**实验报告**

|  |  |  |
| --- | --- | --- |
| **课程名称**： **应用回归分析实验** | **学期：2023年秋季学期** | **成绩**： |
| **指导教师**： 陈丹 | **学生姓名**： **枫叶** | **学生学号**： |
| **实验名称**： **统计诊断** | | |
| **实验编号**：5 | **实验日期**： | **实验学时**： |
| **学院： 数学与统计学院** | **专业： 统计学** | **年级**： **2021级** |

**一、实验目的**

掌握统计诊断的基本方法

**二、使用环境**

R4.3.1

**三、实验内容**

基于cars数据集练习统计诊断方法

1. **算法介绍及结果**

首先进行全模型回归，由于长度问题不展示全模型回归结果。全模型回归结果显示存在完全共线性，根据R的提示，应当删去Saturn, coupe, wagon，这实质上是因为归属于商标、车门类型、车辆类型三个属性下的各个二分类变量之和都应该等于1，删去三个分类变量后进行基准回归，结果如表3所示，

删去导致完全共线性的变量之后，各变量的VIF值如表1所示，可以看到所有变量的VIF都小于10，故多重共线性问题在可接受范围内

表 1

| 变量 | Mileage | Cylinder | Doors | Cruise | Sound | Leather | Buick |
| --- | --- | --- | --- | --- | --- | --- | --- |
| VIF | 1.01 | 2.35 | 4.61 | 1.55 | 1.14 | 1.19 | 2.6 |
| 变量 | Cadillac | Chevy | Pontiac | Saab | convertible | hatchback | sedan |
| VIF | 3.33 | 4.41 | 3.42 | 3.56 | 1.63 | 2.45 | 4.51 |

由于变量过多，我们希望进一步选择变量，首先进行正态性检验，残差QQ图如图1所示，有较多的点不在直线上，这说明数据与正态分布偏离较远。同时我们从图2可以看到，残差随拟合值增大呈现喇叭状分布，这说明存在异方差问题，Spearman相关系数结果也显示在0.05显著性水平下认为存在异方差问题。而从图3可以看出195,196,200三个数据点的学生化残差绝对值大于2，表明它们为异常点。

