Statistics and Application 统计学与应用, 2022, 11(2), 206-215
Published Online April 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.112022

Hans汉斯

基于多元线性回归分析方法的汽车油耗(MPG)预测模型

曲培元[1]，赵志斌[2,3]，陈　浩[2,3]，徐东昕[2,3]，刘国军[2,3]

[1]金景(海南)科技发展有限公司，海南 海口
[2]海南省院士团队创新中心，海南省激光技术与光电功能材料重点实验室，海南 海口
[3]海南师范大学物理与电子工程学院，海南 海口

收稿日期：2022年3月5日；录用日期：2022年3月31日；发布日期：2022年4月6日

摘　要

多元线性回归是回归分析中应用非常广泛的一种，当今的社会经济现象与人们的日常生活变化大多都受到多个因素的影响，所以在做分析计算开发时一般都要进行多元线性回归。本文利用多元线性回归模型，使用R语言开发实现了一个对汽车MPG预测的模型，该模型应用于统计预测汽车的续航能力，可进一步为客户提供一定的购车参考。

关键词

多元线性回归，汽车MPG，回归器，应用统计

Prediction Model of Vehicle MPG Based on Multiple Linear Regression Analysis

Peiyuan Qu[1], Zhibin Zhao[2,3], Hao Chen[2,3], Dongxin Xu[2,3], Guojun Liu[2,3]

[1]Hainan Brisight Science & Technology Co., Ltd., Haikou Hainan
[2]Key Laboratory of Laser Technology and Optoelectronic Functional Materials of Hainan Province, Hainan Academician Team Innovation Center, Haikou Hainan
[3]School of Physics and Electronic Engineering, Hainan Normal University, Haikou Hainan

Received: Mar. 5th, 2022; accepted: Mar. 31st, 2022; published: Apr. 6th, 2022

Abstract

Multiple linear regression is widely used in regression analysis. Today's socio-economic pheno-

文章引用: 曲培元, 赵志斌, 陈浩, 徐东昕, 刘国军. 基于多元线性回归分析方法的汽车油耗(MPG)预测模型[J]. 统计学与应用, 2022, 11(2): 206-215. DOI: 10.12677/sa.2022.112022

mena and changes in people's daily life are mostly affected by multiple factors, so it is generally necessary to carry out multiple linear regression when doing analysis, calculation, and development. Using multiple linear regression model, this paper develops and implements a model for automobile MPG prediction. The model is applied to statistically predict the endurance of automobile, which can further provide customers with a certain reference for car purchase.

Keywords

Multiple Linear Regression, Automobile MPG, Regressors, Applied Statistics

1. 引言

线性回归分析，也称为线性建模，需要将直线、平面或多项式拟合到数据中。与大多数机器学习算法一样，线性回归的目标是使用其他变量预测输出变量。线性回归将输出变量或因变量表示为自变量或预测变量的函数或线性组合。线性回归是一种广泛用于建模和理解现实世界现象的技术。它易于使用和直观理解。在简单线性回归中，模型只是一条直线，而对于多元回归，模型可以是多项式或平面；本文主要用到的是多元线性回归。多元线性回归模型是由多个自变量与因变量之间呈线性关系时，进行的回归分析设定的模型。多元线性回归模型的应用非常广泛，主要为建立最优多元线性回归方程再评定各个自变量对因变量的影响大小，并对有关模型的有效性进行检测[1]。汽车 MPG 意为汽车的油耗，是每加仑可以跑多少英里的一个数据，每加仑约等于 3.785 升，而每英里约等于 1.61 千米。衡量的是一辆汽车在一加仑燃油的情况下能走多远。汽车 MPG 的数据建模受到气缸数、加速度、马力、排量与重量等影响。在世界各地汽车使用非常普遍的情况下，消费者有时会考虑购买之前想购买的汽车的效率和燃油经济性。每个人都想买一辆能行驶很远、耗油更少的车。预测汽车 MPG 同样也是属于机器学习模块的一个问题，说到机器学习就不得不提到 python，该预测模型同样也可以使用 python 来实现，使用 python 来完成需要用到更多的算法，需要进行测试的模型也非常的多，比如 Adaboost、XGboost 和 GradientBoosting 等，最后还要对这些不同的模型进行各种方面的对比，比较耗时耗力，也容易在对比的过程中出现错误的判断。相比于 python，R 更适合对数据进行统计分析，在这个问题上则有着更简单的方法，作为专门为统计和数据分析开发的语言，R 语言在数据分析的输出上有着更直观的表现，虽然在处理过于巨大的数据面前对比 python 有着一定的劣势，但在对数据的深入分析能力方面也是 python 无法比拟的。R 语言拥有优秀的数据可视化工具，同时还有包罗万象的统计函数与相应的包可以直接使用调用。由于本文需要做的只是对少数的数据集进行分析建模，所以意在使用 R 语言通过分析上述提到的数据作为变量，生成一个良好的模型，该模型可以预测汽车的每加仑英里数，同时考虑到汽车的其他特性，使误差最小。使用的方法是将这些变量作为回归器，成为多元线性回归模型中的基本模型，进行多元统计分析，找出变量之间的依存关系与规律，从而建立多元线性回归模型[2]，来实现该汽车 MPG 预测模型的开发。

