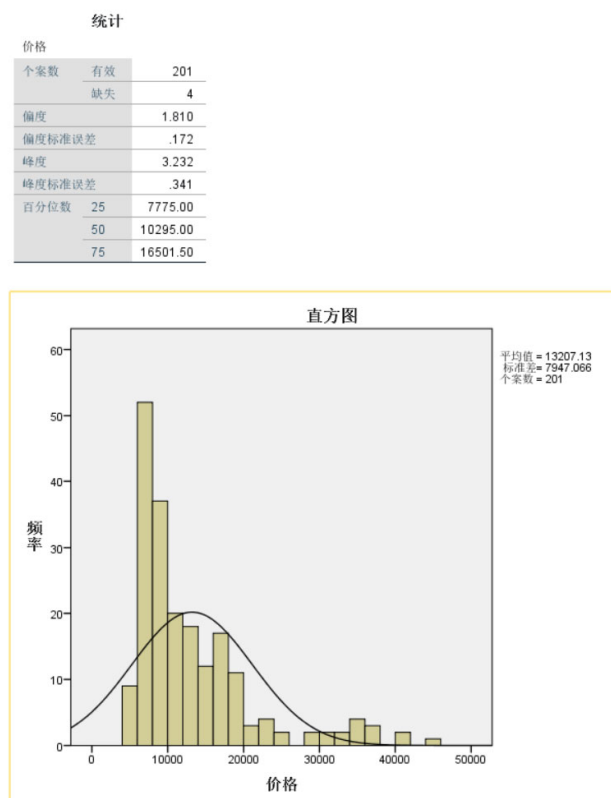


图 4 车辆价格分布直方图

2、大众认知的参数检验

(1) 目的：检验转速峰值在 5500 转之内的汽车是否达到了 90%

当发动机高速运转，增加缸体内各部件的摩擦，如果车辆没有充分的机油进行润滑时，对汽车发动机的损害较大，严重情况下还会造成“拉缸”。所以，对于日常使用的轿车，汽车制造商一般会将转速红线设置为 5500 转左右，为了检验该判断是否成立，此处对转速峰值进行参数检验。

I. 采用方法：非参数检验——二项

零假设 H0：样本来自的总体 90% 的车辆转速峰值低于 5500 转

II. 操作步骤：

- a) 选择菜单：【分析】→【非参数检验】→【旧对话框】→【二项】；
- b) 将“转速峰值”变量选入【检验变量列表】框；
- c) 设置【定义二分法】为分割点“5500”；
- d) 设置【检验比例】为“0.9”；
- e) 单击【确认】完成。

分析结果如表 3 所示

III. 结果分析：

该检验使用 5500 转作为二项分布的分界点，将样本数据分为了两组，分别为“≤5500 转”和“>5500 转”。假设显著性水平 α 为 0.05，因为得出的单尾概率 P-值为 0.295 (>0.05)，所以接受原假设，认为 90% 的车辆转速峰值低于 5500 转。

二项检验						
		类别	个案数	实测比例	检验比例	精确显著性 (单尾)
转速峰值	组 1	≤ 5500	180	.9	.9	.295 ^a
	组 2	> 5500	23	.1		
	总计		203	1.0		

a. 备用假设指出第一个组中的个案比例 < .9。

表 3 转速峰值的二项检验表

(2) 目的：检验引擎大小（单位：CID 立方英寸）的均值是否为 122CID (=2L)

I. 采用方法：单样本 t 检验

零假设 H0：引擎大小的均值与 122CID 无显著差异

II. 操作步骤：

- 选择菜单：【分析】→【比较均值】→【单样本 T 检验】；
- 将“引擎大小”变量选入【检验变量】框；
- 设置【检验值】为“122”；
- 点击【选项】按钮，选择【按分析顺序排除个案】，并且将【置信区间】设定为 95%，即显著性水平 α 为 0.05；
- 单击【确认】完成。

分析结果如表 4 所示

III. 结果分析：

对于该数据集，推断引擎的排量一般约为 2 升（122CID）。由于引擎排量可以近似认为服从正态分布，因此，可采用单样本 t 检验来进行分析。

对该问题应采用单侧检验方法，进而比较 α 和 $p/2$ 。如果 α 取 0.05，由于 $p/2$ 大于 α ，因此应接受原假设，认为引擎排量的均值与 2L 无显著差异。95% 的置信区间告诉我们有 95% 的把握认为引擎的排量均值在 121.17~132.64 之间，也证实了上述推断。

单样本检验

检验值 = 122

	t	自由度	显著性（双尾）	平均值差值	差值 95% 置信区间	
					下限	上限
引擎大小	1.687	204	.093	4.907	-.83	10.64

表 4 引擎排量的单样本 t 检验表

(3) 目的：检验不同门数是否会对整车重量造成影响

在传统认知下，两门与四门的汽车在车身重量上是存在较大的区别的，并且同种车型的两门会具有更小的车身和更轻的车重，以此来提高驾驶操作的灵活性。

I. 采用方法：独立样本 t 检验

零假设 H0：不同门数汽车的车重均值无显著差异

II. 操作步骤：

- 选择菜单：【分析】→【比较均值】→【独立样本 T 检验】；
- 将“车重”变量选入【检验变量】框；
- 将“门数”变量选入【分组变量】框；
- 点击【定义组】按钮，选择【使用指定值】并指定为“two”和“four”；
- 单击【确认】完成。

分析结果如表 5 所示

III. 结果分析：

由表 5 中的组统计表可以看出，不同门数的车重的样本平均值存在一定差异。通过检验来推断这种差异是由抽样误差造成还是系统性的。

第一步，两总体方差是否相等的 F 检验。该检验的 F 统计量的观测值为 0.974，对应的概率 P-值为 0.325。如果显著性水平 α 为 0.05，由于概率 P 值大于 0.05，可以认为两总体的方差无显著差异；第二步，两总体均值差的检验。在第一步中，由于两总体方差无显著差异，因此应看第一列（假设方差相等）检验的结果。其中，t 统计量的观测值为-2.829，对应的双侧概率 P-值为 0.005。如果显著性水平 α 为 0.05，由于概率里 P-值小于 0.05，拒绝原假设，认为两总体的均值有显著差异，即不同门数的车重的总体均值存在显著差异，且四门的重量大于两门的车重。该置信区间不跨零，也从另一个角度证实了上述推断。

组统计					
	门数	个案数	平均值	标准差	标准误差平均值
车重	two	89	2442.47	498.384	52.829
	four	114	2648.04	525.367	49.205

独立样本检验									
莱文方差等同性检验				平均值等同性 t 检验					
		F	显著性	t	自由度	显著性（双尾）	平均值差值	标准误差差值	差值 95% 置信区间
车重	假定等方差	.974	.325	-2.829	201	.005	-205.572	72.666	-348.858 -62.286
	不假定等方差			-2.847	193.502	.005	-205.572	72.194	-347.960 -63.183