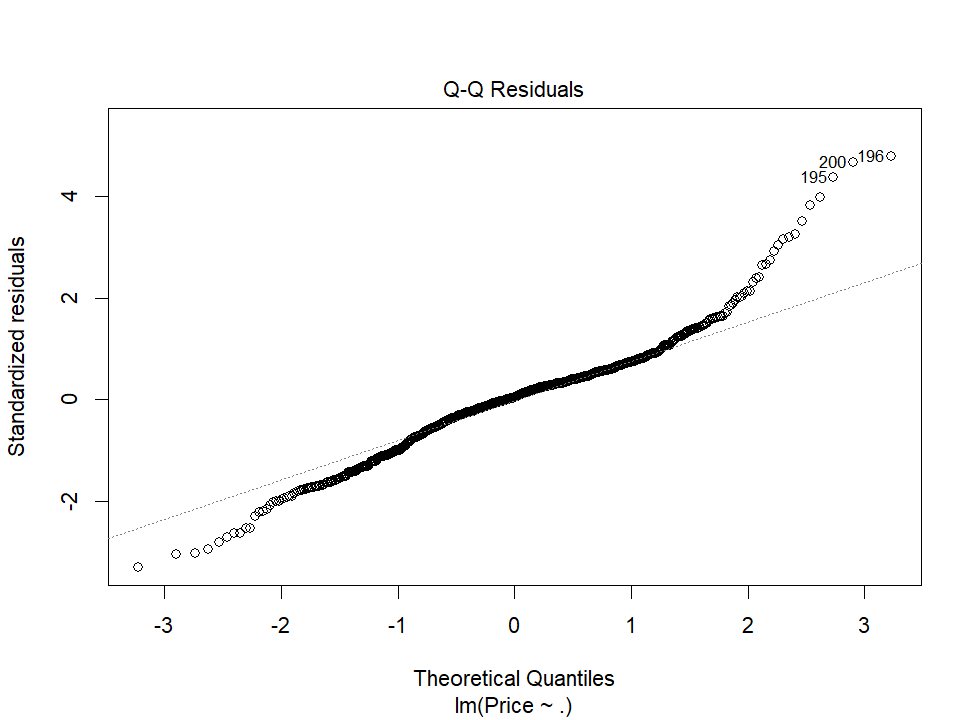


图 1

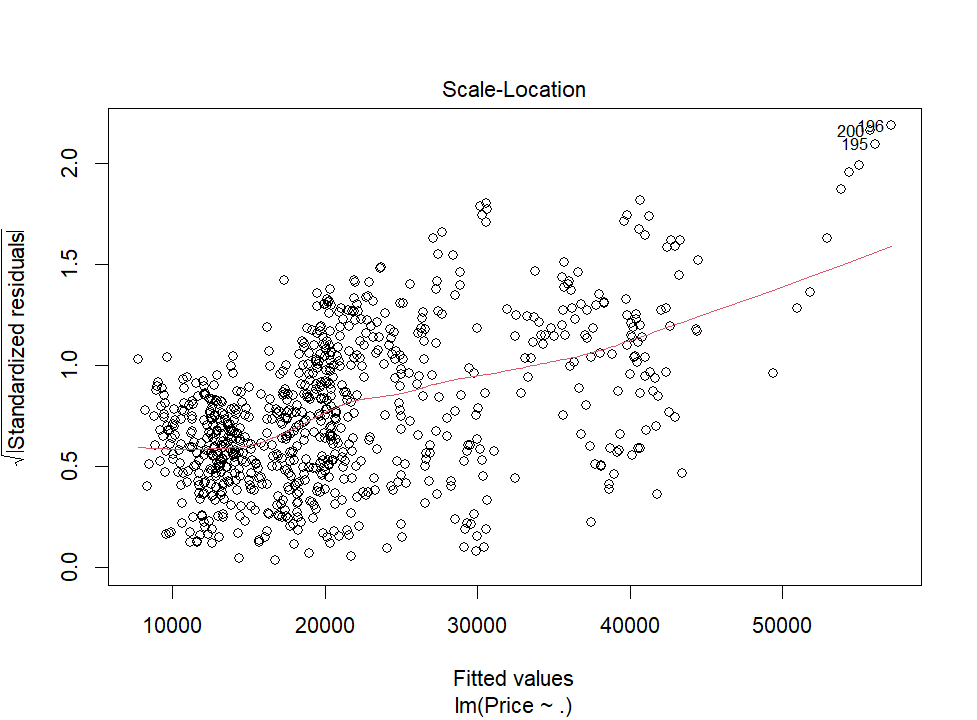
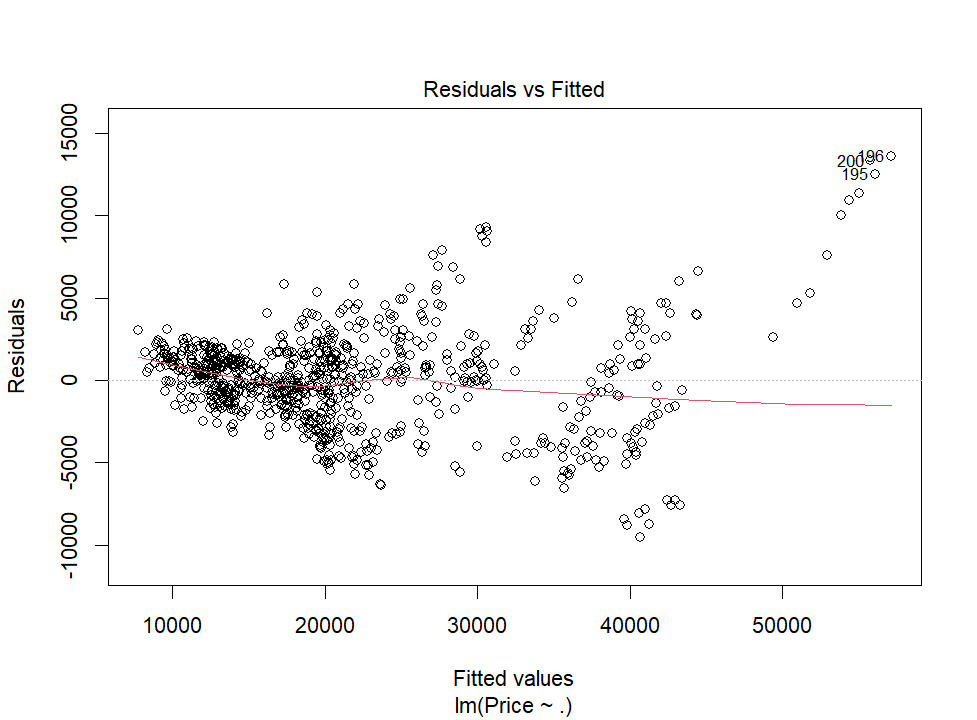