2. 方法

首先最基本的多元线性回归模型可以设为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

其中 y 为随机变量，$x_1, x_2, \cdots, x_n$ 为一般变量，本多元线性回归模型用到了六个不同的回归器，具体为 MPG，气缸数，马力，排量，加速度与重量。具体概念罗列如下：

MPG——英里数/每加仑；

气缸——汽车的动力装置，汽油在其中转化为动力；

排量——汽车的发动机排量；

马力——发动机性能的比率；

重量——汽车的重量；

加速度——汽车的加速度。

导入准备好的不同品牌的汽车的各种数据，我们要做的就是读入数据文件，合并它们并执行一些数据清理。在数据分析领域，数据清理是一件非常重要的事情。它需要检测、纠正或从数据集中删除不准确的记录。它可以提高数据质量，从而提高整体生产效率。当清理数据时，所有过时或不正确的信息都会消失，以便提供最高质量的信息。完成数据的导入后使用“pairs.r”来实现所有数值的散点图(图 1)。

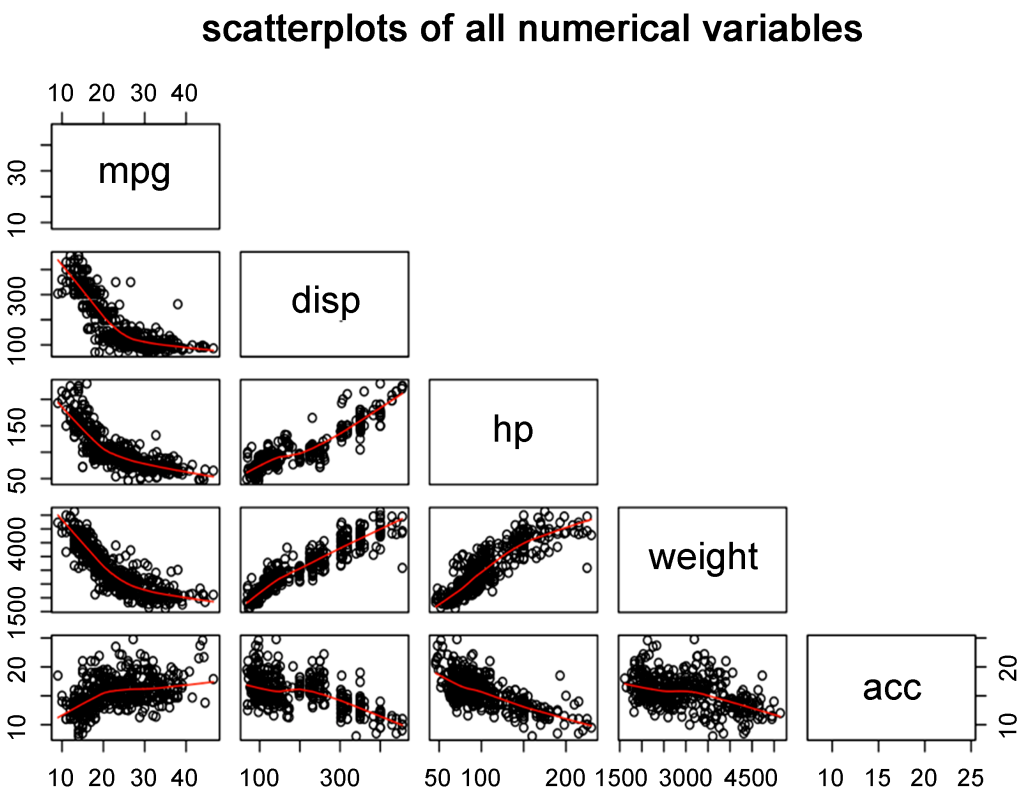

Figure 1. Scatter plot of all values
图 1. 所有数值的散点图

之后使用气缸，马力，排量，重量和加速度五个回归因子拟合多元线性回归模型，得出以下的数据如表 1，表 2。

输入：

lm(formula = MPG ~ as.factor(cyl) + hp + disp + weight + acc, data = autoMPG)

残差：

Table 1. Residuals data of multiple linear regression mode
表 1. 多元线性回归模型残差数据

Min	1Q	Median	3Q	Max
−9.6406	−2.5277	−0.4809	1.9960	16.1944

系数：

Table 2. Coefficients data of multiple linear regression model
表 2. 多元线性回归模型系数数据

	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	38.8797135	3.0568097	12.719	<2e−16
as.factor(cyl)4	7.2195952	2.1036081	3.432	0.000664
as.factor(cyl)5	9.2635725	3.1611670	2.930	0.003588
as.factor(cyl)6	3.4582042	2.3390418	1.478	0.140104
as.factor(cyl)8	6.0961268	2.7080015	2.251	0.024942
hp	−0.0673907	0.0163484	−4.122	4.60e−05
disp	−0.0008322	0.0086455	−0.096	0.923364
weight	−0.0043707	0.0007889	−5.540	5.63e−08
acc	−0.0828389	0.1197356	−0.692	0.489452

