南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

统计学习基础课程项目

题　　目：汽车定价数据的回归分析

姓　　名：康国钰 11611706

张　昀 11610606

专　　业：数学系统计专业

2019 年 8 月 5 日

目录

一、 背景介绍

随着经济的快速发展，人们生活水平也在不断提升，对出行的的要求越来越高，汽车成为当今不可或缺的出行工具。汽车的需求量的大幅上涨，会导致汽车行业竞争愈加激烈，研究汽车定价相关的因素，一方面有利于促进汽车企业技术革新，制造出更满足消费者需求的产品；另一方面有利于消费者在购车时更准确方便的了解汽车市场的行情。研究汽车定价

二、 数据介绍

2.1 数据来源

本文研究汽车价格影响因素所使用的数据来自“UCI Machine Learning Repository”网站(https://archive.ics.uci.edu/ml/datasets/Automobile)，数据包括 1985 年沃德汽车年鉴数据的进口车辆规格、操作手册和保险碰撞报告。

原始数据集包含 204 条数据，26 列变量，25 列含连续型和分类型的自变量，第 26 列为连续型的汽车价格数据。本报告使用的数据纳入 16 个自变量，包括所有连续型变量（共 14 个），及 2 个分类变量。此处人为选择分类变量，是基于分类变量的含义并结合评价汽车性能的指标而确定。假设未纳入研究的定性变量对表征汽车性能好坏的作用较小，如“车的品牌”、“车门数目”、“汽油类型”、“发动机位置”等，默认留下的自变量均是有助于表征汽车性能的评价指标。

2.2 变量含义

2.2.1 因变量

因变量是车辆价格，连续型变量，取值范围是 5118 -- 45400 美元。

2.2.2 自变量

自变量包括共 16 个变量，其中包括 14 个连续型变量，2 个分类变量，大致分为描述汽车驾驶性能，安全性，操控性,乘坐舒适性,动力性，燃油经济性等方面。16 个变量含义如下（参考来自百度词条）：

指标	变量	含义	取值范围
安全性指标	评估风险等级（symboling）	以汽车出厂评估的风险等级为基准，向上为危险系数增加，向下为安全系数增加	-3,-2,-1,0,1,2,3
	年均车辆赔付（normalized-losses）	表示排除车型影响，汽车年均赔付的金额	65 - 256
操控性/舒适性指标	车身长度（length）	车前后保险杆最凸出的两点之间的距离	141.1 -- 208.1
	车身宽度（width）	车身左右最凸出位置的距离，不包含后视镜伸出的宽度	60.3 -- 72.3
	轴距（wheel base）	前后车轮轴之间的距离。轴距短，灵活性好，轴距长，直线行驶稳定性好	86.6 -- 120.9
	车身高度（height）	从地面算起到车身顶部最高的位置，不包括天线长度	47.8 -- 59.8
	整备重量（curb-weight）	汽车完全装备的空车质量，包括燃料，润滑油，冷却液，随车工具，备用轮胎及备品等的质量，不包括乘客和货物的重量。	1488 -- 4066
动力性指标	引擎尺寸（engine-size）	发动机大小	61 -- 326
	缸径（bore）	汽缸本体上用来让活塞做运动的圆筒空间的直径	2.54 -- 3.94
	冲程（stroke）	活塞在汽缸本体内运动时的起点与重点的距离。	2.07 -- 4.17
	压缩比（compression ratio）	内燃机气缸中最大和最小体积之比。压缩比高，发动机最大功率越大，最高车速越高，驱动力越强。	7 -- 23
	马力（horse power）	功率单位，1 马力=745.7 瓦特	48 -- 288
	最大转速（peak-rpm）	发动机每分钟最大转速	4150 -- 6600
车型	车体类型（body size）	硬顶，车皮，四门轿车，两厢车，敞篷车 (hardtop, wagon, sedan, hatchback, convertible)	5 种
燃油经济性	城区油耗（city-mpg）	城市路中，每消耗一加仑燃油，汽车行驶的英里数	13 -- 49
	高速路油耗（highway-mpg）	高速路中，每消耗一加仑燃油，汽车行驶的英里数	16 -- 54

安全性指标：车辆进行碰撞试验评估的危险系数和年均理赔金额表征安全性能。

操控性指标：长宽较大的车型拥有较为宽敞的车内空间，乘坐舒适性能高，但是降低在狭窄街道中的行驶灵活性。车身低，运动性能高，车身高，乘客可以在车内活动。轮距宽，行驶稳定性好，操纵性能好。

动力性指标：压缩比越大，发动机排量越大，最大功率越大，动力性越好，最高车速越高，汽车在行驶过程中不进行换挡加速和爬坡的能力就越大。

燃油经济性：衡量汽车是否省油，一加仑能行驶的里程数越多，说明油耗越低。

车型：

hardtop

wagon

sedan

hatchback

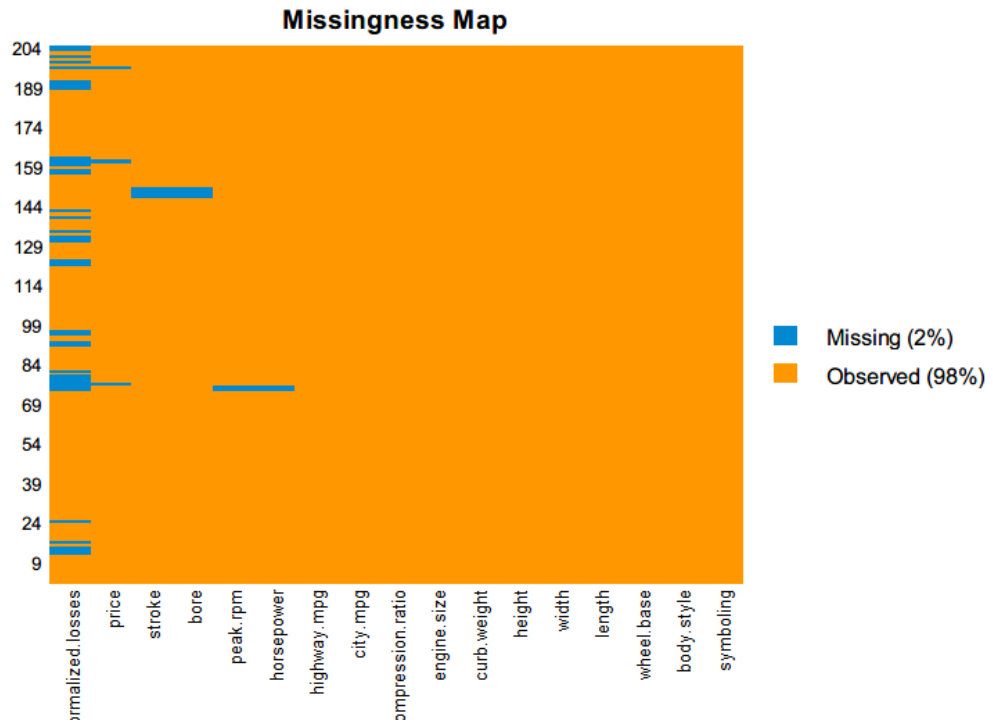
convertible car

三、 数据预处理：

3.1 缺失值

3.1.1 缺失值情况

响应变量汽车价格中存在 4 条缺失值，占原始数据 2%。 16 个自变量中含有缺失值的变量包括相对年均车辆赔付(normalized-losses),缸径（bore），冲程（stroke）， 最大转速（peak-rpm），马力（horsepower），分别占原始 204 条总数据的 19.6%，1.96%，1.96%，0.98%，0.98%。 下图显示各变量的含有缺失值情况：