图 2 图 3

为了解决上述问题，使用box-cox变换，变换后的回归结果如表3所示。对其再次画出残差图（图4和图5），注意到残差大致呈现均匀分布， 且所有数据点的学生化残差绝对值均小于2，以此为标准则无异常点，但从cook距离图（图6）可以看到第630,738,800个数据点为强影响点，需要专门分析。从残差QQ图（图7）可以知道，经Box-Cox变换后的数据已经在很大程度上接近于正态，故使用AIC准则进行逐步回归以挑选变量， 结果如表3所示。

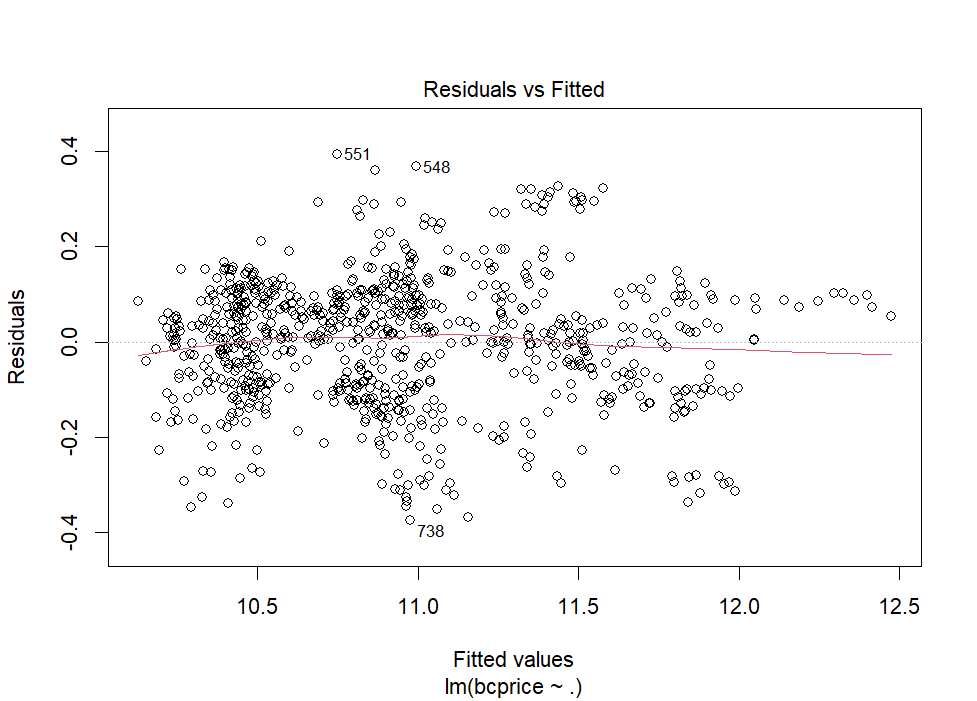


图 4

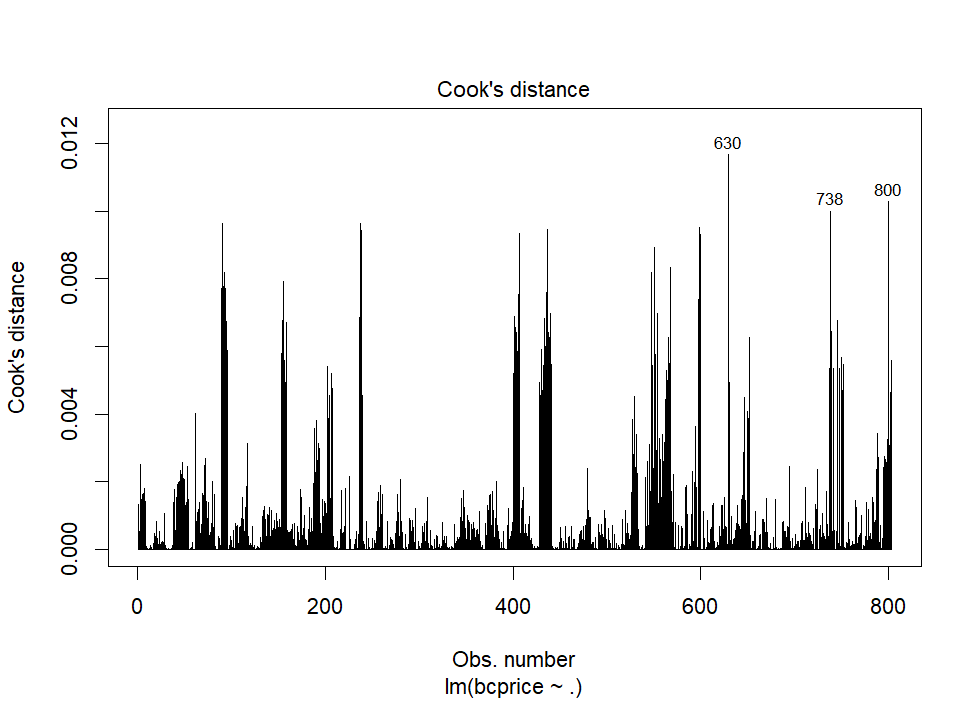
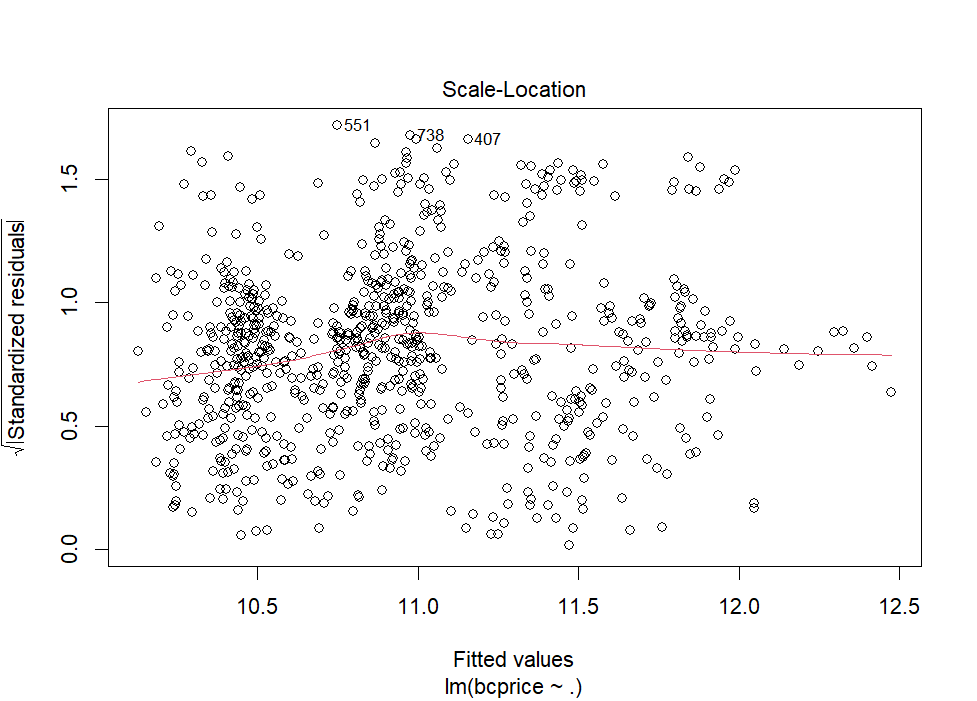


