问题背景分析：

财政收入是指一个国家政府凭借政府的特殊权力，按照有关的法律和法规在

一定时期内（一般为一年）取得的各种形式收入的总和，包括税收，企事业收入，

国家能源交通重点建设基金收入，债务收入，规费收入，罚没收入等。财政收入

水平高低是反映一国经济实力的重要标志。

在一定时期内，财政收入规模大小受许多因素的影响，如国民生产总值大小，

社会从业人数多少，税收规模大小，税率高低等。在本案例中，我们认为，一个

国家税收水平高低，国民生产总值规模的大小，社会从业人数多少，其他收入的

多少，是决定一个国家一定时期内财政收入规模的主要影响因素。

相关分析：

同分析原始变量间的相关系数和原始变量的散点图，有大部分线性关系明显，通过标准化处理后线性关系将更显著，进行回归分析时效果更好

建立线性回归模型：

1. 筛选变量，
2. 看相关指数。拟合优度R方，F统计量、显著性、误差平方和等。
3. 由参数估计值得出方程。
4. 对模型作残差分析，找出异常观测点和强影响点。

预测：

预测并与真实值比较，判断模型好坏。第二个模型与第一个模型相比，去掉了第六个观测点，拟合的效果较模型一好，但预测时效果稍差，是因为去掉第六个观测时，并没有考虑它是一个强影响点，一个观测点既是异常点，同时也是强影响点，则处理它时需要谨慎。从总体来说，两个模型的效果都比较好。财政收入是一个综合性较强的指标，税收和国内生产总值之时对财政收入水平影响较大的两个只要因素，哈油气他的以下因素，如平均税率，社会劳动生产率等，对财政收入水平也有一定的影响，如果将他们也加入到回归模型中俩，变量的增加必然使模型的拟合度提高，具体表现就是决定系数的提高。但是由于这些变量同财政收入的关系可能并不密切，因此新加入的变量可能无法通过t检验，模型的预测效果也可能不理想。与第一个问题相似，同学们可以尝试一下，观察引入其他变量后的方程拟合效果。应该注意的是，回归方程建立的过程中，并非是变量越多越好，拟合度越高越好，“最佳”的回归方程应该是在一定的拟合水平下，用较少的变量，较简单的方程形式将变量之间的关系描述出来。

模型一



图1——原始变量的描述性统计量的输出结果

从描述性统计量中不能清楚地看到变变量的关系,所以需进行进一步的分析，我们的目的是看各变量的取值范围差别是不是较大，如果较大，则要考虑标准化变量。由输出结果知道原始变量的取值范围较大，需要标准化变量。

