VITU ACADEMIC
PUBLISHING PTE.LTD

四种模型预测结果的对比分析

邹　齐
北方工业大学　北京　100144

摘　要：两个多世纪以来，汽车工业的发展导致了燃料价格的上涨和消费者对汽车功能的需求增加。制造商努力优化工艺以提高燃油效率。本文对多元线性回归、贝叶斯多元线性回归（GLS）、广义最小二乘法、随机森林和Lasso回归等预测模型进行了比较和分析，以确定基于气缸数量、排量、马力、重量、加速度、年份和产地的最准确的MPG预测模型，并将进一步选择对MPG影响最大的变量。

关键词：多元线性回归；贝叶斯多元线性回归；随机森林模型；Lasso回归

引言

尽管在热效率方面取得了进步，但整体车辆设计和使用模式对燃油经济性产生了重大影响。由于不同的测试方案，燃油经济性数据在全球范围内有所不同。本研究通过缸数、排量、马力、重量、加速度、年份和产地来评估四种预测MPG的模型。

一、方法

（一）多元线性回归模型

在前几章中，通过相关分析、描述性统计分析和特征选择，我们发现使用多元线性回归模型预测我们选择的自变量和因变量的MPG是合适的。多元线性回归模型可表示为：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p+\varepsilon \quad (1)$$

在式中，Y为因变量（MPG），X_1，X_2，…，X_p是自变量（如对数马力，对数重量，圆柱体4，圆柱体8，原点3，年份），β_0，β_1，β_2，…，β_p是系数，而ε是误差项。

（二）贝叶斯多元线性回归模型

在前面的分析中，通过Durbin-Watson检验可以发现变量之间存在自相关，这会影响回归模型的准确性和稳定性。为了解决多重共线性问题，增强回归模型的稳定性和预测能力，本节采用贝叶斯多元线性回归。

贝叶斯多元线性回归采用正则化来防止过拟合，从而提高回归模型的稳健性。具体来说，贝叶斯回归为回归系数β引入正态分布，并根据数据计算后验分布。

假设线性回归模型：

$$y=X\beta+\varepsilon \quad (3)$$

其中，X为自变量矩阵，β为回归系数向量，λ为误差项。在贝叶斯回归中，我们假设回归系数β服从正态先验分布：

$$\beta \square N(\mu_0, \Sigma_0) \quad (4)$$

先验分布的选择可以基于先验知识或假设，有效地将先验知识纳入回归模型。

（三）随机森林

随机森林模型的预测公式如下：

$$\hat{y}=\frac{1}{N}\sum_{i=1}^{N}f_i(x) \quad (10)$$

其中，y为最终预测，N为决策树的数量，$f_i(x)$为输入x的第i棵决策树的预测。

（四）Lasso回归

Lasso回归通过向目标函数添加L1正则化项来实现正则化。其数学表达式为：

$$\min_{\beta}\frac{1}{2n}\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2+\lambda\sum_{j=1}^{p}|\beta_j| \quad (11)$$