图 5 图 6

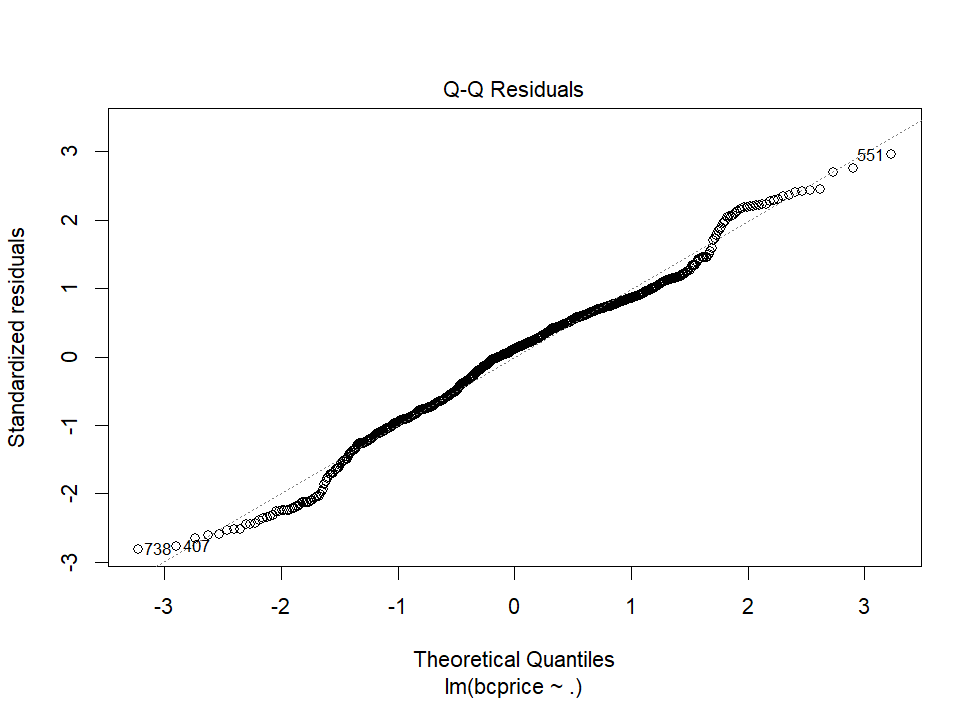


图 7

从回归结果来看，除Pontiac外，其余变量均在0.1的显著性水平下显著，且大部分变量具有极高的显著性。值得质疑的是，二分类变量的系数仅代表控制其他变量的情况下两种水平之间的差异，而该数据集的二分类变量占据了绝大多数，故而回归模型的预测意义有待商榷。此外，由于该数据集为截面数据，故而对序列相关性的检验是无意义的（除非样本数据的排序具有某种内涵），故不作序列相关性的检验。

由于Box-Cox变换的lambda值为0.02，其解释性较差，这里仅描述正向和负向作用，除Mileage,Chevy,Pontiac,hatchback,sedan对价格具有负向作用外，其余变量均对价格具有正向作用。用第一个样本所作的预测如表2所示,真实值为11.11647,相差较大