表 5 独立样本 t 检验表

3、品牌的价格策略

(1) 目的：观察不同厂商的整体的价格策略是否相同

I. 采用方法：单因素方差分析

本例以价格为观测变量，厂商为控制变量，通过单因素方差分析方法对厂商对价格的影响进行分析，进而研究不同厂商的总体定价策略。

零假设 H₀：不同厂商没有对价格的平均值产生显著影响（即不同厂商对价格的效应同时为 0）。

II. 操作步骤：

首先，因为因子变量（分组变量）需要为数值型，所以对厂商进行重编码，使其全部变为数值型变量。

预处理：

- a) 选择菜单：【转换】→【重新编码成不同变量】→【描述】；
- b) 将“厂商”变量选入【输入变量】框；
- c) 在【输出变量】框中将【名称】设置为“make_num”，将标签设置为“厂商编号”并点击【变化量】按钮；
- d) 点击【旧值和新值】按钮，分别将每个厂商设置为 1~22 的有序数列；
- e) 单击【确认】完成。

ANOVA 检验：

- a) 选择菜单：【分析】→【比较均值】→【单因素 ANOVA】；
- b) 将“价格”变量选入【因变量列表】框；
- c) 将“厂商编号”变量选入【因子】框；
- d) 点击【选项】按钮，选择【方差齐性检验】和【平均值图】；
- e) 点击【事后比较】按钮，选择【LSD】检验；
- f) 单击【确认】完成。

分析结果如表 6 与图 5 所示

III. 结果分析：

由于原始样本数据的限制，SPSS 给出警告，“至少有一个组的个案数不足两个”，所以此处无法进行价格的事后检验。

方差齐性的检验表（表 6）表明，不同厂商的汽车价格的方差齐性检验

的检验统计量的观测值为 4.502，概率 P-值为 0.000。如果显著性水平 α 为 0.05，由于概率 P-值小于显著性水平，因此拒绝原假设，认为不同厂商的汽车价格的总体方差存在显著差异，即不满足方差分析的前提要求。所以我们在此放弃不同厂商的价格统计的方差检验。

虽然该数据集无法满足方差分析的前提要求，但是我们仍可以在平均值图（图 5）中看出不同厂商的价格分层，如 3、8、10、16 号厂商（即 bmw、jaguar、mercedes-benz、porsche）明显处于高价区间，平均价格超过£20000 甚至£25000，而 4、5、6、7、15、19 号厂商（即 chevrolet、dodge、honda、isuzu、plymouth、subaru）明显处于低价区间，平均价格低于£10000。这与我们正常的产品认知相符，可以见得不同厂商的价格策略是比较清晰的，高价区间的厂商主要售卖豪车，赚取溢价和满足个性化差异需求；而低价区间的厂商主要靠性价比取胜，面向价格敏感型客户。

方差齐性检验

价格			
莱文统计	自由度 1	自由度 2	显著性
4.502	20	179	.000

ANOVA

价格					
	平方和	自由度	均方	F	显著性
组间	1.005E+10	21	478700876.6	33.232	.000
组内	2578454279	179	14404772.51		
总计	1.263E+10	200			

表 6 方差分析表

平均值图

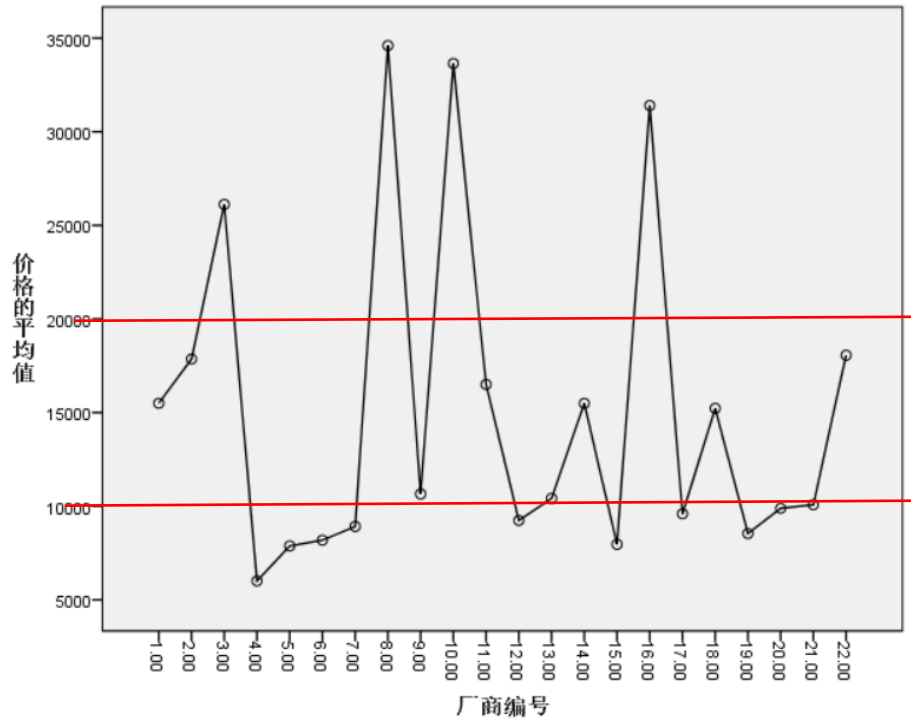


图5 价格平均值折线图

(2) 目的：从分析金额是否有大量异常值入手，分析不同品牌的产品线分布情况

I. 采用方法：描述性统计分析——描述

II. 操作步骤：

- 选择菜单：【分析】→【描述统计】→【描述】；
- 将“价格”变量选入【变量】框；
- 点击【将标准化值另存为变量】选择框；
- 单击【确认】完成；（此时生成了“价格”变量的标准化值“Zprice”）
- 选择菜单：【数据】→【个案排序】；
- 将“Zprice”变量选入【变量】框；
- 设置【排列顺序】为“升序”；
- 单击【确认】完成；
- 选择菜单：【转换】→【对个案中的值进行计数】；
- 设置【目标变量】为“outliers”，【目标便签】为“价格异常值”；

- k) 将“Zprice”变量选入【变量】框；
- l) 点击【定义值】按钮，选择范围为 $(-\infty, -3]$ 和 $[3, +\infty)$ ；
- m) 单击【确认】完成。