残差标准误差：4.003 on 383 degrees of freedom；

多重 R 平方：0.7424, Adjusted R-squared: 0.737；

F-统计量：138 on 8 and 383 DF, p-value: <2.2e−16。

得出上组模型数据后为了可以更清晰的得到预测结论，需要移除不显著的变量，这一数据可以定在低于 20%的等级，移除时从最不显著的变量依次移除，得到下表 3，表 4。

输入：

lm(formula = MPG ~ disp + hp, data = autoMPG)

残差：

Table 3. Model residuals data after removing insignificant variables (mod1)
表 3. 移除不显著变量后的模型残差数据(mod1)

Min	1Q	Median	3Q	Max
−11.3674	−3.1721	−0.4338	2.3349	16.4288

系数：

Table 4. Model coefficients data after removing insignificant variables of (mod1)
表 4. 移除不显著变量后的模型系数数据(mod1)

	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	37.469488	0.727716	51.489	<2e−16
disp	−0.040818	0.004963	−8.225	2.95e−15
hp	−0.058275	0.013491	−4.319	1.99e−05

残差标准误差：4.534 on 389 degrees of freedom；

多重 R 平方：0.6643, Adjusted R-squared: 0.6626；

F-统计量：384.9 on 2 and 389 DF, p-value: <2.2e−16。

可以将上述模型命名为 mod1，方便之后的计算与统计，再对这个 mod1 进行修改，开发出模型并命名 mod2，其中包括与气缸数的交互项，用来检查这些交互项中有无对预测有帮助的该模型的项。见表 5，表 6。

输入：

lm(formula = MPG ~ as.factor(cyl) + disp + hp, data = autoMPG)

残差：

Table 5. Modify mod1 to develop a model that add interaction terms with cyl residuals data (mod2)
表 5. 修改 mod1 后增加了与 cyl 交互项的模型残差数据(mod2)

Min	1Q	Median	3Q	Max
−9.4411	−2.6495	−0.6635	1.9999	17.8515

系数：

Table 6. Modify mod1 to develop a model that add interaction terms with cyl coefficients (mod2)
表 6. 修改 mod1 后增加了与 cyl 交互项的模型系数(mod2)