表 2

| 拟合值 | 置信下限 | 置信上限 |
| --- | --- | --- |
| 10.97072 | 10.9372 | 11.00423 |

表 3

| 模型 | 基准回归 | Box-Cox | 逐步回归 |
| --- | --- | --- | --- |
| (Intercept) | -1123.114 | 9.582\*\*\* | 9.582\*\*\* |
|  | (992.654) | (0.046) | (0.046) |
| Mileage | -0.184\*\*\* | 0.000\*\*\* | 0.000\*\*\* |
|  | (0.013) | (0.000) | (0.000) |
| Cylinder | 3657.982\*\*\* | 0.225\*\*\* | 0.225\*\*\* |
|  | (113.363) | (0.005) | (0.005) |
| Doors | 1568.464\*\*\* | 0.079\*\*\* | 0.079\*\*\* |
|  | (258.939) | (0.012) | (0.012) |
| Cruise | 327.153 | 0.033\* | 0.033\* |
|  | (296.333) | (0.014) | (0.014) |
| Sound | 442.435+ | 0.021+ | 0.021+ |
|  | (234.504) | (0.011) | (0.011) |
| Leather | 794.342\*\* | 0.045\*\*\* | 0.045\*\*\* |
|  | (249.788) | (0.012) | (0.012) |
| Buick | 954.438+ | 0.108\*\*\* | 0.108\*\*\* |
|  | (552.611) | (0.026) | (0.026) |
| Cadillac | 13369.782\*\*\* | 0.495\*\*\* | 0.495\*\*\* |
|  | (624.904) | (0.029) | (0.029) |
| Chevy | -536.253 | -0.036+ | -0.036+ |
|  | (439.954) | (0.020) | (0.020) |
| Pontiac | -1394.578\*\* | -0.032 | -0.032 |
|  | (486.911) | (0.022) | (0.022) |
| Saab | 12283.601\*\*\* | 0.811\*\*\* | 0.811\*\*\* |
|  | (554.668) | (0.026) | (0.026) |
| convertible | 11029.760\*\*\* | 0.348\*\*\* | 0.348\*\*\* |
|  | (541.991) | (0.025) | (0.025) |
| hatchback | -6373.755\*\*\* | -0.354\*\*\* | -0.354\*\*\* |
|  | (610.578) | (0.028) | (0.028) |
| sedan | -4442.832\*\*\* | -0.215\*\*\* | -0.215\*\*\* |
|  | (446.390) | (0.021) | (0.021) |
| Num.Obs. | 803 | 803 | 803 |
| R2 | 0.915 | 0.930 | 0.930 |
| R2 Adj. | 0.914 | 0.928 | 0.928 |
| + p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001 | | | |

代码

|  |
| --- |
| library(caret)  library(dplyr)  library(car)  library(modelsummary)  library(flextable)  library(MASS)  data("cars")  #转二分类变量为因子型  cars[,5:18] <- lapply(cars[,5:18],factor)  model0 <- lm(data=cars,Price~.)  summary(model0)  alias(model0)#存在完全共线性  #留最后一个样本用于预测  cars\_drop <- cars[1:803,-which(names(cars) %in% c("Saturn","coupe","wagon"))]  model1 <- lm(data=cars\_drop,Price~.)  summary(model1)  #多重共线性  VIF <- data.frame(vif(model1))  options(digits = 3)  data.frame(t(VIF)) %>%  `rownames<-`("VIF") %>%  flextable() %>%  align(align = "center",part = "all") %>%  save\_as\_docx(path = "D:/预删除文件夹/大三上/应用回归分析/临时.docx")  #正态性  plot(model1,2)  #异方差  plot(model1,1)  plot(model1,3)  cor.test(abs(ei),cars$Mileage[1:803],method = "spearman")  #变换  lambda <- data.frame(boxcox(model1))  lambda\_max <- lambda[which.max(lambda$y),]$x  options(digits = 7)  cars\_boxcox <- data.frame(bcprice = bcPower(cars\_drop$Price,lambda\_max),cars\_drop[-1])  model2 <- lm(data = cars\_boxcox,bcprice~.)  summary(model2)  #变换结果  plot(model2,1)  plot(model2,2)  plot(model2,3)  plot(model2,4)  #筛选变量  model3 <- step(model2,direction = "both")  #回归结果  options(digits = 2)  models <- list("基准回归" = model1,  "Box-Cox" = model2,  "逐步回归" = model3)  modelsummary(models,output = "flextable",stars = T,  gof\_omit = "IC|F|Log|RMSE",estimate = "estimate",) %>%  align(align = "center",part = "all") %>%  save\_as\_docx(path = "D:/预删除文件夹/大三上/应用回归分析/临时.docx")  #预测  options(digits = 7)  predict(model2,newdata = data.frame(cars[-1][1,]),interval = "confidence") %>%  data.frame() %>%  `names<-`(c("拟合值","置信下限","置信上限")) %>%  flextable() %>%  align(align = "center",part = "all") %>%  save\_as\_docx(path = "D:/预删除文件夹/大三上/应用回归分析/临时.docx")  bcPower(cars$Price[1],lambda\_max) |

**六、参考文献**

**七、教师评语**