分析结果如图 6 所示

III. 结果分析：

我们发现所有样本中只有 4 个样本的“计数”值为 1，分别为 2 辆 mercedes-benz，一辆 bmw，一辆 porsche，且均为高端车型。可以见得相比于其他公司，这三家公司具备顶级豪车、超跑的产品线，具体见图 6。





	 make	 fuel_type	 Zprice	 outliers
182	volvo	gas	1.03084	.00
183	bmw	gas	1.03437	.00
184	bmw	gas	1.05094	.00
185	volvo	gas	1.09758	.00
186	porsche	gas	1.16300	.00
187	audi	gas	1.23312	.00
188	bmw	gas	1.47560	.00
189	mercedes-benz	diesel	1.50449	.00
190	mercedes-benz	diesel	1.92912	.00
191	mercedes-benz	diesel	1.94077	.00
192	bmw	gas	2.23593	.00
193	jaguar	gas	2.41881	.00
194	porsche	gas	2.45293	.00
195	mercedes-benz	diesel	2.48320	.00
196	porsche	gas	2.63703	.00
197	mercedes-benz	gas	2.65617	.00
198	mercedes-benz	gas	2.76320	.00
199	jaguar	gas	2.82383	.00
200	jaguar	gas	2.87906	.00
201	bmw	gas	2.98706	.00
202	porsche	gas	3.00523	1.00
203	mercedes-benz	gas	3.48781	1.00
204	bmw	gas	3.53138	1.00
205	mercedes-benz	gas	4.03275	1.00

图 6 价格异常值图

(3) 目的：以马力为核心指标分析不同品牌的性价比

假设汽车的目标消费群体只关注马力这个指标，需要基于马力比较不同品牌的性价比来决定是否购买该品牌的车辆。

I. 采用方法：描述统计——比率

II. 操作步骤：

- a) 选择菜单：【分析】→【描述统计】→【比率】；
- b) 将“马力”变量选入【分子】框；
- c) 将“价格”变量选入【分母】框；
- d) 将“厂商”变量选入【组变量】框；
- e) 点击【统计量】按钮，选择【均值】和【均值居中 COV】；
- f) 单击【确认】完成。

分析结果如表 7 所示

III. 结果分析：

在比率统计中，dodge、mercury、mitsubishi、plymouth 的比值最高(0.011)，mercedes-benz 的比值最低(0.04)，具体见表 7。

并且平均绝对离差(AAD)和离散系数(COD)总的情况为 0.002 和 0.207，基于均值和中位数的变异系数分别为 25.9%和 25.4%。相比较而言，chevrolet（雪佛兰）、bmw、mercedes-benz、porsche 的变异系数都明显低于平均水平，即离散程度低，AAD 和 COD 也可以证明这点。

总体上大致说明了，在汽车马力的领域，雪佛兰车型基本上生产的是性价比较高的车型，而 bmw、mercedes-benz、porsche 这些高端品牌的车辆性价比都相对较低。

马力/价格 的比率统计

组	平均值	平均绝对偏差	离差系数	差异系数	
				平均值居中	中位数居中
alfa-romero	.008	.001	.106	16.2%	16.0%
audi	.007	.001	.084	10.1%	10.4%
bmw	.005	.000	.083	11.5%	12.1%
chevrolet	.010	.001	.056	9.0%	9.4%
dodge	.011	.001	.087	11.2%	11.2%
honda	.010	.001	.088	11.5%	11.7%
isuzu	.010	.002	.171	24.1%	24.1%
jaguar	.006	.001	.142	20.8%	24.5%
mazda	.009	.002	.179	24.4%	24.4%
mercedes-benz	.004	.000	.045	6.5%	6.5%
mercury	.011	.000	.000	.	.
mitsubishi	.011	.001	.088	10.3%	10.3%
nissan	.010	.001	.100	12.3%	12.4%
peugot	.007	.001	.146	16.1%	18.9%
plymouth	.011	.001	.097	12.5%	12.4%
porsche	.006	.000	.048	6.5%	6.6%
saab	.008	.001	.079	11.1%	11.5%
subaru	.010	.001	.088	12.2%	12.8%
toyota	.009	.001	.141	18.4%	18.1%
volkswagen	.008	.002	.189	22.1%	22.8%
volvo	.007	.001	.171	20.5%	20.9%
总体	.009	.002	.207	25.9%	25.4%

表 7 车辆马力性价比的比率分析表

4、价格的影响因素：

(1) 目的：研究价格与高速油耗、马力之间是否存在关系

I. 采用方法：多因素方差分析

II. 操作步骤：

- 选择菜单：【分析】→【一般线性模型】→【单变量】；
- 将“价格”变量选入【因变量】框；
- 将“高速油耗”和“马力”变量选入【固定因子】框；
- 点击【选项】按钮，选择【方差齐性检验】；
- 单击【确认】完成。

分析结果如表 8、9 所示

III. 结果分析：

如表 8 所示，首先要进行方差齐性检验，因为方差齐性是多因素方差分析的前提，这里采用了方差同质性（Homogeneity of Variance）的检验方法，并且 $p=0.024<0.05$ ，所以拒绝 H_0 ，认为方差不相等。如果不同被试组的方差不齐性，也就是方差之比显著不等于 1，就说明被试之间原本就差异很大，那我们的方差分析就得不到准确的结论，不知道究竟是实验处理造成了不同被试组间的差异，还是说这里面也混淆了个体差异。

误差方差的莱文等同性检验 ^a			
因变量: 价格			
F	自由度 1	自由度 2	显著性
1.491	93	105	.024
检验“各个组中的因变量误差方差相等”这一原假设。			
a. 设计: 截距 + horsepower + highway_mpg + horsepower * highway_mpg			

表 8 方差齐性检验表

虽然不满足方差齐性前提，原则上不能进行方差分析，但 spss 里的方差分析是在最小二乘法的框架下做的，好处是这样的方差分析比较稳健，对于方差齐性的问题不敏感，即使违反了，也还是能用，结果也还是比较可信的。所以在这里勉强使用多因素方差来演示分析过程。

由表 9 所示，观测变量（价格）的总变差 SST 为 $1.26E+10$ ，它被分解为 4 个部分：

1. 马力不同引起的变差（2237657020）
2. 由高速油耗差异引起的变差（113440583.7）
3. 由马力和高速油耗交互作用引起的变差（65454369.53）
4. 由随机因素引起的变差（337867607.1）

这些变差除以各自的自由度后，得到各自的方差，并可计算出各 F 检验统计量的观测值和一定自由度下的概率 P-值。其中 F_{x1} ， F_{x2} ， F_{x1*x2} 的概率 P-值分别为 0.000, 0.046, 0.183。如果显著性水平 α 为 0.05，由于 F_{x1} ， F_{x2} 的 P-值小于显著性水平，所以应拒绝原假设，认为不同马力、高速油耗下的

价格总体均值存在显著差异，对价格的影响不同时为 0，各自不同的水平对价格有着显著影响。

同时，由于 $F_{x1 \times x2}$ 的 P-值大于显著性水平，因此不应拒绝原假设，不同马力和高速油耗没有对价格产生显著的交互作用，不会对价格产生显著影响，不应成为方差分析数学模型的一部分，而是合并到 SSE 中。所以此处可建立非饱和模型，具体操作步骤不再赘述。

另外，修正模型所对应的平方和（ $1.227E+10$ ）是 $x_1, x_2, x_1 \times x_2$ 对应变差相加的结果（ $1.227E+10=2237657020+113440583.7+65454369.53$ ），是线性模型整体对观测变量变差解释的部分，其对应的 F 检验统计量和 P 值说明，观测变量的变动主要是由控制变量整体的不同水平引起的，控制变量能较好地反映观测变量的变动，模型对观测变量有一定的解释能力。

