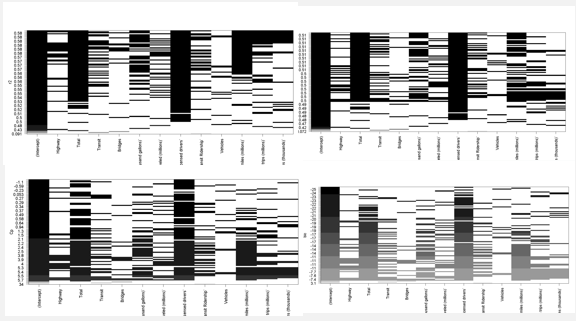
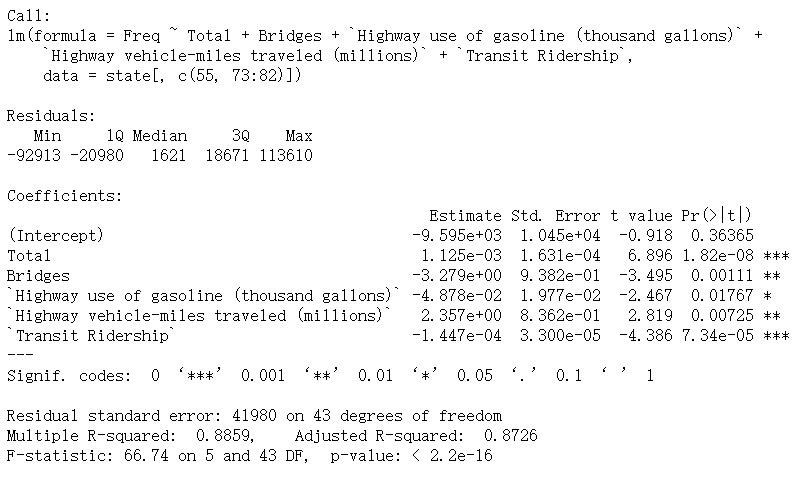
影响因子分析

爬取整合United States Department of Transportation等多个数据源的数据获得了与车祸事故相关的交通财政投入、桥梁数、公路历程数、州平均每公里耗油量、公共交通乘客数、通勤模式等10余个变量并结合源数据集天气相关变量获得扩展数据集。

选用各州车祸数作为因变量进行回归，首先结合逐步寻优法和全子集回归方法进行特征选择，其结果如下：



根据全子集回归的结果最终选定交通财政投入数、桥梁数、公路汽油使用量、公路汽车行驶里程、公路交通乘客数等变量进行回归。首先采用OLS模型进行拟合得到的回归结果如下：



在进行模型评价的过程中发现原始模型存在一定的多重共线性问题，采用方差膨胀系数来对多重共线性进行衡量，最终结果如下表所示：

|  |  |
| --- | --- |
| **变量** | **VIF** |
| 交通财政投入 | 9.450597 |
| 桥梁数 | 2.453150 |
| 公路汽油使用量（千加仑） | 79.968214 |
| 公路车辆行驶里程（百万英里计） | 77.075205 |
| 公共交通乘客数 | 2.46331 |