3.1.2 缺失值处理

1、删除：由于响应变量汽车价格，自变量孔数，里程，最大转速，马力缺失值占比较小，选择直接删除对应的观测值，此项操作共删除了 10 条观测值,有效观测值从 204 变为 194。

2、多重插补：使用多重插补法对“年均车辆索赔”变量所含有的缺失值进行补充。首先，该变量含有 19.6%缺失值，占比较大，直接去掉会导致信息受损；此外，“年均车辆赔付”表征某款车型每年由于事故等报修理赔的费用，并不是硬性的机械指标，所以考虑采用贝叶斯线性回归的多重插补法，利用后验分布作为缺失值在已有观察值下的估值。

经过上述两步缺失值处理之后，共有 194 条数据。

3.2 标准化

由第二部分“变量含义”变量的取值范围可知，变量之间取值大小悬殊。为去除量纲和数量级影响，需对样本数据做标准化处理。

1. 数据划分： 本报告建立的模型，将从模型拟合和预测两个方面进行建模和评估。因此预先对上述处理过缺失值的 194 条数据进行划分，其中 80%（共 156 个）用于训练集数据（training data），20%（共 38 个）测试集数据(testing data)。

2. 训练和预测数据的标准化：对 80%训练集数据进行标准化，20%测试集的标准化使用前面 80%训练集数据的均值标准差进行标准化。

3.3 多重共线性

通过第二部分数据“变量含义”，可以发现描述汽车规格的多个变量之间有可能存在共线性，比如变量“城区油耗”和“高速路油耗”，可以理解为车辆在城区路消耗 1 加仑跑的公里数越多，相应在高速路上消耗 1 加仑跑的公里数也会越多。自变量之间很强的线性关系会出现回归方程显著但是回归系数不能通过显著性检验。

直接对 16 个变量进行线性回归结果如下：

R-squared	Adjusted R2	P-value	显著的变量
			***: body.style, engine.size
0.8825	0.862	< 2.2e-16	**: height, stroke, horsepower
			*: Intercept,normalized.losses,compression.ratio

从R^2 和 p-value 可知，回归方程通过显著性检验，但是回归系数显著的变量数目较少。

3.3.1 共线性检验

1. 相关系数矩阵

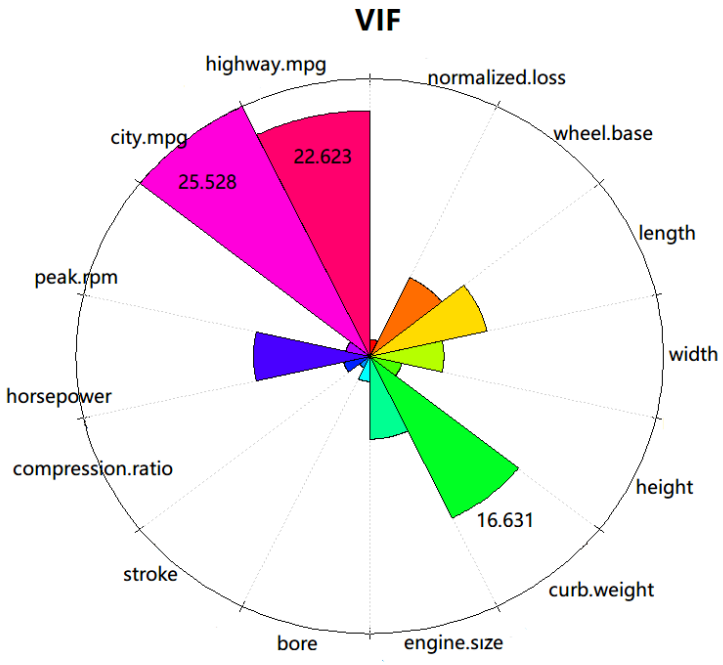
下图表示连续变量之间的相关系数矩阵，矩阵对称只显示下三角矩阵，其中相关系数绝对值高于 0.6 的关系由红色标出。可见操控性和动力性指标内的变量（车宽与轴距，车辆尺寸与整备质量等）存在高度相关，燃油经济性的两个变量（城区油耗与高速路油耗）存在高度相关（相关系数 0.969）。

	normalized.losses	wheel.base	length	width	height	curb.weight	engine.size	bore	stroke	compression.ratio	horsepower	peak.rpm	city.mpg	highway.mpg
normalized.losses	1													
wheel.base	0.036	1												
length	0.119	0.884	1											
width	0.179	0.818	0.863	1										
height	-0.372	0.598	0.510	0.309	1									
curb.weight	0.233	0.779	0.881	0.880	0.312	1								
engine.size	0.300	0.548	0.678	0.748	0.016	0.852	1							
bore	0.113	0.517	0.636	0.570	0.192	0.684	0.622	1						
stroke	0.068	0.151	0.082	0.175	-0.078	0.143	0.157	0.001	1					
compression.ratio	-0.216	0.295	0.202	0.223	0.279	0.209	0.045	0.047	0.218	1				
horsepower	0.400	0.338	0.558	0.608	-0.124	0.740	0.843	0.609	0.049	-0.198	1			
peak.rpm	0.222	-0.380	-0.310	-0.292	-0.287	-0.308	-0.222	-0.271	-0.118	-0.488	0.117	1		
city.mpg	-0.365	-0.490	-0.681	-0.656	-0.088	-0.766	-0.722	-0.625	-0.003	0.289	-0.839	-0.076	1	
highway.mpg	-0.322	-0.564	-0.716	-0.707	-0.144	-0.812	-0.743	-0.629	-0.017	0.210	-0.811	-0.012	0.969	1

2. VIF 方差膨胀因子（VIF）

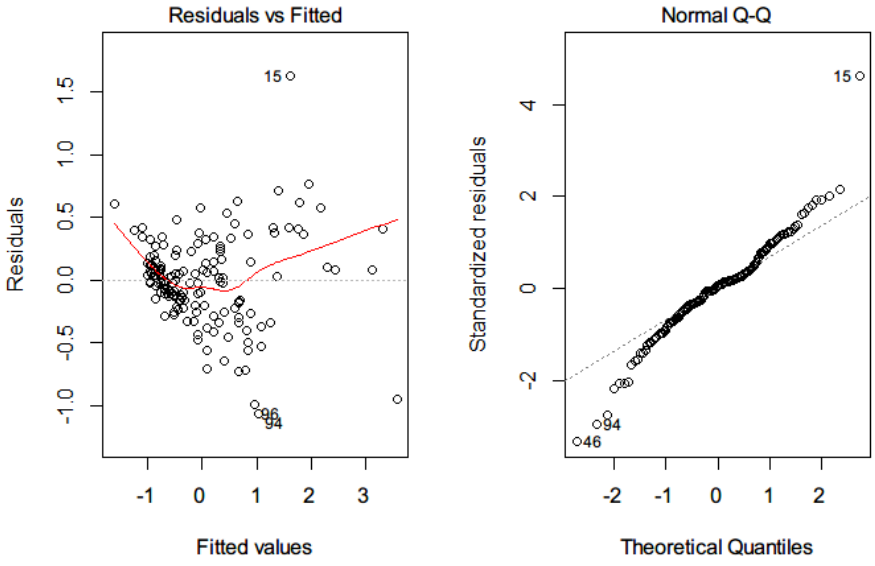
$$VIF_j = \frac{1}{1 - R_j^2}$$

其中，R_j^2是由X_j作为因变量时对其余自变量进行回归得到，R_j^2越接近 1，VIF_j越大，X_j与其余自变量的线性相关程度越强。一般认为，$VIF_j \geq 10$，表示自变量X_j与其余自变量之间存在严重的多重共线性，并影响最小二乘法得到的回归系数估计值。