	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	30.189044	2.392397	12.619	<2e−16
as.factor(cyl)4	8.043303	2.194154	3.666	0.000281
as.factor(cyl)5	7.349540	3.308410	2.221	0.026900
as.factor(cyl)6	3.340332	2.454866	1.361	0.174405
as.factor(cyl)8	6.029386	2.850316	2.115	0.035043
disp	−0.025641	0.008108	−3.162	0.001690
hp	−0.078389	0.013644	−5.745	1.87e−08

残差标准误差：4.221 on 385 degrees of freedom；

多重 R 平方：0.7121, Adjusted R-squared: 0.7076；

F-统计量：158.7 on 6 and 385 DF, p-value: <2.2e−16。

可以看出排量、马力、重量和气缸之间存在强烈的负相关。这意味着，随着这些变量中的任何一个增加，MPG 降低。位移、马力、重量和气缸之间有很强的正相关性，这违反了线性回归的非多重共线假设。多重共线性妨碍了回归模型的性能和准确性。为了避免这种情况，必须通过特征选择来消除其中的一些变量。而其他变量加速度、模型和原点之间则没有高度的相关性。由于还需要更进一步的提高预测准确性，所以可以设置一个只有两个值的 mycyl，如下(表 7)。

Table 7. mycyl value used to change the mod2
表 7. 用来更改 mod2 的 mycyl 值

Mycyl	
0	1
206	186

这样就可以利用此表来更新 mod2，因为只有两个值，甚至用不到 as.factor()的计算方法。再依次排除掉不显著变量(低于 20%)，得到新的 mod2 模型。见表 8，表 9。

输入：

lm(formula = MPG ~ mycyl + disp + hp, data = autoMPG)

残差：

Table 8. Residuals data of new mod2 of after mycyl value change
表 8. mycyl 值更改后的新 mod2 模型残差数据

Min	1Q	Median	3Q	Max
−10.5285	−2.6283	−0.5148	1.9674	16.5964

系数：

Table 9. Coefficients of new mod2 of after mycyl value change
表 9. mycyl 值更改后的新 mod2 模型系数

	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	36.609582	0.703755	52.020	<2e−16
mycyl	−5.735708	0.875424	−6.552	1.81e−10
disp	−0.009384	0.006727	−1.395	0.164
hp	−0.082492	0.013340	−6.184	1.59e−09

残差标准误差：4.307 on 388 degrees of freedom；

多重 R 平方：0.6978, Adjusted R-squared: 0.6954；

F-统计量：298.6 on 3 and 388 DF, p-value: <2.2e−16。

通过观察 MPG 与数值变量的散点图，还考虑改变 mod1，包括多项式变量的数值变量。从 8 次多项式开始，根据输出中看到的内容清理模型。得到全新的 mod3，如下(表 10，表 11)。

输入：

lm(formula = MPG ~ disp + hp + poly(weight, 2), data = autoMPG)

残差：

Table 10. Residuals data of mod3 obtained by changing the numerical variable of polynomial variable from mod1
表 10. 由 mod1 更改多项式变量的数值变量后得出的模型 mod3 残差数据

Min	1Q	Median	3Q	Max
−11.7943	−2.3727	−0.4529	1.9596	15.4264

系数：

Table 11. Coefficients of mod3 obtained by changing the numerical variable of polynomial variable from mod1
表 11. 由 mod1 更改多项式变量的数值变量后得出的模型 mod3 系数

	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	30.243277	1.271720	23.781	<2e−16
disp	−0.004566	0.006239	−0.732	0.465

Continued

hp	−0.056568	0.012342	−4.583	6.18e−06
poly(weight, 2)1	−82.405941	11.391152	−7.234	2.53e−12
poly(weight, 2)2	27.526552	4.095624	6.721	6.48e−11

残差标准误差：4.019 on 387 degrees of freedom；

多重 R 平方：0.7376, Adjusted R-squared: 0.7349；

F-统计量：271.9 on 4 and 387 DF, p-value: <2.2e−16。

得出 mod3 后再通过包含与 mycyl 的交互项来修改上述模型。需要做的是减少多项式的次数，直到大多数项变得重要，把这个模型称为 mod4。见表 12，表 13。

