**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**： **自变量选择** | | |
| **实验编号**：4 | **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**

**掌握自变量选择方法**

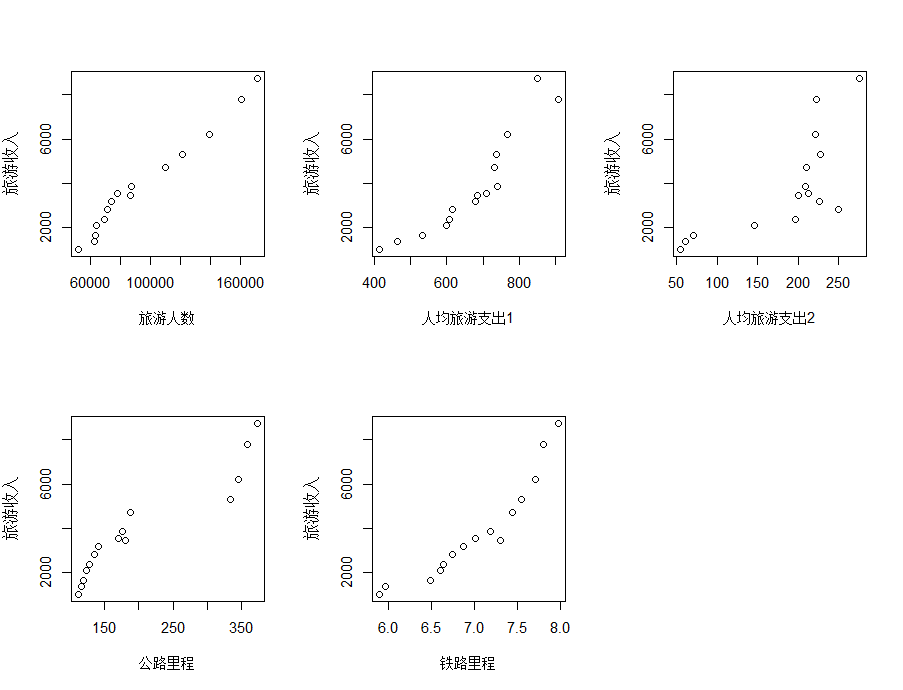
**二、使用环境**

**R4.3.1**

**三、实验内容**

**旅游数据分析**

1. **算法介绍及结果**

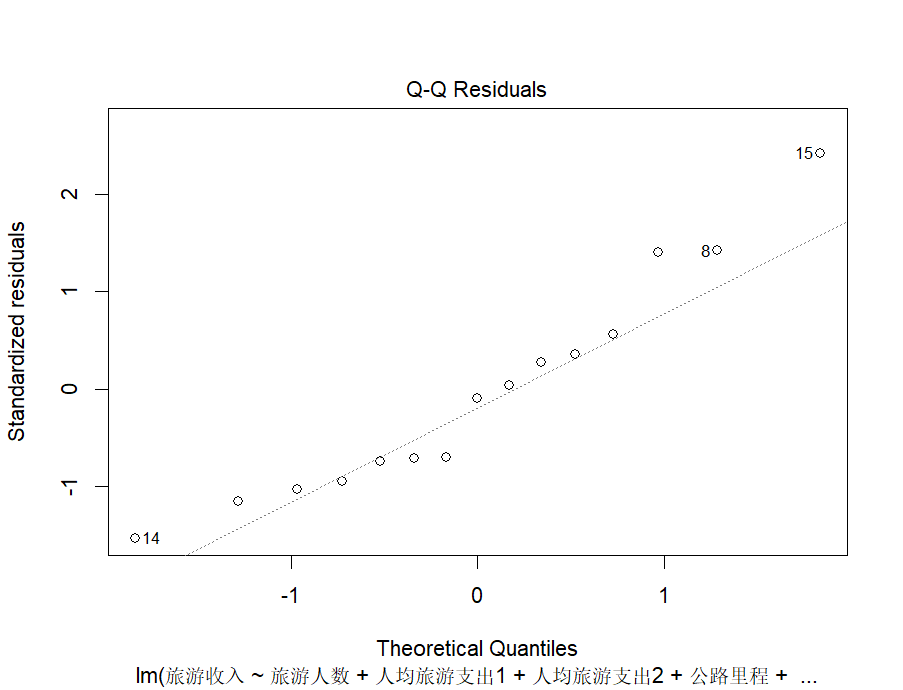


从散点图可以看出旅游人数与国内旅游收入存在一定线性关系，而其他变量是否存在线性关系较不明显（其中人均旅游支出1代表城镇，2代表农村），下面进行基准回归

基准回归结果

|  | (1) |
| --- | --- |
| (Intercept) | -599.023 |
|  | (1807.132) |
| 旅游人数 | 0.054\*\*\* |
|  | (0.007) |
| 人均旅游支出1 | 3.300+ |
|  | (1.663) |
| 人均旅游支出2 | 4.329\* |
|  | (1.607) |
| 公路里程 | -0.676 |
|  | (2.175) |
| 铁路里程 | -502.729 |
|  | (390.310) |
| R2 | 0.995 |
| R2 Adj. | 0.992 |
| + p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001 | |