上图为 14 个连续变量的 VIF 值，整备重量（curb.weight），城区油耗（city.mpg），高速路油耗（highway.mpg）对应的 VIF 远大于 11，表示这三个变量与其余变量存在相当强的相关性，可以选择删除上述三个变量进行回归。

去掉上述三个变量之后的回归模型残差图，体现出异方差性，残差不服从正态分布：



3.3.2 共线性处理

本文中在第四部分建模中会使用以下方法建立回归模型

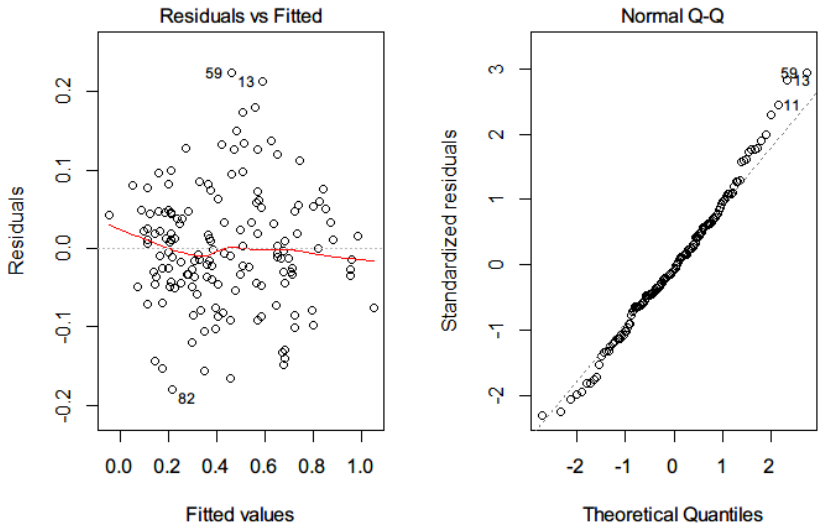
1. VIF 删除变量：去掉方差膨胀因子大于 10 的变量，建立回归模型
2. 正则化处理：Ridge 和 Lasso 回归
3. 主成分回归:降维，归纳主成分代替原有变量

3.4 BOX-COX 变换

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases}$$

当模型存在异方差性，且残差不满足正态分布，回归模型假设不满足。可以对因变量 y 的 BOX-COX 变换 BOX-COX 变化。

下图为上一步 VIF 删除变量之后的数据再进行 BOX-COX 变换得到的残差图：



图中可以看到此时残差均匀的分布在零周围，且 qq 图近似正态。对进行 BOX-COX 变换前后的残差正态性进行检验：Shapiro-Wilk 正态性检验表明变换之后的残差服从正态分布。

Shapiro-Wilk normality test			
Data	W	P-value	Result
Before Box Cox	0.958	1.072E-04	Reject Normal
After Box Cox	0.990	0.338	Normal

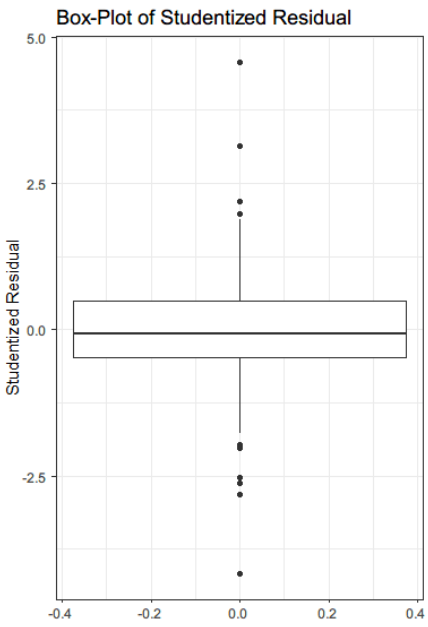
3.5 异常值

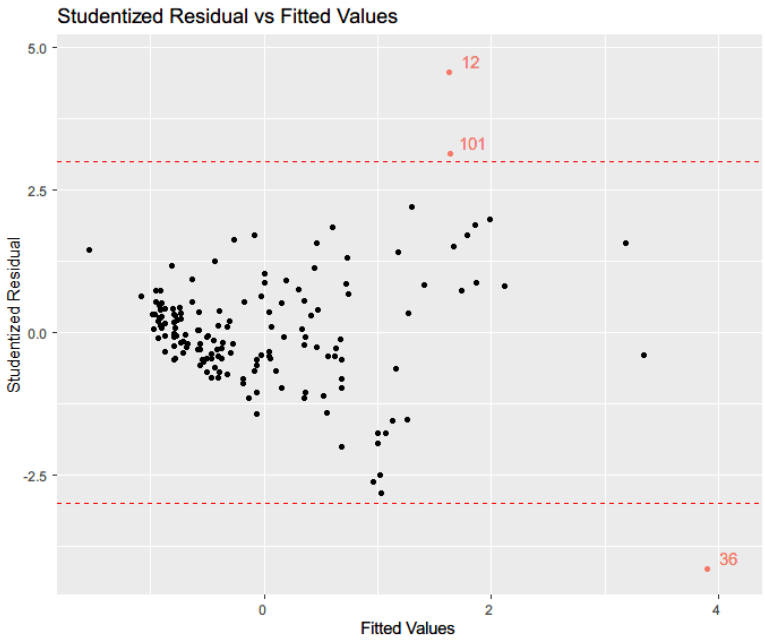
在回归分析中，异常的观测值有可能会引起较大残差，影响回归拟合的效果。由于因变量 Y 的异常值造成的离群点（Outlier），可以通过箱线图和学生化残差进行识别；由于自变量 X 异常造成的高杠杆点（leverage Point）或强影响点（Influence Point），可以结合 cook distance, DFFITS(difference in fits)加以判断。

3.5.1 关于因变量 y 的异常值

下图是训练集中因变量 y 汽车价格的箱线图，以及拟合数据的学生化残差。学生化残差表征第 i 个观测值的异常性，通过 n-1 个观测值拟合回归方程进而得到第 i 个观测值的预测值$\hat{y}_{(i)}$，此时残差为$e_{(i)} = y_i - \hat{y}_{(i)}$。学生化残差的绝对值＞3 将被判断为异常值。