式中，y_i为观测响应变量，x_{ij}为第i次观测的第j个预测变量，β_j为预测变量相关系数，β_0为截距，λ为控制收缩量的正则化参数。

二、实验

（一）多元线性回归模型训练结果

从图1和表中可以看出，虽然大多数数据点都接近于理想的预测，但整体预测精度低于其他模型，RMSE最低。

（二）多元线性回归分析

1.回归系数显著性检验

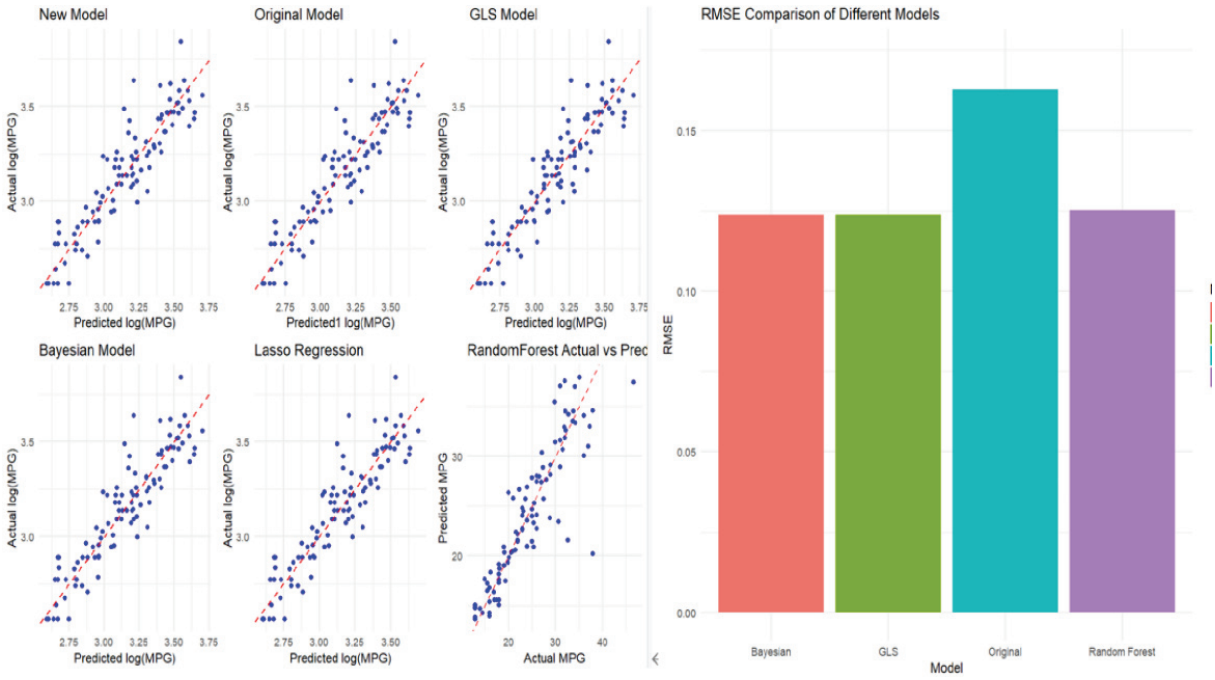

图1　实际值与预测值和RMSE的比较

表格　t检验结果

变量	P-值<0.05	是否拒绝假设
log displacement	No	Do not reject H_0
log horsepower	Yes	Reject H_0
log weight	Yes	Reject H_0
cylinders 4	Yes	Reject H_0
cylinders 6	No	Do not reject H_0
cylinders 8	No	Do not reject H_0
origin 1	No	Do not reject H_0
origin 2	No	Do not reject H_0
year	Yes	Reject H_0

t检验结果表明，马力，重量，气缸（4）和年份显著影响MPG。根据t检验结果，我们剔除不显著变量，重构回归模型，只保留对因变量有显著影响的变量。

（三）贝叶斯多元线性回归模型训练结果

贝叶斯模型的RMSE为0.1237163，表现较好。

（四）随机森林模型训练结果

随机森林模型的RMSE为0.1251177。随机森林模型也显示出较高的预测精度，虽然其RMSE略高于贝叶斯模型、GLS模型和Lasso回归，但仍在可接受的范围内。这表明，虽然随机森林模型在RMSE方面不是最好的，但它的性能对于预测目的仍然是鲁棒和可靠的。

（五）Lasso模型训练结果

LASSO回归模型在所有评估模型中表现最好，RMSE为0.1230052。这表明LASSO模型提供了最准确的预测。预测精度的显著提高是由于LASSO执行变量选择的能力，通过将一些系数缩小到零，有效地减少了多重共线性的影响。这导致了一个更简单、更可解释的模型，并保持了较高的预测能力。实验结果表明，原木马力、原木重量、4缸、8缸、3缸和年份等特征对MPG预测有显著影响。（见下图2）

结论

在本研究中，使用了五个模型来预测MPG。结果表明，对数马力、对数重、缸4、缸8、产地3和年份对MPG的预测有显著影响，Lasso回归模型是最有效的预测模型。这些变量在帮助汽车制造商优化设计、提高燃油效率和经济性以及为消费者提供更高效的汽车方面发挥了至关重要的作用。

参考文献

[1]Shirbhayye V, Kurmi D, Dyavanapalli S, et al. An accurate prediction of MPG (Miles per Gallon) using linear regression model of machine learning[C]//2020 International Conference on Computer Communication and Informatics

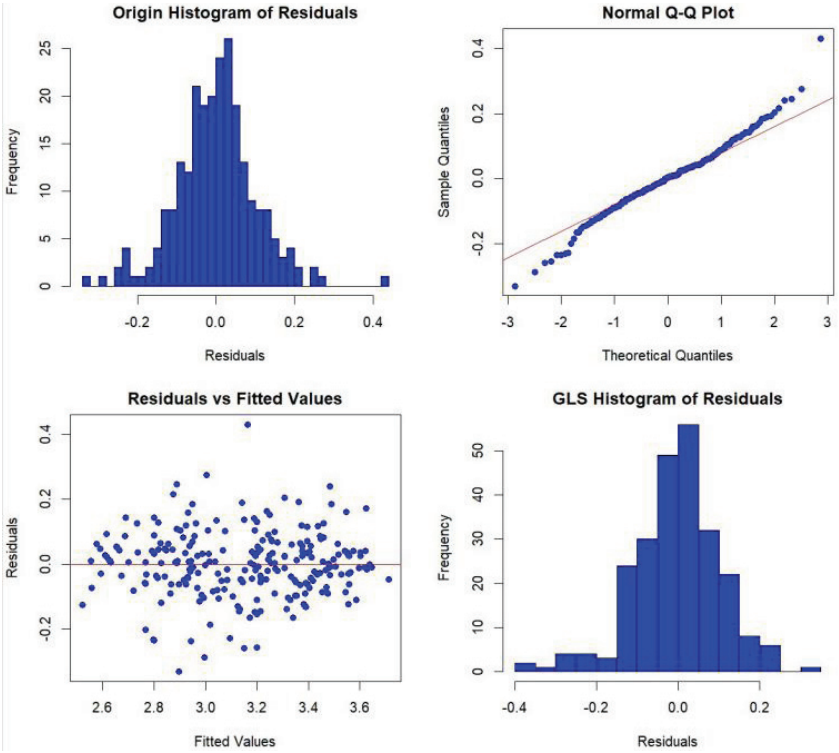

图2　残差分析

(ICCCI). IEEE, 2020: 1-5.

[2]Jamala M N, Abu-Naser S S. Predicting MPG for automobile using artificial neural network analysis[J]. International Journal of Academic Information Systems Research (IJAISR), 2018, 2(10): 5-21.

[3]Meng J, Liu X. MPG prediction based on BP Neural Network[C]//2006 1ST IEEE Conference on Industrial Electronics and Applications. IEEE, 2006: 1-3.

[4]Syahputra R. Aapplication of neuro-fuzzy method for prediction of vehicle fuel consumption[J]. Journal of Theoretical & Applied Information Technology, 2016, 86(1).