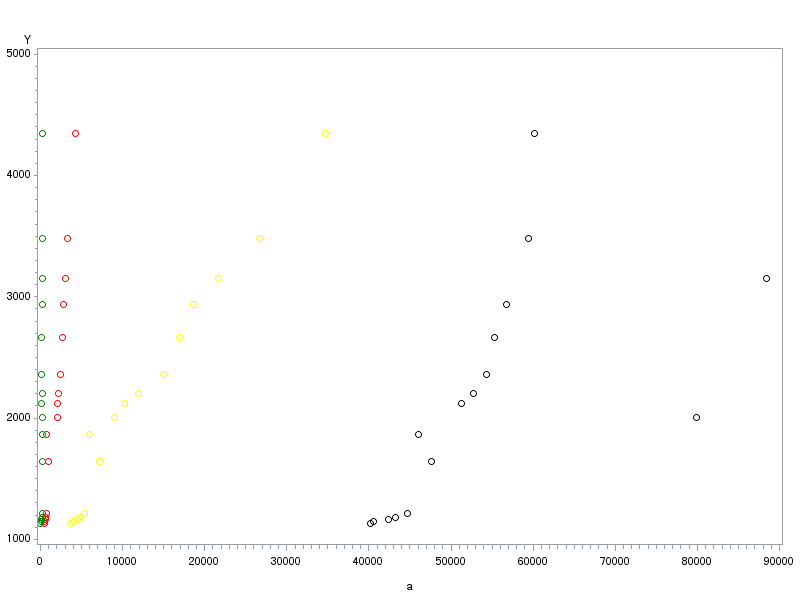


图2——未标准化前变量的散点图

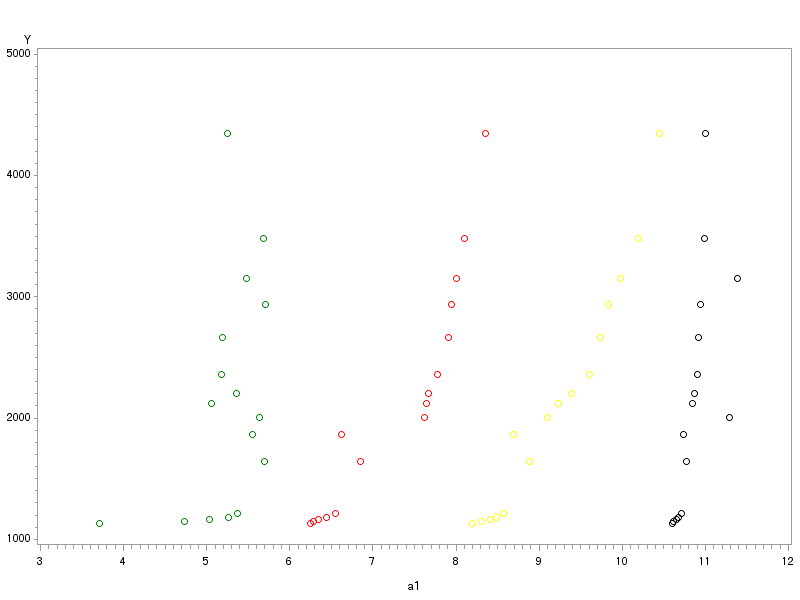


图三——描述性统计量



图4——相关系数矩阵

可以看出变量间的相关性都比较显著，所以它们的相关性大。由图形可以看出它们的关系是线性的，所以可据此建立多元线性回归方程。



**图5——变量标准化后的曲线图**

它们存在着相关性的变化趋势，只是绿色折线：其他收入（X3）曲线波动较大，但因其值相对较小，故保留。



图6



图7



图8

三个模型：逐步回归法、向前选择法和向后消除法对变量进行筛选。根据显著性水平判断，保留变量a和b1.



图9——模型总体

F统计量为319.58，Pr>F<0.0001，模型中变量的作用极为显著，自由度为2，即变量的个数。误差平方和等于274112.。



图10——参数估计值输出。

拟合得到模型的各项参数。截距项为-969.56691，a的系数等于0.08212，log(b)的系数等于289.43460。拟合得方程：yp=-969.56691+0.08212\*a+289.43460\*log(b)



图11——共线性诊断

共线性诊断：条件数K越大，说明数据矩阵有较大的共线性，使结果不稳定。一般在0<K<10则认为没有多重共线性，10<K<30则认为有中等程度或较强的多重共线性。这里最大的条件数是46.45363，说明变量间有严重的多重共线性。容易导致结果不稳定。

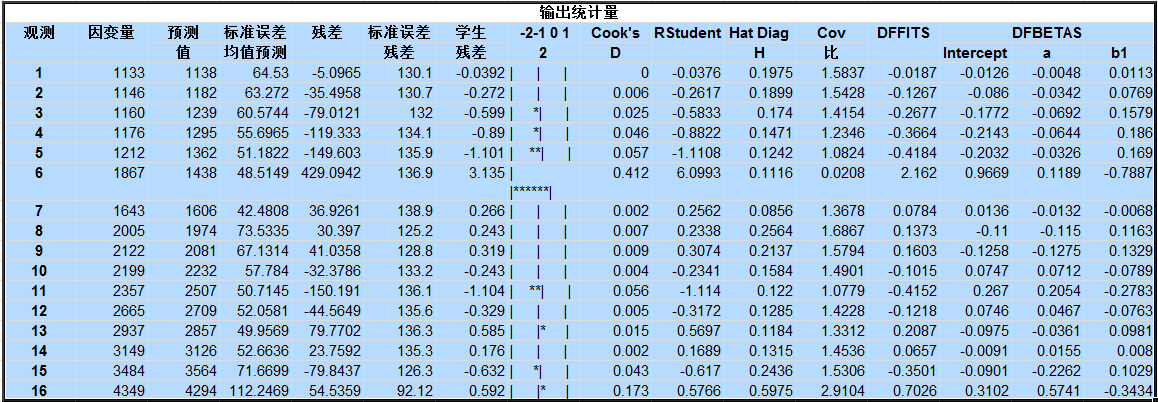


图12——残差分析与预测结果

由残差值得知，第六个观测点的观测值异常，偏离了预测曲线，可以仔细检查数据，看这个原因是什么造成的，是否可以去掉它后再重新进行拟合。第六个观测点的COOK’D统计量为0.412，较大，说明第六个观测点是个强影响点。