图中显示在训练集中的第 12, 36, 101 条数据是异常值。

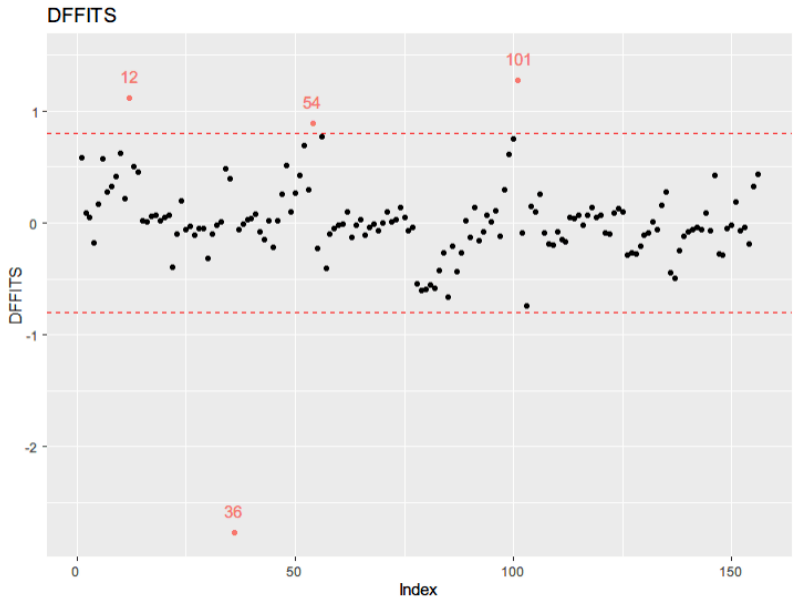




3.5.2 关于自变量 x 的异常值

DFFITS 值表征去掉了某一个数据值之后，新建立的模型对于其余点的拟合残差的变化。判断 DFFITS 的阈值可以通过 DFFITS 图主观确定。下图使用 0.8 作为阈值，第 12,36，54,101 数据被判断为强影响点。

Cook 距离反应去除某一个数据点之后，其余样本点拟合值的变化。Cook distance ＞ 1,判断为异常值点。本例中第 16 条数据被判断为异常值点。



Cook Distance	Hat Values	DFFITS	Studentized Residual
36	36, 55, 56, 89,103	12, 36, 54, 101	12, 36, 101
Outliers			
12, 36, 54, 55, 56, 89, 101, 103			

综合上述诊断方法，共去掉原 194 条训练集中的 8 条数据（第 12,36,54,55,89,101,103 条数据）。

注：上述结果在设置种子（seed(523)）后随机抽取 80%训练数据得到，因为设置不同种子以及相同种子下的抽取样本具有随机性，都会导致 80%训练集数据组成不同，因此结果会有差异。

四、模型建立：拟合角度

4.1 线性模型

4.1.1 逐步回归变量选择

经过第三部分数据预处理步骤，使用 VIF 去除掉三个最强共线性的变量，使用 BOX-COX 变换得到满足回归假设的数据（残差同方差且正态分布），仍剩有 13 个包含显著和不显著的变量。下面使用前进法（forward）、后退法（backward）、逐步回归法（stepwise）三种方法， 以 AIC 信息统计量为准则，通过选择过程中最小的 AIC 信息统计量，来达到剔除或者添加变量的目的。

1. 前进法（forward）

初始模型不包含任何变量，只有常数项，每次增加一个使的模型 AIC 最小的变量，模型中变量由少到多。直至引入新变量所建立的回归方程的 AIC 不会更小，则得到最终的回归方程。

下图表示使用前进法选出来回归系数显著的变量有：年均赔付金额（normalized losses），车型（body size），压缩比（compression ratio），马力（horse power），车身长度、宽度（length、width）。

Forward Selection

Variable	Estimate	Pr(>|t|)	
(Intercept)	0.656	1.040E-12	***
symboling-1	0.062	0.234	
symboling0	0.019	0.700	
symboling1	-0.019	0.717	
symboling2	-0.007	0.900	
symboling3	0.001	0.991	
normalized.losses	0.034	1.830E-03	**
body.stylehardtop	-0.229	1.907E-03	**
body.stylehatchback	-0.246	1.900E-04	***
body.stylesedan	-0.225	1.011E-03	**
body.stylewagon	-0.260	2.240E-04	***
wheel.base	-0.001	0.955	
length	0.048	0.031	*
width	0.038	0.027	*
height	0.021	0.084	.
engine.size	0.018	0.429	
bore	0.007	0.572	
stroke	-0.012	0.161	
compression.ratio	0.037	1.730E-05	***
horsepower	0.120	4.370E-08	***
peak.rpm	0.005	0.592	

模型评估指标显示，adj $-R^2$为 0.884， 训练集上的 RMSE（rooted mean square error）为 0.071，测试集 RMSE 为 0.079，训练集的误差小不排除有过拟合的现象存在。

model	r2	adj.r2	aic	bic	RMSE.train	RMSE.test
Forward Selection	0.900	0.884	-319.952	-254.013	0.071	0.079

2. 后退法（backward）

先通过全部变量建立回归模型，然后计算剔除一个变量使回归模型 AIC 最小的变量，直至任意剔除变量都会使回归模型 AIC 增加，即得到最终的回归方程。

下图表示使用后退法选出来回归系数显著的变量有：年均赔付金额（normalized losses）,车型（body size）,压缩比（compression ratio），马力（horse power），车身长度、宽度（length、width）。

Backward Selection

Variable	Estimate	Pr(>|t|)	
(Intercept)	0.661	<2e-16	***
normalized.losses	0.025	0.002	**
body.stylehardtop	-0.232	0.000	***
body.stylehatchback	-0.243	3.27E-05	***
body.stylesedan	-0.222	1.54E-04	***
body.stylewagon	-0.252	5.87E-05	***
length	0.051	0.004	**
width	0.039	0.011	*
height	0.023	0.037	*
engine.size	0.030	0.126	
stroke	-0.015	0.058	.
compression.ratio	0.039	2.55E-06	***
horsepower	0.124	1.35E-12	***

模型评估指标显示，adj $-R^2$为 0.883， 训练集上的 RMSE（rooted mean square error）为 0.073，测试集 RMSE 为 0.079。本例中，前进法和后退法得到的回归模型在拟合和预测角度几乎没有区别。

model	r2	adj.r2	aic	bic	RMSE.train	RMSE.test
Backward Selection	0.893	0.883	-326.001	-284.04	0.073	0.079

3. 逐步回归法（stepwise）

model	r2	adj.r2	aic	bic	RMSE.train	RMSE.test
Stepwise	0.893	0.883	-326.001	-284.04	0.073	0.079

本例中，逐步回归法和后退法结果一致，变量选择结果参考“后退法”，此处不赘述。

4.1.2 Ridge 和 lasso 回归

当自变量之间存在较强线性关系，设计矩阵X病态，X^TX求逆不存在，且存在$\lambda_j \approx 0$，使得均方误差$\mathrm{MSE} = \sigma^2 \sum \frac{1}{\lambda_j}$会异常大，导致普通最小二乘法失效。通过给 RSS(square sum of residual)增加惩罚项 L2-norm $\|\beta\|_2^2$ 或者 L1-norm $\|\beta\|_1$，得到:

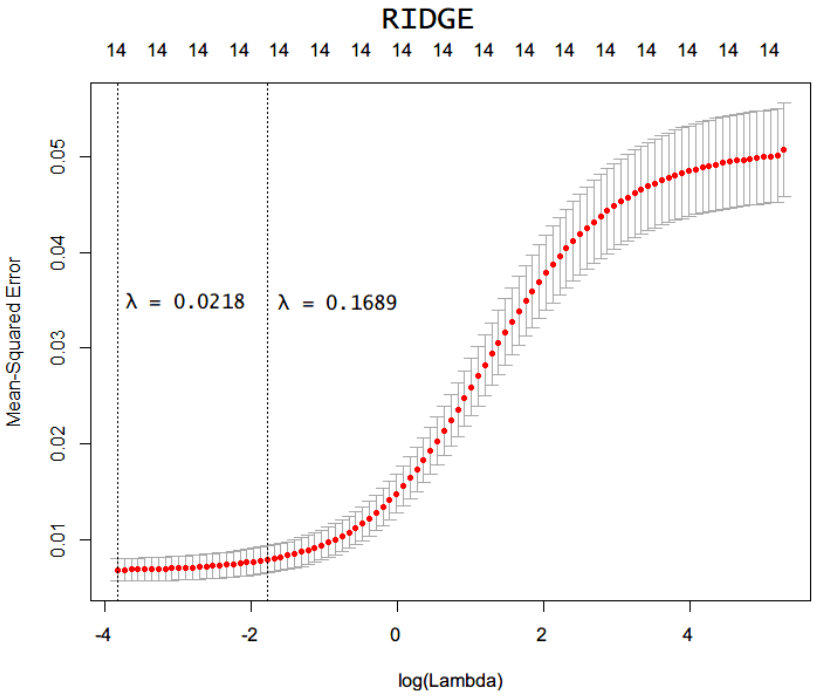
Ridge Regression:

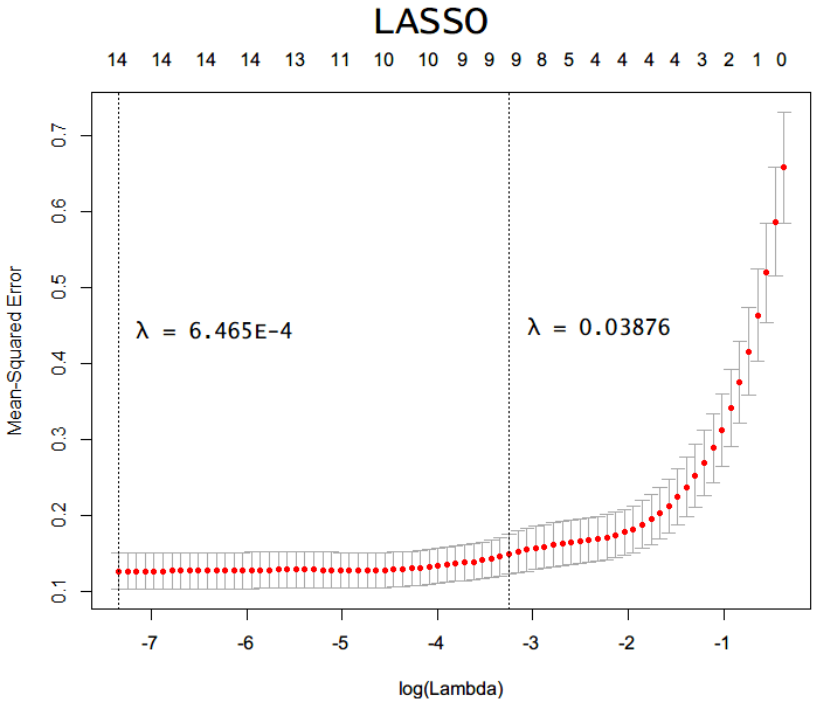
$$\hat{\beta}(\lambda)^{ridge} = argmin(\sum_{i=1}^{n}(y-\hat{y})^2 + \lambda\|\beta\|_2^2)$$

LASSO Regression:

$$\hat{\beta}(\lambda)^{lasso} = argmin(\sum_{i=1}^{n}(y-\hat{y})^2 + \lambda\|\beta\|_1)$$

Ridge 和 LASSO 中参数λ 是使得 MSE($\hat{\beta}(\lambda)$)达到最小的λ，本例中 λ^{ridge} = 0.1689，λ^{lasso} = 0.03875。





1、Ridge 回归模型：

下图表示岭回归得到的变量及回归系数，回归系数显著的变量有：轮距（wheel.base），车身宽度（width）、整备质量（curb.weight）、发动机尺寸（engine.size）、冲程（stroke）、压缩比（compression ratio），马力（horse power）、城区油耗（city.mpg）。

Ridge

Vairables	Coef	Pr(>|t|)	
(Intercept)	-0.043	NA	
normalized.losses	0.043	0.062	.
wheel.base	0.055	0.010	*
length	0.033	0.101	
width	0.114	0.000	***
height	0.023	0.318	
curb.weight	0.114	0.000	***
engine.size	0.269	<2E-16	***
bore	-0.001	0.967	
stroke	-0.069	0.005	**
compression.ratio	0.092	0.000	***
horsepower	0.191	<2E-16	***
peak.rpm	0.042	0.061	.
city.mpg	-0.047	0.006	**
highway.mpg	-0.021	0.262	

模型评估指标显示，训练集上的 RMSE（rooted mean square error）为 0.845，测试集 RMSE 为 0.834。

model	RMSE.train	RMSE.test
Ridge	0.845	0.834

2、LASSO 回归模型：

下图表示 LASSO 回归得到的变量及回归系数，LASSO 有变量选择的功能，其中回归系数为零的变量（如：车身长度（length）、缸径（bore）、城市/高速路油耗（city/highway-mpg））则不能纳入回归模型。

LASSO 回归得到的变量有：引擎尺寸（engine size）、压缩比（compression ratio），马力（horse power）、车身宽度/高度（width/height）、轴距（wheel.base）、

LASSO

Variable	Coefficient
engine.size	0.482
horsepower	0.239
width	0.151
compression.ratio	0.092
peak.rpm	0.040
wheel.base	0.038
height	0.027
curb.weight	0.011
normalized.losses	6.54E-05
(Intercept)	-0.028
stroke	-0.083
length	0
bore	0
city.mpg	0
highway.mpg	0

模型评估指标显示，训练集上的 RMSE（rooted mean square error）为 0.341，测试集 RMSE 为 0.266。相较于 Ridge regression 有较小的预测误差，表明预测效果较好。

model	RMSE.train	RMSE.test
LASSO	0.341	0.266

4.2 主成分回归线性模型

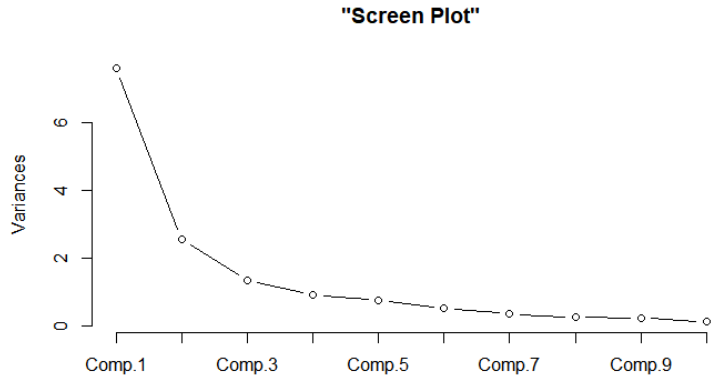
变量之间存在高相关性，会使观测信息有重叠，且最小二乘法估计参数失效。可以将多个变量转化为少数几个主成分，实现降维，简化模型。

第三部分中处理过缺失值和标准化后的数据，自变量之间存在共线性、模型残差存在异方差性，并且不服从正态分布。此处欲引入主成分回归方法，解决自变量共线性问题；观测到主成分回归后的残差图含有二次项趋势，进一步进行多项式回归增加二次项，之后使用逐步选择筛选变量（stepwise），最终得到的模型可以去除异方差性。