R 方（0.973）和调整后的 R 方（0.949）反映的是多因素方差模型对观测数据的总体拟合程度，越接近 1 说明拟合程度更高。在该问题中有两个控制变量，所以应参考调整后的 R 方（0.949），可以看到该模型对数据的拟合程度比较理想。

主体间效应检验

因变量: 价格					
源	III 类平方和	自由度	均方	F	显著性
修正模型	1.227E+10 ^a	93	131900729.7	40.991	.000
截距	2.213E+10	1	2.213E+10	6877.038	.000
highway_mpg	113440583.7	21	5401932.557	1.679	.046
horsepower	2237657020	49	45666469.80	14.192	.000
highway_mpg * horsepower	65454369.53	15	4363624.635	1.356	.183
误差	337867607.1	105	3217786.734		
总计	4.751E+10	199			
修正后总计	1.260E+10	198			

a. R 方 = .973 (调整后 R 方 = .949)

表 9 价格方差分析表

(2) 目的：进一步探究价格与高速油耗、马力之间的关系
既然发现高速油耗和马力很有可能对于价格产生影响，那么本例则使用线性回归的方式来探究具体影响。

I. 采用方法：线性模型

II. 操作步骤 1（散点图）：

- 选择菜单：【图形】→【旧对话框】→【散点图/点图】→【矩阵散点图】；
- 将“高速油耗”、“马力”和“价格”变量选入【矩阵变量】框；
- 单击【确认】完成。

分析结果如图 7 所示

III. 结果 1 分析：

因为我们的目标是观察马力和高速油耗对于价格的影响，所以这里只关注价格与其他两个变量的关系（即红色方框中的散点图）。由矩阵散点图可以粗略的看出，价格与马力之间存在较强的正相关关系，价格与高速油耗之间存在较强的负相关关系。

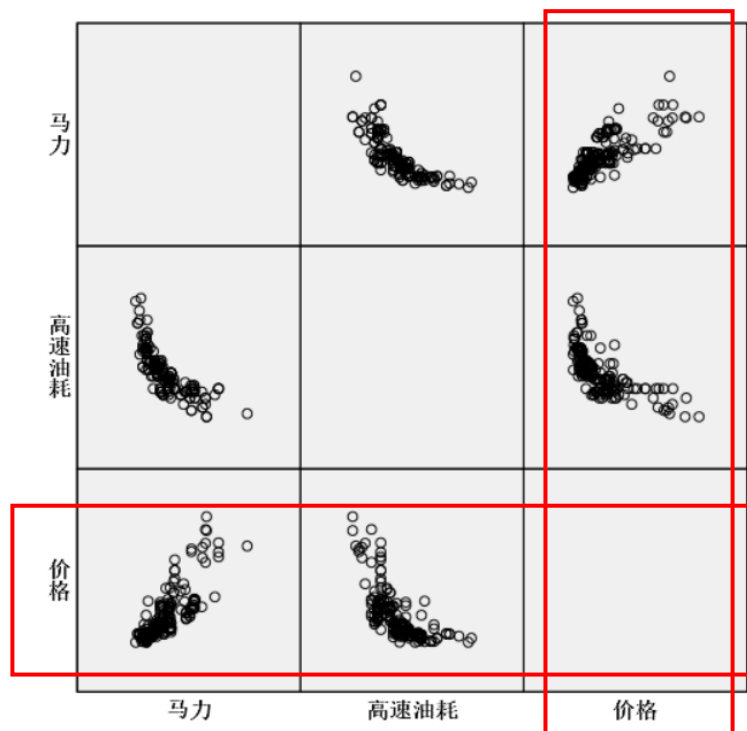


图 7 高速油耗、马力与价格的矩阵散点图

其次，相关分析是分析客观事物之间关系的数量分析法，由于收集到的马力、高速油耗、价格数据全部为定距数据，因此可通过计算 pearson 简单相关系数分析变量间线性相关性的强弱。

IV. 操作步骤 2（相关系数）：

- 选择菜单：【分析】→【相关】→【双变量】；
- 将“高速油耗”、“马力”和“价格”变量选入【变量】框；

- c) 设置【相关系数】为“皮尔逊”，设置【显著性检验】为“双尾”；
- d) 选择【标记显著性相关性】；
- e) 单击【确认】完成。

分析结果如表 10 所示

V. 结果 2 分析：

价格和马力、高速油耗之间的简单相关系数分别为 0.811、-0.705，说明价格和马力、高速油耗之间分别存在正的强相关性和负的较强相关性，其相关系数检验的概率 P-值近似为 0。因此，当显著性水平 α 为 0.05 或 0.01 时，应拒绝相关系数检验的原假设，认为两总体不是零相关。

相关性		马力	高速油耗	价格
马力	皮尔逊相关性	1	-.771**	.811**
	显著性（双尾）		.000	.000
	个案数	203	203	199
高速油耗	皮尔逊相关性	-.771**	1	-.705**
	显著性（双尾）	.000		.000
	个案数	203	205	201
价格	皮尔逊相关性	.811**	-.705**	1
	显著性（双尾）	.000	.000	
	个案数	199	201	201

** 在 0.01 级别（双尾），相关性显著。

表 10 相关系数表

最后，因为直觉认为这种相关性可能会受到转速峰值的影响。为此，可将转速峰值作为控制变量，进行偏相关分析。

相关性			马力	高速油耗	价格
控制变量	马力	相关性	1.000	-.804	.831
转速峰值		显著性（双尾）	.	.000	.000
		自由度	0	196	196
	高速油耗	相关性	-.804	1.000	-.716
		显著性（双尾）	.000	.	.000
		自由度	196	0	196
	价格	相关性	.831	-.716	1.000
		显著性（双尾）	.000	.000	.
		自由度	196	196	0

表 11 偏相关系数表

很有趣的是，在控制了转速峰值后，价格与马力、高速油耗的相关性都略有上升，所以更能证明了马力与高速油耗是影响价格的重要因素。

VI. 操作步骤 3（线性拟合）：

以高速油耗和马力作为自变量，价格作为因变量建立回归方程来判断是否马力和高速油耗是影响价格的重要因素。

- 选择菜单：【分析】→【回归】→【线性】；
- 将“价格”变量选入【因变量】框；
- 将“高速油耗”和“马力”变量选入【自变量】框；
- 设置【方法】为“步进”
- 点击【统计】按钮，选择【模型拟合】和【共线性诊断】；
- 点击【图】按钮，将“*ZRESID”选入【Y】框，将“*ZPRED”选入【X】框；并选择【直方图】和【正态概率图】；
- 单击【确认】完成。

分析结果如下所示

VII. 结果 3 分析：

SPSS 的逐步回归策略是通过逐一建立多个模型实现的。因为此处有两个解释变量，SPSS 需建立两个回归模型。依据逐步回归策略，第一个模型为一元线性回归模型，并在此基础上建立第二个模型（最终的分析结果）。