输入：

lm(formula = MPG ~ mycyl + disp + hp + poly(weight, 2), data = autoMPG)

残差：

Table 12. Residuals data of mod4 after modification containing the interactive items with mycyl
表 12. 包含与 mycyl 的交互项修改后 mod4 残差数据

Min	1Q	Median	3Q	Max
−11.2382	−2.2908	−0.4385	1.8893	15.5598

系数：

Table 13. Coefficients of mod4 after modification containing the interactive items with mycyl
表 13. 包含与 mycyl 的交互项修改后 mod4 系数

	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	30.150451	1.253435	24.054	<2e−16
mycyl	−3.133488	0.884613	−3.542	0.000445
disp	0.009615	0.007337	1.311	0.190772
hp	−0.067839	0.012571	−5.396	1.19e−07
poly(weight, 2)1	−77.039910	11.326679	−6.802	3.95e−11
poly(weight, 2)2	21.516735	4.377976	4.915	1.32e−06

残差标准误差：3.96 on 386 degrees of freedom；

多重 R 平方：0.7458, Adjusted R-squared: 0.7426；

F-统计量：226.6 on 5 and 386 DF, p-value: <2.2e−16。

得出四个模型后，就可以比较它们之间的 adjR2 and s 两个数值，得出下表 14。

Table 14. Comparison of two values adjR2 and s between mod1, 2, 3, 4
表 14. mod1、2、3、4 之间 adjR2 和 s 两个数值的对比

$adjR^2$	0.6626	0.7076	0.7349	0.7426
s	4.534	4.221	4.019	3.96

通过对比，根据简约原则，首先排除 mod3 与 mod4，对比 mod1 与 mod2 有更多的数据计算项，在

计算中并不是不可或缺选项，所以排除；mod1 与 mod2 相对比，mod1 的误差值要比 mod2 更大，所以对比后可以选出 mod2 为更合适的模型。

3. 结果与检测

通过对比，根据简约原则可以选出 mod2 为更合适的模型。还需要对选定的模型进行检测，检查假设是否存在正态性和恒定方差冲突，得到如下输出图 2，图 3。

```
##
##  studentized Breusch-Pagan test
##
## data:  mod2
## BP = 23.326, df = 3, p-value = 3.454e-05
```

Figure 2. Bptest detection output
图 2. Bptest 检测输出

```
##
##  Shapiro-Wilk normality test
##
## data:  mod2$residuals
## W = 0.96482, p-value = 4.281e-08
```

Figure 3. Shapiro.test detection output
图 3. Shapiro.test 检测输出

最后进行异常值与影像的检测，使用 influencePlot()从数据中检测出杠杆点，需要检测的有三项，分别是数据中的杠杆点但不为影响点，不具影响力的影响点与具有影响力的影响点。得到图 4 与表 15 的输出。