4.2.1 主成分解释

对所有连续变量使用主成分回归，累积贡献率前 4 个主成分达到 84.5%，前

五个主成分达到 88.8%，前六个主成分达到 91.8%。结合碎石图，从解释性角度表明取前五个主成分即可。



	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	…	Comp.15
Cumulative Proportion	0.495	0.687	0.784	0.845	0.888	0.918	0.939	…	1.000

下图热力图表示前五个主成分的载荷矩阵，数值代表相关系数，颜色表示相关程度。根据载荷的正负及取值大小，结合第二部分“变量含义”，对 5 个主成分进行解释：

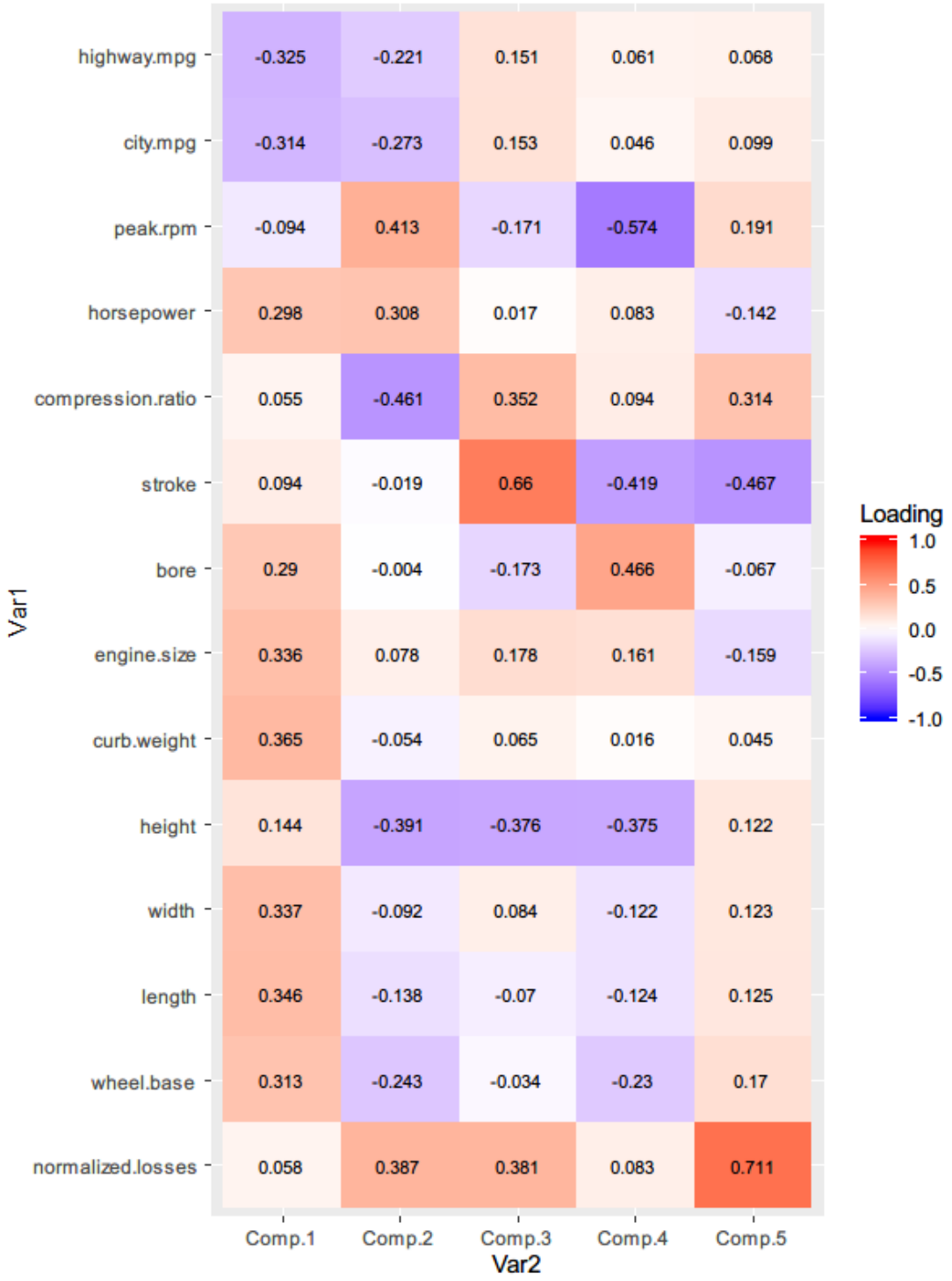
第一主成分 z1：油耗。第一主成分在“城市/高速路油耗”中等程度负载荷，每加仑跑的公里数越多，油耗越小；在汽车的质量长度参数“整备重量”、“车身长度/宽度”上有中等程度正载荷，表示车型越大，油耗越多。

第二主成分 z2：操控灵活性 。第二主成分在汽车质量长度参数“整备重量”、“车身长度/宽度”、“轴距”均体现为负载荷，长宽较大的车型在狭窄街道中的行驶灵活性降低；在“年均赔付”体现中等正载荷，灵活性好的车辆安全性下降，则年均赔付金额会增加。

第三主成分 z3:动力性。第三主成分在“冲程”、“压缩比”分别体现高等和中等正载荷，这两个变量数值越大，发动机效率越高，动力越大；在“年均赔付”体现中等负载荷，动力性能好的车辆稳定性低，安全性低，导致需要报修的金额增加。

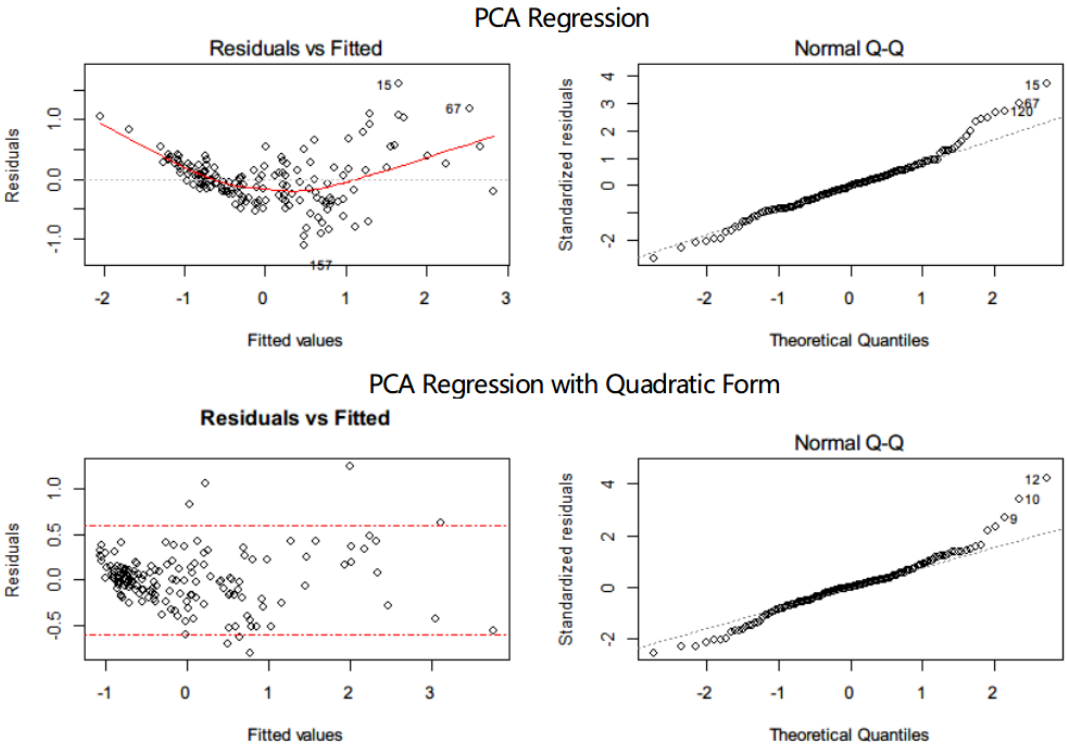
第四主成分 z4:稳定性。第四主成分在“最大转速”、“冲程”、“车身高度”呈现中等负载荷，在“缸径”呈现中等正载荷；大缸径短冲程，可以降低引擎高度和车辆重心，则车辆的稳定性高。