由表 11 可知，第一个模型是以马力为解释变量的一元线性回归方程。该

模型的判定系数为 0.657，回归方程的估计标准误为 4684.918。第二个模型是包含马力和高速油耗的二元线性回归方程，其判定系数增加至 0.665，且调整的判定系数也有所增加，回归方程的估计标准误减小。从拟合优度角度看，第二个模型的拟合效果更佳。

模型摘要 ^c				
模型	R	R 方	调整后 R 方	标准估算的误差
1	.811 ^a	.657	.655	4684.918
2	.815 ^b	.665	.662	4641.836

a. 预测变量: (常量), 马力
b. 预测变量: (常量), 马力, 高速油耗
c. 因变量: 价格

表 11 线性模型摘要表

表 12 是回归方程显著性检验结果。由表 12 可知，被解释变量（腰围）的总离差平方和 SST 为 1.260E+10。一元模型（第一个模型）的回归平方和（SSR）为 8280790190，剩余平方和（SSE）为 4323845281；二元模型（第二个模型）增加了一个解释变量，剩余平方和减少为 4223140908，回归平方和增大为 8381494563。对于二元模型，回归方程显著性检验的 F 统计量的观测值为 194.497，其对应的概率 P 值近似为 0。若显著性水平 α 为 0.05，因概率 P-值小于 α ，拒绝回归方程显著性检验的原假设，即回归系数不同时为 0，解释变量全体与被解释变量间存在显著的线性关系，选择线性模型具有合理性。

ANOVA ^a						
模型		平方和	自由度	均方	F	显著性
1	回归	8280790190	1	8280790190	377.284	.000 ^b
	残差	4323845281	197	21948453.20		
	总计	1.260E+10	198			
2	回归	8381494563	2	4190747281	194.497	.000 ^c
	残差	4223140908	196	21546637.29		
	总计	1.260E+10	198			

a. 因变量: 价格
b. 预测变量: (常量), 马力
c. 预测变量: (常量), 马力, 高速油耗

表 12 回归方程显著性检验表

表 13 中表明，对于第一个模型，因马力与价格的相关性高于高速油耗，所以首先进入模型得到一元线性回归方程。此时，马力的回归系数显著性检验的 t 统计量的观测值为 19.424，概率 P -值近似为零。当显著性水平 α 为 0.05 时，应拒绝回归系数检验的原假设，认为马力与价格有显著的线性关系，应保留在模型中。此时按照逐步回归策略，高速油耗尚未进入回归模型，被列在表 9—1 (d) 中。如果高速油耗被加入到第一个模型中（即建立二元模型，其回归系数显著性检验的 t 统计量的观测值和概率 P -值将为 -2.162 和 0.032（即线性关系显著），可以引入到第二个模型中。对于第二个模型，两者的回归系数显著性检验均显著，无应被剔除的解释变量，此时建模过程结束。多元线性回归模型中标准化回归系数用于比较解释变量对被解释变量的重要性大小。本例中，马力对价格的贡献大于高速油耗，这是个合理的结论。

进一步，表 13 还列出了多重共线性的相关计算结果。对于第二个模型，马力的容忍度为 0.353，方差膨胀因子为 2.836，存在一定的多重共线性。由于模型中只有马力和高速油耗两个解释变量，所以它们的容忍度和方差膨胀因子值都是相等的。

系数 ^a								
模型		未标准化系数		标准化系数	t	显著性	共线性统计	
		B	标准误差	Beta			容差	VIF
1	(常量)	-4562.175	974.995		-4.679	.000		
	马力	172.206	8.866	.811	19.424	.000	1.000	1.000
2	(常量)	3478.383	3842.630		.905	.366		
	马力	146.475	14.793	.689	9.902	.000	.353	2.836
	高速油耗	-175.340	81.105	-.151	-2.162	.032	.353	2.836

a. 因变量：价格

排除的变量 ^a							
模型	输入	Beta	t	显著性	偏相关	容差	共线性统计
						VIF	最小容差
1	高速油耗	-.151 ^b	-2.162	.032	-.153	.353	2.836

a. 因变量：价格

b. 模型中的预测变量：(常量), 马力

表 13 模型系数表

表 14 是多重共线性检验的特征值以及条件指数。对于第二个模型，最大特征值为 2.851，其余依次快速减小。第三列的各个条件指数均不大，可以认为存在较弱或中等程度的多重共线性。

共线性诊断^a

模型	维	特征值	条件指标	(常量)	方差比例	
					马力	高速油耗
1	1	1.940	1.000	.03	.03	
	2	.060	5.696	.97	.97	
2	1	2.851	1.000	.00	.00	.00
	2	.145	4.437	.00	.16	.05
	3	.005	24.675	1.00	.83	.95

a. 因变量：价格

表 14 多重共线性检验表

图 8 是残差的正态性图形结果。可以看到，各数据点基本上分布在中间对角线的两边，但是在残差的散点图（图 9），可以看到存在一些异常值和不均匀的分布，其中距离 0 轴越远表明离散越大。并且残差正态性的非参数检验（单样本 K-S 检验）结果（见表 12）表明拒绝原假设，即认为它与正态分布有显著差异。