用此模型估计出来的每一点的y值和实际值的差的平方之和称为残差平方和，这里的残差平方和为274112。我们希望拟合的残差平方和越小越好。

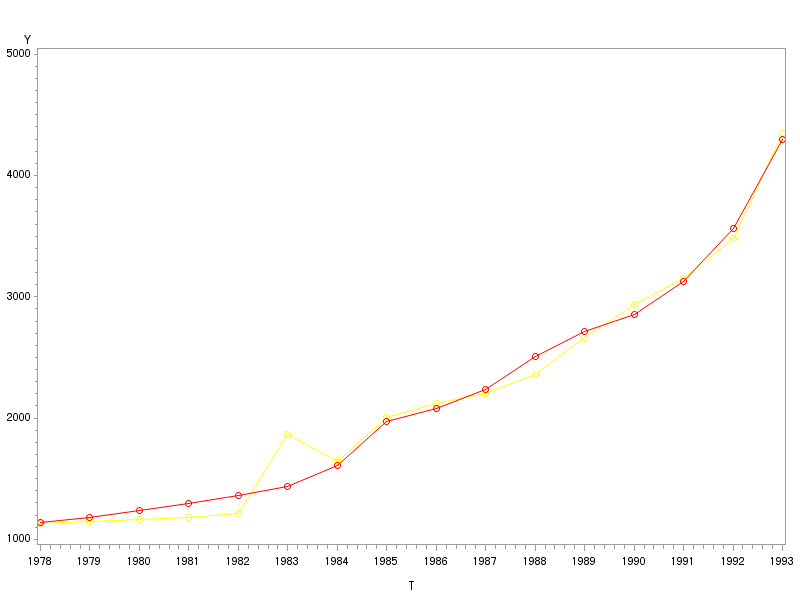


图13——拟合模型预测的曲线图与实际值的曲线图

结果表明，该模型能较好的描述实际变量间的关系。

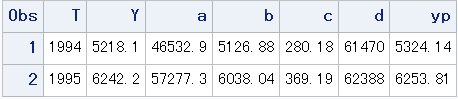


图14

图14是用剩余两年的数据代入拟合模型，得出预测的y值，YP为估计值，而Y为真实值。

模型二

模型二是在模型一的基础上去掉第六个观测点（异常观测点），重新拟合模型。

下面是输出结果。



图1



图2

此模型用不同方法选择的变量为a、b，不同于上一模型。



图3

F统计量是1312.36，相比前面(F=319.58)有很大的改善。误差平方和减少到了62158，拟合优度为0.9954,。这些统计指标都说明这个模型有所改善。



图4

截距项、a和b的Pr>|t|均小于0.0001，说明他们对模型的作用极为显著。拟合得方程：

yp=705.61946+0.06016\*a+0.36553\*b;



图5

最大的条件数减少到了15.17373，说明此模型相对稳定，但仍有较弱的多重共线性。

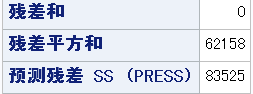


图6

残差平方和由274112减少到了62158。从总体上看，模型二的拟合效果比前面的好。

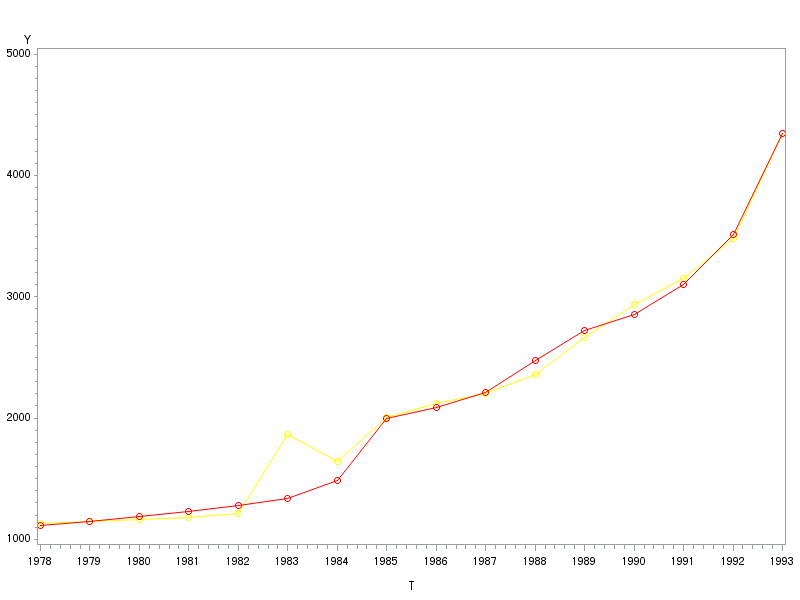


图7

预测曲线与实际曲线吻合度很高。



图8

做预测时，精确度没有上一模型的高，但稳定度较好。