第五主成分 z5：危险性。第五主成分在“年均赔付”拥有高等程度的正载荷，车辆安全系数月底，赔付金额越高。



4.2.2 主成分模型

直接进行主成分回归后，残差图含有二次项趋势，因此考虑引入二次项，进一步进行多项式回归。对引入二次项之后的模型使用逐步回归法（stepwise）选择筛选变量，由残差图可知，此时的模型已经去除异方差性(均匀分布在正负两倍标准差内)，虽然残差正态性仍无法满足。



引入二次项进行逐步回归之后的筛选出的变量有：

z_1, z_3, z_4 ,z_1^2,z_2^2, z_4^2, z_5^2 ,z_1z_2,z_2z_3, z_3z_4, z_2z_4，风险等级（symboling），车型（body style）。

PCA Regression + Quadratic + Stepwise Selection

Variable	Estimate	Pr(>|t|)	
z1	7.827	<2E-16	***
z1^2	2.642	2.230E-16	***
z2	1.320	3.770E-03	**
z1z2	9.629	0.015	*
z2^2	1.169	1.552E-03	**
z2z3	-27.834	1.120E-07	***
z3^2	0.623	0.094	.
z4	-0.721	0.058	.
z2z4	18.913	1.420E-04	***
z3z4	-6.161	0.084	.
z2z5	15.242	1.070E-05	***
z4z5	5.348	0.177	
z5^2	-0.930	1.978E-03	**
body.styleconvertible	0.221	0.280	
body.stylehardtop	-0.026	8.228E-01	
body.stylehatchback	-0.150	7.600E-04	***
body.stylesedan	-0.082	2.437E-02	*
body.stylewagon	-0.292	1.510E-04	***

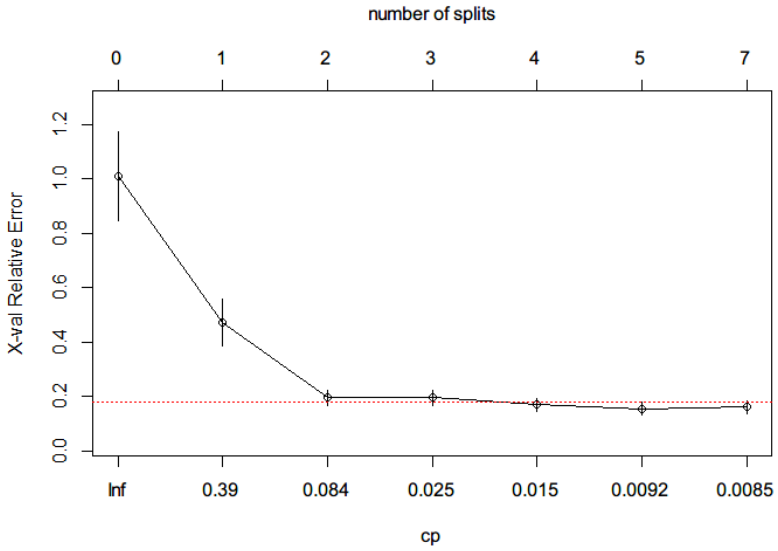
引入二次项之后的回归模型模型拟合效果增强，adj $-R^2$增大为 0.895，训练集上的 RMSE（rooted mean square error）减小至 0.249，测试集 RMSE 为 0.913。

model	r2	adj.r2	aic	bic	RMSE.train	RMSE.test
PCA Regression	0.771	0.747	171.3906	219.346	0.388	0.312
PCA + Quadratic + Step	0.908	0.895	46.81168	103.7587	0.249	0.913