回归 标准化残差的正态 P-P 图

因变量：价格

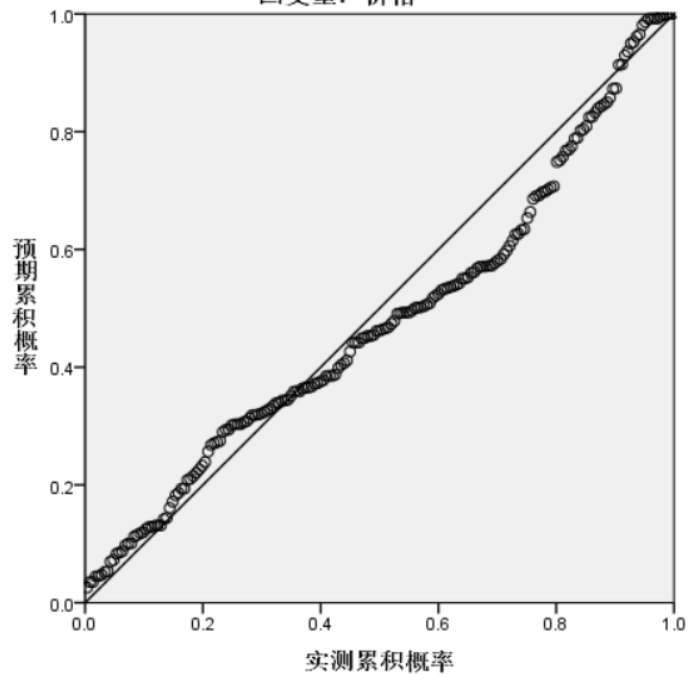


图 8 残差的正态性 P-P 图

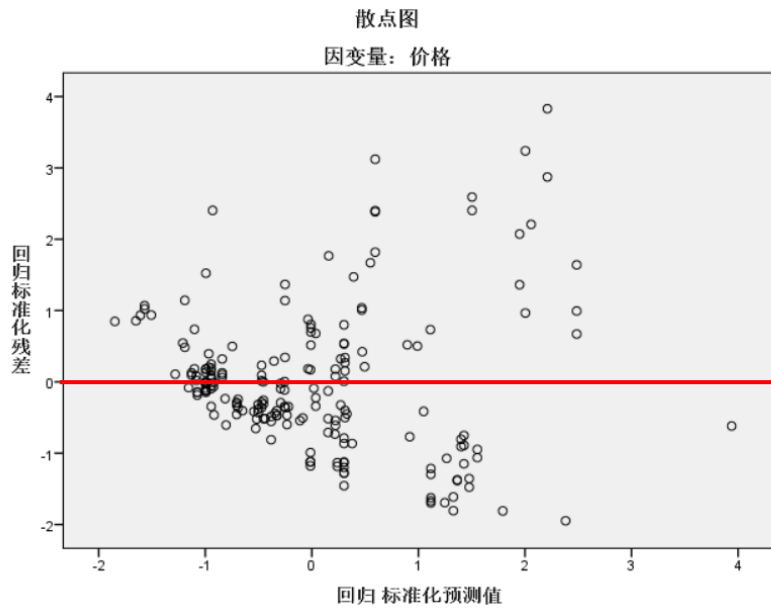


图 9 残差的正态性散点图

单样本柯尔莫戈洛夫-斯米诺夫检验

价格		
个案数		201
正态参数 ^{a,b}	平均值	13207.13
	标准差	7947.066
最极端差值	绝对	.154
	正	.154
	负	-.154
检验统计		.154
渐近显著性（双尾）		.000 ^c

a. 检验分布为正态分布。

b. 根据数据计算。

c. 里利氏显著性修正。

表 12 单样本 K_S 检验表

该模型拟合的回归方程为：

价格=3478.383+146.475×马力-175.340×高速油耗。方程表明，当高速油耗保持不变时，马力提高 1 匹，价格平均增加£146.475。当马力保持不变时，高速油耗增加 1mpg，价格平均减少£175.340。

(3) 目的：模型拓展（Python）

由于该模型拟合存在一定的多重共线性并且在加入“高速油耗”变量后偏相关系数进一步下降，所以为了增强线性模型的可靠性，这里采用 Python 引入多个变量以及多项式来深入探讨回归模型，如图 10-14。（具体代码及分析见文件 SPSS.ipynb 或者附录 4 中的数据链接）

