对于广告投资和曝光率的一些相关分析

——数据思维第四次小组作业

1. 小组成员：崔士强、郑中锐、张博伦
2. 相关分析：

a)线性回归下的C国2020年广告曝光量预测值：2075

b)二次回归下的C国2020年广告曝光量预测值：1802

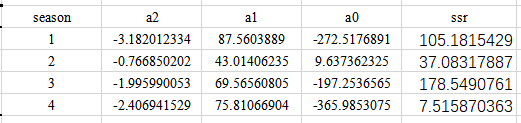
c)在剔除2020年的异常观测值的前提下，我们计算并比较了线性回归和二次回归两种方法的残差平方和SSR，线性的SSR为123万，二次的SSR为55万，显然二次回归函数的拟合精度更高。

而在分析了两种函数的系数后可发现，两种回归函数预测值都是以常数项为主，通过高次项来微调，由此可以推断，在加入了这个三次项后，能够进一步提高拟合性能。

a)在剔除了异常观测值的前提下，我们将三个国家的数据进行合并绘制出散点图，可以发现散点呈近似线性排列，再分别用线性回归和二次回归进行拟合，得到二次回归的二次项系数趋于零，由此可知线性拟合更加准确。

b)线性回归下的C国2020年广告覆盖率预测值：47.18%

c)不能解决因变量对自变量较为敏感的情况，用百分比表示时因变量集中在[0,1]的范围内，相比自变量过小，上面提到的情况更为显著，可以考虑将其放大，例如调整为100倍。或者对自变量和因变量进行差分处理，增强相关性便于分析。



1. 通过对每一个季度的二次回归分析，我们发现第2、4季度的二次回归拟合的较为精确，第1、3季度的回归偏差较大。

b)二次回归下的C国2022年第2季度广告曝光量预测值为：263.4

c)3a中引入了季度对广告效果的影响，使得回归更加精确。分离出了季度的作用效果，也使广告投入作用的梯度更能反映真实作用效果。

1. 分析数据可知，不同年份的季节性固定效应为：1、4季是低谷，第2季达到峰值后逐渐降低。通过对比大流行前后的线性回归函数可发现，一次项系数b1变化不大，说明投资对曝光率的影响没有明显变化，b但常数项b0明显下降，说明大流行使曝光率基数下降。
2. 数据反思：
3. 大流行之前的年份数据较少，回归的可靠性有待提高，若提供到大流行前十年左右的数据，便可以利用时间序列分析或者通过灰色预测模型大致推测大流行期间的理论值，从而更好的分析出大流行对广告曝光的影响。
4. 部分数据相关性不是很明显，尤其在涉及曝光率和季度等变量时，可以先对相关数据进行差分化处理再进行数据分析。
5. 分析者的理论知识和能力还有限，提高后可使用SPSS等工具进一步分析。