4.3 回归树

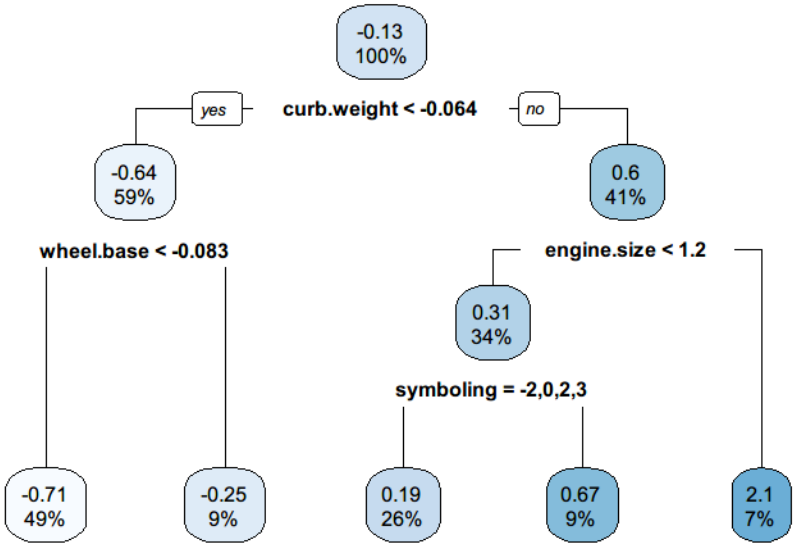
4.3.1 复杂度确定

首先需确定回归树的模型复杂度，在训练集上进行交叉验证，得到交叉验证误差与复杂度参数 Cp 的关系。当 Cp=0.015 时，模型的预测效果较好。



4.3.2 构建回归树模型

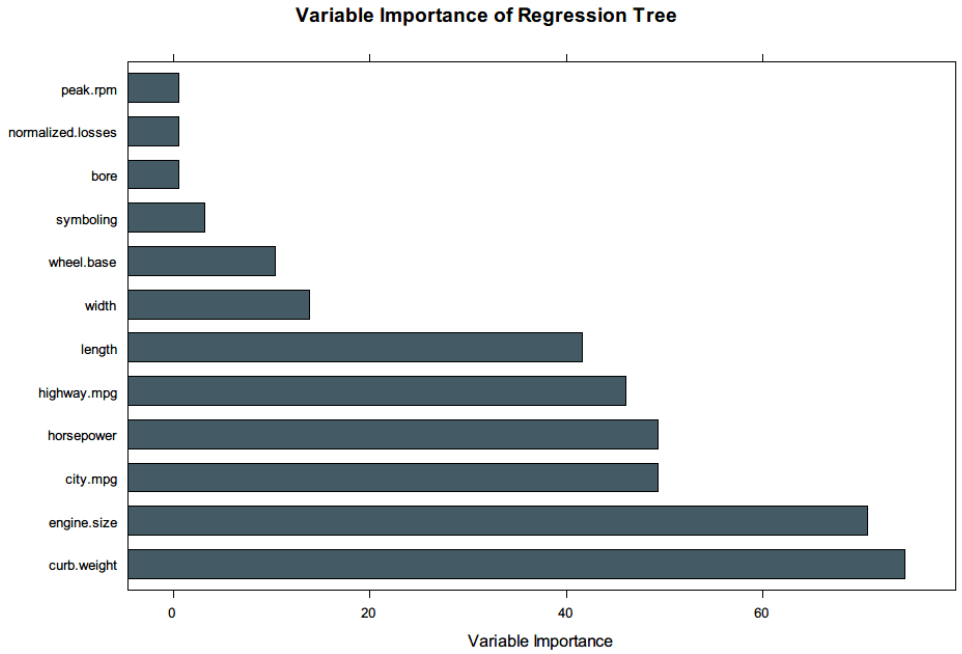
通过观察上述 CV vs Cp 图，此处构建复杂度参数 Cp=0.015 时的回归树模型。该模型的分割方式如下：



模型评估指标显示，训练集上的 RMSE（rooted mean square error）为 0.274，测试集 RMSE 为 0.310。

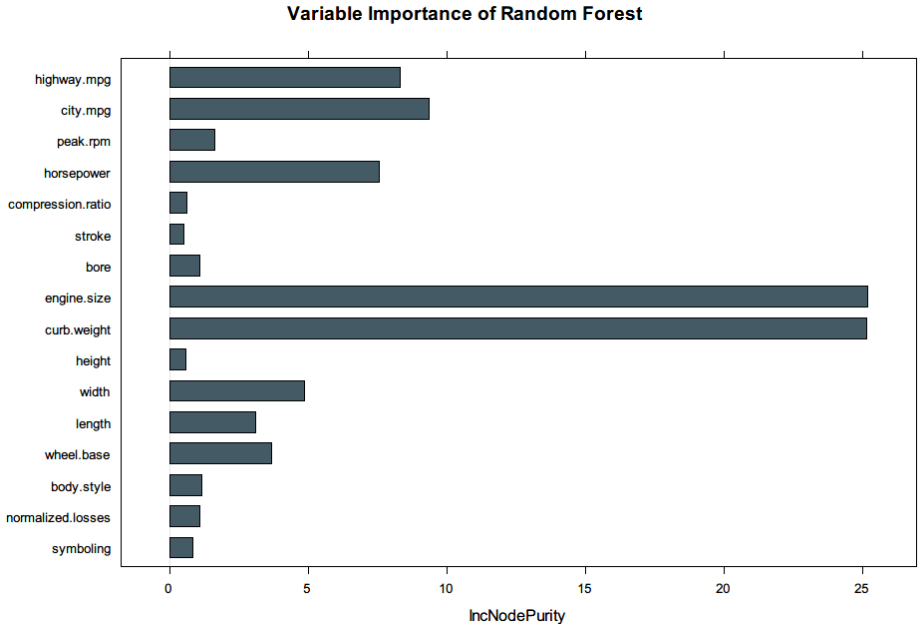
model	RMSE.train	RMSE.test
Regression Tree	0.274	0.310

回归树变量重要性如下图显示，回归树模型结点依次按照变量重要性进行分隔。



4.3.3 随机森林

下图为随机森林模型的变量重要性，相比于回归树模型的变量重要性，前五个最重要的变量都是整备质量（curb weight），引擎尺寸（engine size），城市油耗（city.mpg），高速路油耗（highway.mpg），马力（horse power）。



模型评估指标显示，训练集上的 RMSE（rooted mean square error）为 0.123，测试集 RMSE 为 0.196。

model	RMSE.train	RMSE.test
Random Forest	0.123	0.196

五、模型比较：预测角度

下表为第四部分[模型建立]中各个模型从预测角度的评估值，其中 RMSE.test 表示使用预先划分的 20%测试集数据计算均方根误差；LOOCV 表示在训练集数据上“leave one out”得到的均方误差 MSE。

观察对照 RMSE.test 可知，回归树模型，随机森林模型，lasso 回归模型有较好的预测效果。

model	RMSE.test	LOOCV
Forword	0.079	6.88E-03
Backward	0.079	6.44E-03
Stepwise	0.079	6.44E-03
Pca+Quadratic+Stepwise	0.913	0.083
LASSO	0.266	
Ridge	0.834	0.149
Regression Tre	0.310	
Random Forest	0.196	

六、总结

1、变量：不同模型选择变量各有不同，在模型中都表现显著的变量有：

年均赔付金额（normalized losses）,车型（body size）,压缩比（compression ratio），马力（horse power），车身长度、宽度（length、width）。

2、拟合角度：

（1）逐步回归三种方法使用了 BOX-COX 变换，满足残差正态分布及均值为零，使得拟合均方误差非常小，然而不排除过拟合的问题。

（2）LASSO 回归，不仅可以进行变量选择，本例中模型拟合和预测的方面都体现出较好结果。

（3）主成分回归解释性差，且只能对连续变量进行主成分回归，得到的回归模型只包括连续变量，具有一定的局限性。但本例中拟合的均方误差较小，说明模型拟合效果不错。

（4）回归树与随机森林部分，一旦选择到合适的复杂度参数 Cp,会得到较好的拟合和预测效果。

3、预测角度：回归树模型，随机森林模型，lasso 回归模型具有较好的预测效果。

model	r2	adj.r2	aic	bic	RMSE.train	RMSE.test
Forword	0.900	0.884	-319.952	-254.013	0.071	0.079
Backward	0.893	0.883	-326.001	-284.04	0.073	0.079
Stepwise	0.893	0.883	-326.001	-284.04	0.073	0.079
Pca+Quadratic+Stepwise	0.908	0.895	46.81168	103.7587	0.249	0.913
LASSO	-	-	-	-	0.341	0.266
Ridge	-	-	-	-	0.845	0.834
Regression Tre	-	-	-	-	0.274	0.310
Random Forest	-	-	-	-	0.123	0.196

七、讨论

1、预测模型的时效性：数据来自 1987 年车辆价格数据，40 年后的今天汽车价格有明显提升；此外，自变量车的型号和规格和当下有大差异，比如市面中已经没有“车型”中的 hardtop 型车。因此根据本文数据建立的模型，如果应用于当下数据的预测，会存在偏差。

2、建立回归模型的假设难以满足：数据需要一定的变化才能去除异方差性，如使用 BOX-COX 对因变量做变换，使用加权最小二乘法加权去除异方差，或者使用多项式回归增加二次项等方法。然而，使用上述变换之后仍有可能难以满足残差正态性的假设，此外，变换后数据会导致解释性降低。

3、主成分难以解释：主成分法作为一个中间步骤而非主要目标时，主成分可不必给出解释，（源自参考文献[2]）。本文在第四部分主成分回归建模时结合自变量含义，给出了前五个主成分的解释。解释会存在主观性，使用主成分回归难以像线性模型选出变量并有明确含义。

参考文献

[1]《应用回归分析（R 语言版）》，何晓群，电子工业出版社，2017

[2]《应用多元统计分析》王学民，上海财经大学出版社，2017