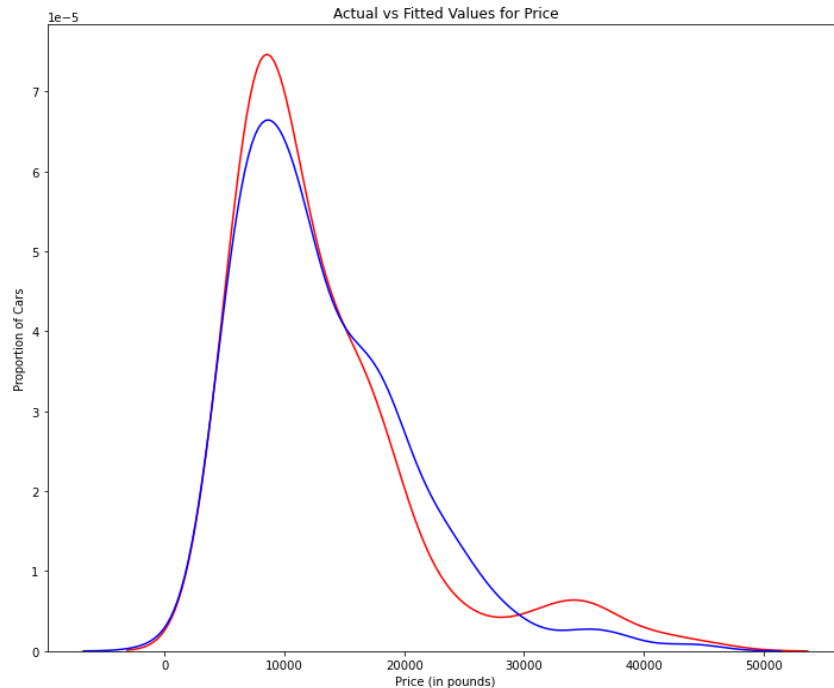


图 10 多元线性回归模型拟合图

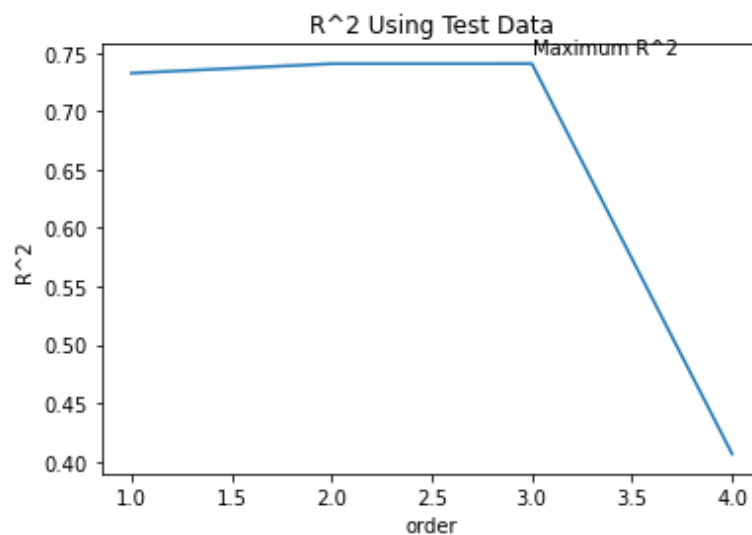


图 11 最优多项式次数测试图

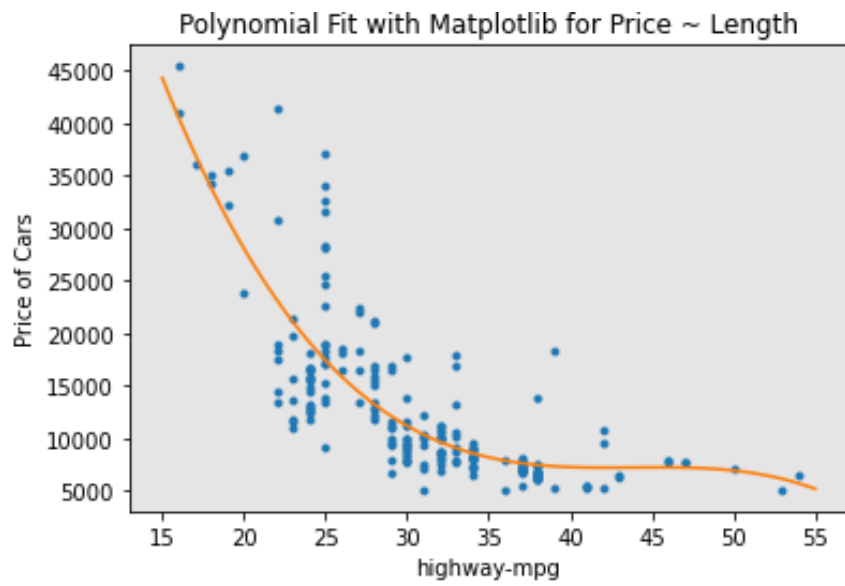


图 12 多项式回归模型拟合图

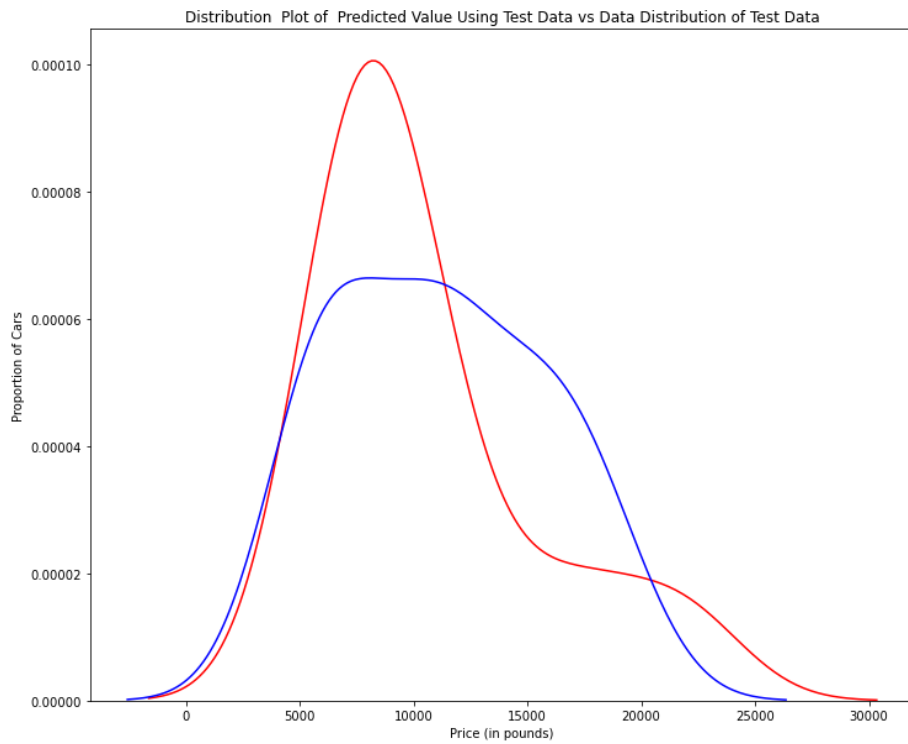


图 13 多元线性回归模型训练图

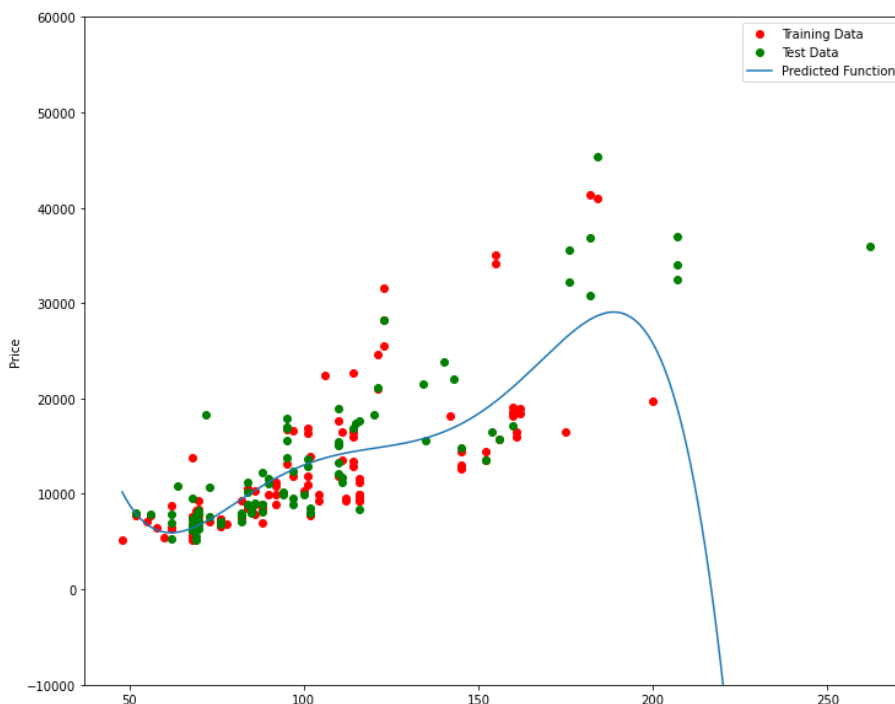


图 14 多项式回归模型训练图

结论

综上所述，在所有的模型训练中，多元线性回归模型是最优的模型，R方为0.8093562806577457,交叉验证评分为0.6645924739294804

最终的线性方程: $\text{Price} = -15678.742628061467 + 52.65851272 \times \text{horsepower} + 4.69878948 \times \text{curb-weight} + 81.95906216 \times \text{engine-size} + 33.58258185 \times \text{highway-mpg}$

图 15 结论图

四、 结论

根据此次对于汽车参数数据集的数据分析，可以得出以下一些结论：

- (1) 目前英国汽车市场的主流车型为 4 缸前驱，自然吸气，排量 2.0L，转速峰值 5500r 的四门家庭轿车，价位主要集中在£10000 左右。
- (2) 总体而言，除应用较少的两缸汽车以外，随着缸数的增加，汽车马力会呈现一定幅度的增加，油耗也会随之下降。
- (3) 调查中汽车厂商的产品策略均较为清晰，其中 bmw、jaguar、mercedes-benz、porsche 明显处于高价区间，并且其中 mercedes-benz, bmw, porsche 这三家厂商都具备顶级豪车的生产线；而 chevrolet、dodge、honda、isuzu、plymouth、subaru 的汽车则明显处于低价区间，主打低端家用车消费市场。
- (4) 作为一名汽车潜在购买者，若核心需求为汽车马力，则 dodge、mercury、

mitsubishi、plymouth 等汽车品牌性价比最高，并且雪佛兰（chevrolet）的大多车型均可满足目标需求。

- (5) 在本案例研究环境中，油耗和马力是影响汽车定价的重要因素。对价格敏感性较强的汽车购买者而言，可以参照“ $\text{价格} = -15678.742628061467 + 52.65851272 \times \text{马力} + 4.69878948 \times \text{车重} + 81.95906216 \times \text{引擎大小} + 33.58258185 \times \text{高速油耗}$ ”来衡量自己的目标车型的价格水平和本人的消费能力，进而提高自己的议价能力，挑选出价廉物美的产品。