可以看到公路汽油使用量（千加仑）与公路车辆行驶里程（百万英里计）有较高的VIF值，但去除其中某一变量模型拟合效果大大降低。

图表, 折线图

描述已自动生成图表, 折线图

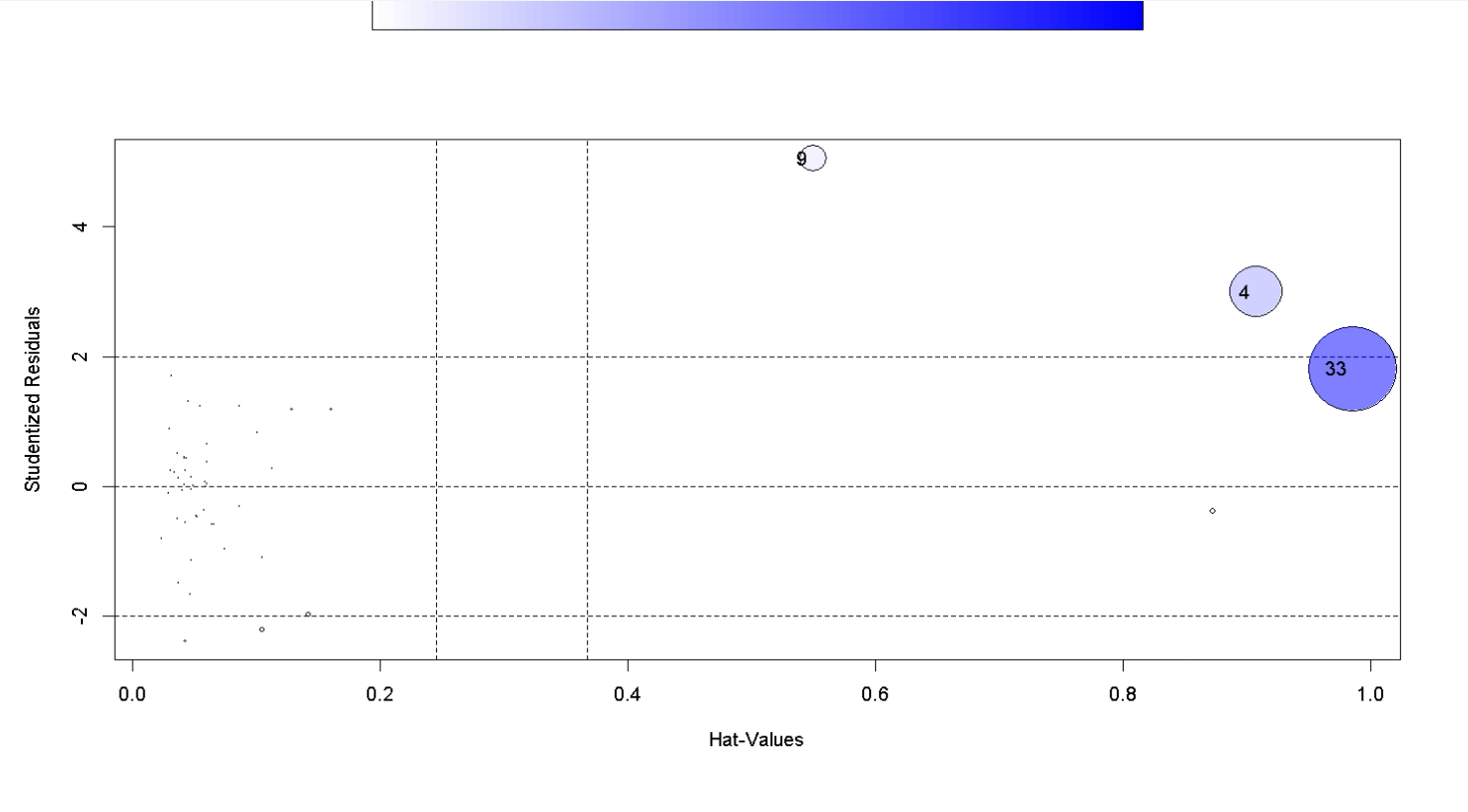
描述已自动生成图表, 折线图

描述已自动生成图表, 折线图

描述已自动生成同时观察其plot（）绘图结果：

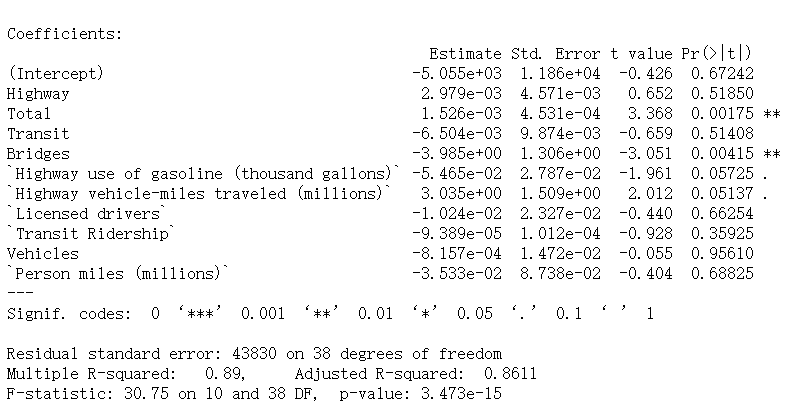
从残差与拟合图中可以看出存在部分规律未被识别，从正态Q-Q图中可以看出部分点不符合正态性假设，同时有较多高影响点。