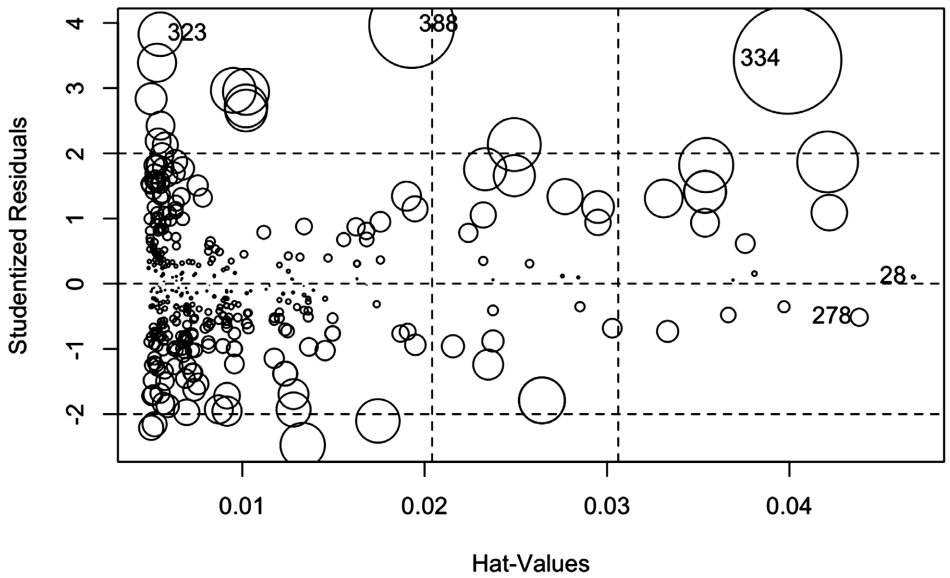

Figure 4. Detect leverage points and influential points detection output
图 4. 杠杆点与影响点检测输出

Table 15. Detect leverage points and influential points detection output
表 15. 杠杆点与影响点检测输出

	StudRes	Hat	CookD
28	0.1029127	0.046800957	0.0001303341
278	−0.5154153	0.043826940	0.0030498747
323	3.8274994	0.005507743	0.0195941641
334	3.4328540	0.039899753	0.1191235887
388	3.9637792	0.019292656	0.0744473346

再用 vif()检测出模型的共线性，查看是否有 vif 大于 10 的条件，小于 10 表示不具有多重共线性，而大于 10 则表示具有较强的多重共线性，如果只有少数大于 10 则不会受到影响，但多数大于 10 且数值偏高则证明会对该模型的预测有着较大的影响；得到输出表 16。

Table 16. Collinearity detection
表 16. 共线性检测

mycyl	disp	hp
4.037366	10.442030	5.556574

最后可以看出异常点图 4 还是有着相对少的异常点与影响点，而气缸数和马力的 vif 都小于 10，所以都不存在多重共线性，只有排量的 vif 大于 10，但也只是刚刚超过，存在较强的多重共线性但也对预测没有较大的影响，相比于其他的 mod1、3、4，mod2 仍然是最为适合的模型。

4. 结论

本文根据多元线性分析方法通过多个自变量形成多元线性回归模型，用到 R 语言实现四个 mod 并进行对比，最终根据简约原则选定并进行检测，确保将误差与影响降到最低。需要注意的是在分析模型的预测中需要依次排除不显著的变量，并要注意各个变量之间的正相关性与负相关性，如果违背了线性回归中的非多重共线假设，那同样需要对一些变量进行排除。多元线性分析在日常生活中的应用非常广泛，可以在各个领域进行分析预测，比如在汽油辛烷值损失的预测上，同样需要降维后进行多元线性回归模型的建立[3]，甚至在热带大气研究中也有着非常重要的应用，可以描述出大气在不同季节的移动过程，未来甚至在气象学中得到广泛的应用[4]。该模型也是机器学习中比较重要的一个学习领域[5]，在未来大数据人工智能领域是一个非常有发展性与应用性的预测模型。

基金项目

海南省重点研发项目(ZDYF2020020)。

参考文献

[1] 冷建飞, 高旭, 朱嘉平. 多元线性回归统计预测模型的应用[J]. 统计与决策, 2016(7): 82-85. https://doi.org/10.13546/j.cnki.tjyjc.2016.07.021

[2] 郭娟. 基于存在交互项的多元线性回归汽车油耗预测模型[J]. 广西质量监督导报, 2019(11): 138-140.

[3] 徐宗煌. 基于多元线性回归分析的汽油辛烷值损失预测建模[J/OL]. 宁夏大学学报(自然科学版), 2022: 1-8. http://kns.cnki.net/kcms/detail/64.1006.N.20220124.1703.030.html, 2022-03-30.

[4] Roundy, P.E. and Frank, W.M. (2004) Applications of a Multiple Linear Regression Model to the Analysis of Rela-

tionships between Eastward-and Westward-Moving Intraseasonal Modes. *Journal of the Atmospheric Sciences*, **61**, 3041-3048. https://doi.org/10.1175/JAS-3349.1

[5] Maulud, D. and Abdulazeez, A.M. (2020) A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, **1**, 140-147. https://doi.org/10.38094/jastt1457