回归结果如上表所示，在0.01显著性水平下，仅旅游人数这一变量的系数显著，下面需要作进一步的变量选择，由于选取AIC准则作为变量选择的判断标准，故首先需要进行正态性检验

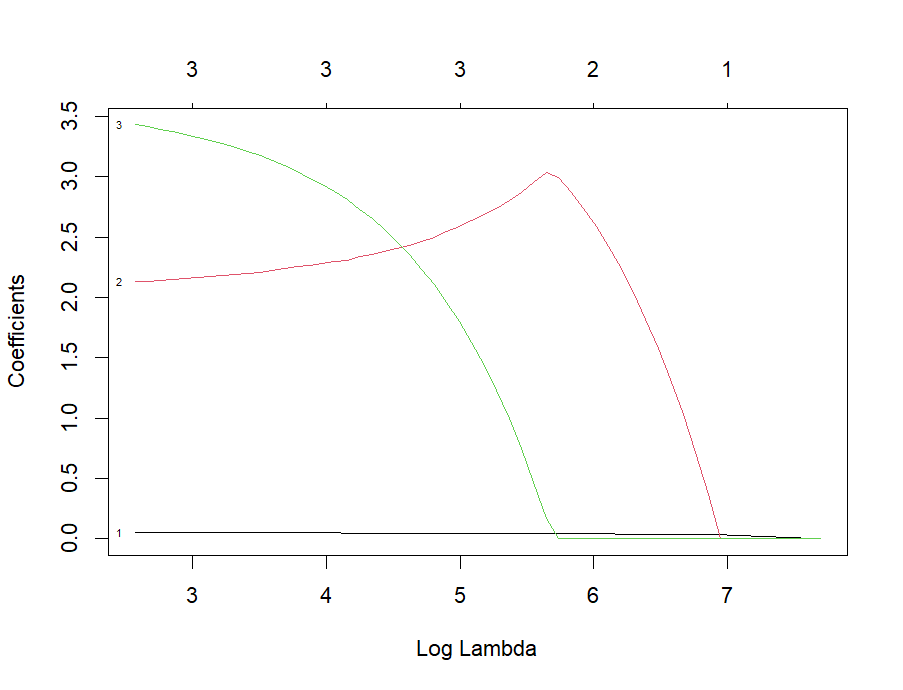


从残差QQ图来看，其近似呈一条直线，但有部分点存在较大偏离，shapiro检验的p值为0.241，在0.01显著性水平上接受原假设，故认为正态性假定成立，以下分别用向前法、向后法和逐步回归法进行变量筛选

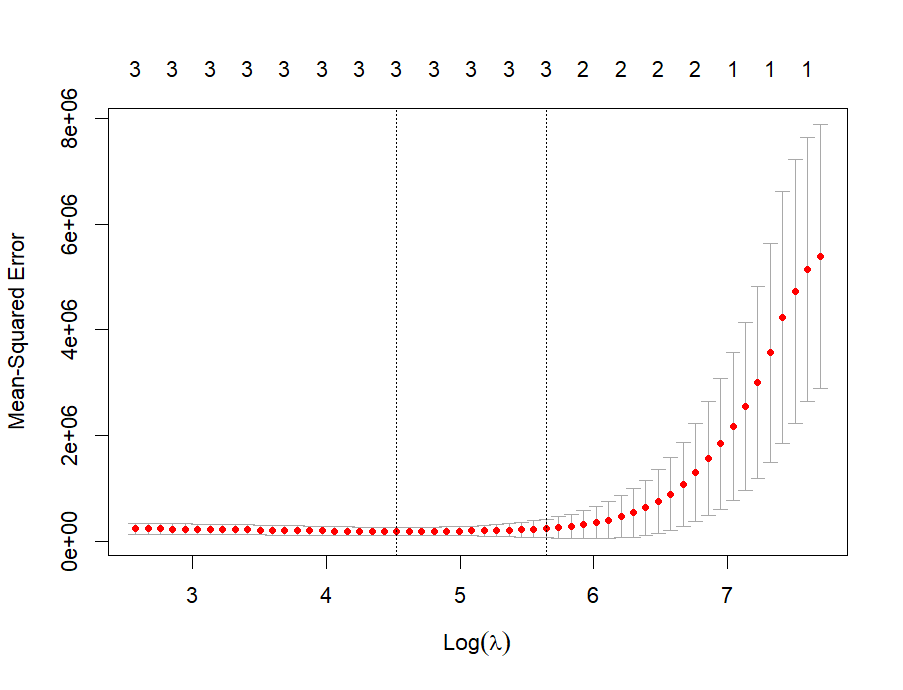
|  | 向前法 | 向后法 | 逐步回归 |
| --- | --- | --- | --- |
| (Intercept) | -599.023 | -470.016 | -470.016 |
|  | (1807.132) | (1677.491) | (1677.491) |
| 旅游人数 | 0.054\*\*\* | 0.053\*\*\* | 0.053\*\*\* |
|  | (0.007) | (0.004) | (0.004) |
| 人均旅游支出1 | 3.300+ | 3.444\* | 3.444\* |
|  | (1.663) | (1.523) | (1.523) |
| 人均旅游支出2 | 4.329\* | 4.335\* | 4.335\* |
|  | (1.607) | (1.533) | (1.533) |
| 公路里程 | -0.676 |  |  |
|  | (2.175) |  |  |
| 铁路里程 | -502.729 | -530.846 | -530.846 |
|  | (390.310) | (362.125) | (362.125) |
| R2 | 0.995 | 0.995 | 0.995 |
| R2 Adj. | 0.992 | 0.993 | 0.993 |
| + p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001 | | | |