通过influencePlot函数得到图像如下：



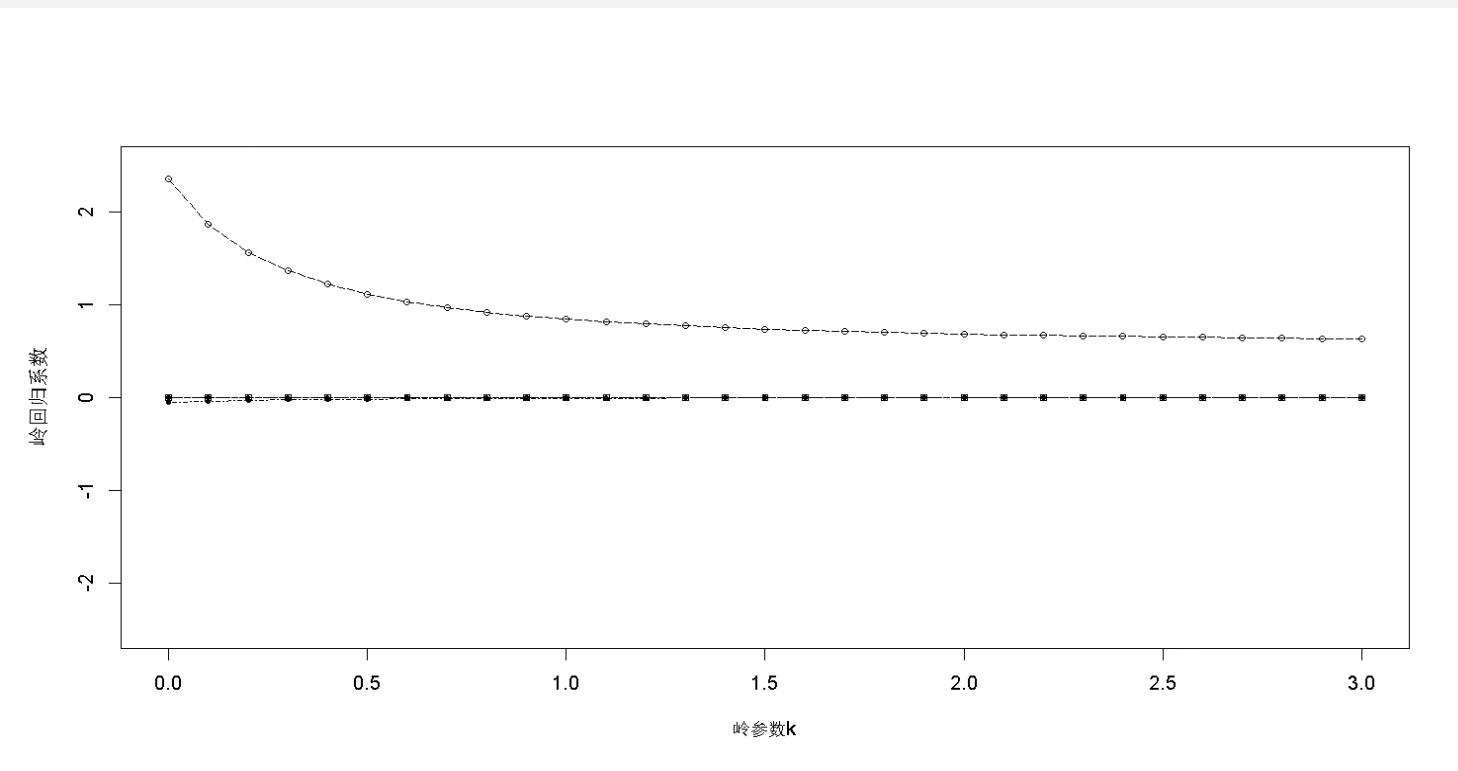
从图中可以明显的观察到几个强影响点分别为佛罗里达州、加利福尼亚州和内华达州。为解决初步进行回归探索中遇到的问题我们进行了如下尝试：

首先尝试采用GWR模型进行改进，GWR是"Geographically Weighted Regression"的缩写，意为"地理加权回归"。它是一种空间数据分析方法，用于探索和建模空间数据中的空间异质性和空间非平稳性。GWR通过在空间上对回归参数进行加权，允许模型参数在不同的空间位置上有所变化。它根据样本点周围的邻域信息为每个样本点赋予权重，将距离较近的样本点对参数估计产生更大的影响。这种加权方法可以更好地捕捉空间数据中的局部空间关系和异质性。试验多次所得效果有限：



最终选用岭回归对模型进行修正：

岭回归是专门用于共线性数据分析的有偏估计的回归方法，实际上是一种改良的最小二乘法，但它放弃了最小二乘的无偏性，损失部分信息，放弃部分精确度为代价来寻求效果稍差但更符合实际的回归方程。



根据岭迹图以及ridge包中提供的linearRidge()自动寻优函数，最终确定岭回归K值为0.1189619。经过岭回归拟合后模型的显著性得到了进一步的提升。

表格

描述已自动生成