附录

附录 1：变量类型及取值

编号	变量名	变量类型	取值	单位
1	make	字符串	alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo	
2	fuel-type	字符串	diesel, gas	
3	aspiration	字符串	std, turbo	
4	num_of_doors	字符串	four, two	
5	drive_wheels	字符串	4wd, fwd, rwd	
6	engine_location	字符串	front, rear	
7	engine_type	字符串	dohc, dohc, l, ohc, ohcf, ohcv, rotor	
8	num_of_cylinders	字符串	eight, five, four, six, three, twelve, two	
9	curb_weight	数字	continuous from 1488 to 4066	pound (磅)
10	engine_size	数字	continuous from 61 to 326	CID (立方英寸)
11	compression_ratio	数字	continuous from 7 to 23	无
12	horsepower	数字	continuous from 48 to 288	hp (匹)
13	peak_rpm	数字	continuous from 4150 to 6600	r (转)
14	city_mpg	数字	continuous from 13 to 49	mpg (加仑/英里)
15	highway_mpg	数字	continuous from 16 to 54	mpg (加仑/英里)
16	price	数字	continuous from 5118 to 45400	£ (英镑)

附录 2：源文件（import_data.csv）

附录 3：SPSS 文件（auto mobile pricing.sav）

	make	fuel_type	aspiration	num_of_doors	drive_wheels	engine_location	curb_weight	engine_type	num_of_cylinders	engine_size	compression_ratio	horsepower	peak_rpm	city_mpg	highway_mpg	price
1	mazda	gas	std	two	rwd	front	2380	rotor	2	70	9.40	101	6000	17	23	10945
2	mazda	gas	std	two	rwd	front	2380	rotor	2	70	9.40	101	6000	17	23	11845
3	mazda	gas	std	two	rwd	front	2385	rotor	2	70	9.40	101	6000	17	23	13645
4	mazda	gas	std	two	rwd	front	2500	rotor	2	80	9.40	135	6000	16	23	15645
5	chevrolet	gas	std	two	fwd	front	1488	l	3	61	9.50	48	5100	47	53	5151
6	alfa-romero	gas	std	two	rwd	front	2548	dohc	4	130	9.00	111	5000	21	27	13495
7	alfa-romero	gas	std	two	rwd	front	2548	dohc	4	130	9.00	111	5000	21	27	16500
8	audi	gas	std	four	fwd	front	2337	ohc	4	109	10.00	102	5500	24	30	13950
9	bmw	gas	std	two	rwd	front	2395	ohc	4	108	8.80	101	5800	23	29	16430
10	bmw	gas	std	four	rwd	front	2395	ohc	4	108	8.80	101	5800	23	29	16925
11	chevrolet	gas	std	two	fwd	front	1874	ohc	4	90	9.60	70	5400	38	43	6295
12	chevrolet	gas	std	four	fwd	front	1909	ohc	4	90	9.60	70	5400	38	43	6575
13	dodge	gas	turbo	two	fwd	front	2128	ohc	4	98	7.60	102	5500	24	30	7957
14	dodge	gas	turbo	?	fwd	front	2191	ohc	4	98	7.60	102	5500	24	30	8558
15	dodge	gas	std	two	fwd	front	1876	ohc	4	90	9.41	68	5500	37	41	5572
16	dodge	gas	std	four	fwd	front	1967	ohc	4	90	9.40	68	5500	31	38	6229
17	dodge	gas	std	two	fwd	front	1876	ohc	4	90	9.40	68	5500	31	38	6377
18	dodge	gas	std	four	fwd	front	1989	ohc	4	90	9.40	68	5500	31	38	6692
19	dodge	gas	std	four	fwd	front	1989	ohc	4	90	9.40	68	5500	31	38	7609
20	dodge	gas	turbo	two	fwd	front	2811	ohc	4	156	7.00	145	5000	19	24	12964
21	dodge	gas	std	four	fwd	front	2535	ohc	4	122	8.50	88	5000	24	30	8921
22	honda	gas	std	two	fwd	front	1837	ohc	4	79	10.10	60	5500	38	42	5399
23	honda	gas	std	two	fwd	front	1713	ohc	4	92	9.60	58	4800	49	54	6479
24	honda	gas	std	two	fwd	front	1940	ohc	4	92	9.20	76	6000	30	34	6529
25	honda	gas	std	two	fwd	front	1819	ohc	4	92	9.20	76	6000	31	38	6855

SPSS 数据截取图

名称	类型	宽度	小数位数	标签	值	缺失	列	对齐	测量	角色
make	字符串	13	0	厂商	无	无	7	左	名义	输入
fuel_type	字符串	6	0	燃料类型	无	无	6	左	名义	输入
aspiration	字符串	5	0	吸气	无	无	8	左	名义	输入
num_of_doors	字符串	4	0	门数	无	无	12	左	名义	输入
drive_wheels	字符串	3	0	驱动轮	无	无	10	左	名义	输入
engine_location	字符串	5	0	引擎位置	无	无	12	左	名义	输入
curb_weight	数字	4	0	车重	无	无	10	右	标度	输入
engine_type	字符串	5	0	引擎类型	无	无	6	左	名义	输入
num_of_cylinders	数字	6	0	缸数	无	无	14	右	标度	输入
engine_size	数字	3	0	引擎大小	无	无	7	右	标度	输入
compression_ratio	数字	5	2	压缩比	无	无	10	右	标度	输入
horsepower	数字	3	0	马力	无	无	9	右	标度	输入
peak_rpm	数字	4	0	转速峰值	无	无	8	右	标度	输入
city_mpg	数字	2	0	城市油耗	无	无	8	右	标度	输入
highway_mpg	数字	2	0	高速油耗	无	无	8	右	标度	输入
price	数字	5	0	价格	无	无	6	右	标度	输入
make_num	数字	8	2	厂商编号	无	无	10	右	有序	输入
Zprice	数字	11	5	Zscore: 价格	无	无	7	右	标度	输入
outliers	数字	8	2	价格异常值	无	无	7	右	名义	输入

SPSS 变量定义图

附录 4:

Python 文件 (SPSS.ipynb)

数据链接:

https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/e260b352-70da-4c1d-a2cf-1cb3df1ce4f2/view?access_token=adf94173075739e2d5c95adc40d94e680e261fcd05b65ce81362ac44abf35a8f

附录 5: 引擎大小单位换算表

ENGINE SIZE CHART

Liters (L)	Cubic Centimeters (CC)	Cubic Inches (CID)
1.0	1,000	61
1.5	1,500	92
1.6	1,606	98
1.7	1,721	105
1.8	1,836	112
2.0	1,983	121
2.0	2,000	122

引擎大小单位换算表

数据来源: <https://www.cjponyparts.com/resources/engine-size-chart>