上表给出了变量选择的结果，其中向前法保留了所有变量，而向后法和逐步回归法均删去了公路里程变量，除铁路里程外剩余变量均在0.05的显著性水平上显著。

下面使用LASSO回归进行变量筛选



从系数分布图注意到第1-3个变量的系数不随的减小而趋近于零，下面基于3折交叉验证找出一个合适的，图中第一条虚线表示在所有的值中，得到最小目标参量均值的那一个，而第二条虚线表示在最小一个标准差范围内能得到最简单模型的那一个，我们选择后者，不过从图中可以注意到两个对应的自变量都是3个



最终筛选出的自变量是旅游人数、城镇居民人均旅游支出和农村居民人均旅游支出，其回归结果如下表所示

|  | (1) |
| --- | --- |
| (Intercept) | -2857.701\*\*\* |
|  | (421.676) |
| 旅游人数 | 0.050\*\*\* |
|  | (0.003) |
| 人均旅游支出1 | 2.065 |
|  | (1.259) |
| 人均旅游支出2 | 3.607\* |
|  | (1.524) |
| R2 | 0.994 |
| R2 Adj. | 0.992 |
| + p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001 | |

注意到除城镇居民人均旅游支出外，其余变量均在0.05显著性水平下显著。我们选择逐步回归法筛选变量后的回归结果进行解读。

从变量筛选后的回归结果中可以看出，旅游人数、城镇居民人均旅游支出和农村居民旅游支出均对旅游收入具有显著的正向作用，在其他条件不变的情况下，旅游人数每上升1万人，旅游收入平均上升0.053亿元，城镇居民人均旅游支出每上升1元，旅游收入平均上升3.444亿元，农村居民旅游支出每上升1元，旅游收入平均上升4.335亿元，模型修正为0.993，较全模型有微小上升。

代码

|  |
| --- |
| library(readxl)  library(glmnet)  library(modelsummary)  library(flextable)  library(dplyr)  data <- read\_xlsx("D:/预删除文件夹/大三上/应用回归分析/task4数据.xlsx")  par(mfrow=c(2,3))  for(i in 1:5){  plot(data[,c(2+i,2)])  }  model0 <- lm(data=data,旅游收入~旅游人数+人均旅游支出1+人均旅游支出2+公路里程+铁路里程)  modelsummary(model0,output = "flextable",stars = T,gof\_omit = 'IC|Log|RMSE|Obs|F') %>%  width(width = 2) %>%  set\_caption("基准回归结果") %>%  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")  plot(model0,2)  shapiro.test(model0$residuals)  model1 <- step(model0,direction = "forward")  model2 <- step(model0,direction = "backward")  model3 <- step(model0,direction = "both")  models <- list("向前法"=model1,  "向后法"=model2,  "逐步回归"=model3)  modelsummary(models,output = "flextable",stars = T,gof\_omit = 'IC|Log|RMSE|Obs|F') %>%  width(width = 1) %>%  set\_caption("变量选择") %>%  align(align = "center", part = "all") %>%  save\_as\_docx(path = "D:/预删除文件夹/大三上/应用回归分析/临时.docx")  #LASSO  Y <- as.matrix(data[,2])  X <- as.matrix(data[,3:7])  lasso\_model <- glmnet(X, Y, family = "gaussian", alpha = 1)  #绘制LASSO图  plot(lasso\_model,xvar = "lambda",label = T)  #交叉验证并绘制可视化结果  cv\_model <- cv.glmnet(X, Y, family = "gaussian",alpha = 1,nfolds = 3)  plot(cv\_model)  #根据交叉验证结果，选择lambda值  lambda\_1se <- cv\_model$lambda.1se  #根据lambda值选择模型  best\_lasso <- glmnet(X, Y, family = "gaussian", alpha = 1,lambda = lambda\_1se)  coef(best\_lasso)  model4 <- lm(data=data,旅游收入~旅游人数+人均旅游支出1+人均旅游支出2)  modelsummary(model4,output = "flextable",stars = T,gof\_omit = 'IC|Log|RMSE|Obs|F') %>%  width(width = 2) %>%  set\_caption("LASSO变量选择模型") %>%  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") |

**六、参考文献**

**七、教师评语**