**线性统计模型**

**线性回归与方差分析**

**小组成员：**

**Ⅰ.模型说明**

**一．数据来源：**

本次作业使用的数据为1995——2010年湖北省的X1（总收入：亿元）,X2（人口增长率%），X3（居民消费价格指数增长率%），X4（少儿抚养系数），X5（老年抚养系数），X6（居民消费比重）以及y（消费率%）的对应数值，年份数据用year表示。希望借此来研究消费率与以上各个变量之间的关系（注：数据来自《湖北省统计年鉴》）

**二．软件使用：**

课本的上机作业由R和python完成，所以这次的大作业选择用spss来做

**Ⅱ.构建模型**

**一．选择变量**

**1.使用前进法：**依次引入了x1，x2，最优回归模型为

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **系数a** | | | | | | |
| 模型 | | 非标准化系数 | | 标准系数 | t | Sig. |
| B | 标准 误差 | 试用版 |
| 1 | (常量) | 55.526 | 1.131 |  | 49.109 | .000 |
| 总收入（亿元） | -.004 | .000 | -.965 | -12.235 | .000 |
| 2 | (常量) | 52.497 | .996 |  | 52.686 | .000 |
| 总收入（亿元） | -.004 | .000 | -.902 | -17.599 | .000 |
| 人口增长率（%） | .565 | .132 | .220 | 4.293 | .002 |
| a. 因变量: 消费率（%） | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Anovaa** | | | | | | |
| 模型 | | 平方和 | df | 均方 | F | Sig. |
| 1 | 回归 | 369.043 | 1 | 369.043 | 149.689 | .000b |
| 残差 | 27.119 | 11 | 2.465 |  |  |
| 总计 | 396.163 | 12 |  |  |  |
| 2 | 回归 | 386.623 | 2 | 193.311 | 202.632 | .000c |
| 残差 | 9.540 | 10 | .954 |  |  |
| 总计 | 396.163 | 12 |  |  |  |
| a. 因变量: 消费率（%） | | | | | | |
| b. 预测变量: (常量), 总收入（亿元）。 | | | | | | |
| c. 预测变量: (常量), 总收入（亿元）, 人口增长率（%）。 | | | | | | |

**2.使用后退法**：依次剔除变量x3，x6，x4，x5，最优回归模型与前进法结果一样

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **系数a** | | | | | | |
| 模型 | | 非标准化系数 | | 标准系数 | t | Sig. |
| B | 标准 误差 | 试用版 |
| 1 | (常量) | -8.248 | 45.193 |  | -.182 | .861 |
| 总收入（亿元） | -.006 | .002 | -1.540 | -2.950 | .026 |
| 人口增长率（%） | .928 | .396 | .361 | 2.345 | .057 |
| 居民消费价格指数增长率（%） | .055 | .124 | .044 | .446 | .671 |
| 少儿抚养系数 | -.377 | .507 | -.365 | -.745 | .484 |
| 老年抚养系数 | 9.135 | 6.089 | .468 | 1.500 | .184 |
| 居民消费比重（%） | -.117 | .195 | -.153 | -.599 | .571 |
| 2 | (常量) | -12.042 | 41.770 |  | -.288 | .781 |
| 总收入（亿元） | -.006 | .002 | -1.440 | -3.246 | .014 |
| 人口增长率（%） | .979 | .356 | .381 | 2.750 | .029 |
| 少儿抚养系数 | -.251 | .395 | -.242 | -.635 | .546 |
| 老年抚养系数 | 8.460 | 5.551 | .434 | 1.524 | .171 |
| 居民消费比重（%） | -.058 | .135 | -.076 | -.428 | .681 |
| 3 | (常量) | -10.500 | 39.434 |  | -.266 | .797 |
| 总收入（亿元） | -.005 | .001 | -1.294 | -4.817 | .001 |
| 人口增长率（%） | .911 | .302 | .354 | 3.020 | .017 |
| 少儿抚养系数 | -.102 | .178 | -.099 | -.573 | .582 |
| 老年抚养系数 | 7.137 | 4.370 | .366 | 1.633 | .141 |
| 4 | (常量) | -7.206 | 37.528 |  | -.192 | .852 |
| 总收入（亿元） | -.005 | .001 | -1.187 | -6.408 | .000 |
| 人口增长率（%） | .775 | .180 | .302 | 4.304 | .002 |
| 老年抚养系数 | 6.360 | 3.997 | .326 | 1.591 | .146 |
| 5 | (常量) | 52.497 | .996 |  | 52.686 | .000 |
| 总收入（亿元） | -.004 | .000 | -.902 | -17.599 | .000 |
| 人口增长率（%） | .565 | .132 | .220 | 4.293 | .002 |
| a. 因变量: 消费率（%） | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **已排除的变量a** | | | | | | |
| 模型 | | Beta In | t | Sig. | 偏相关 | 共线性统计量 |
| 容差 |
| 2 | 居民消费价格指数增长率（%） | .044b | .446 | .671 | .179 | .291 |
| 3 | 居民消费价格指数增长率（%） | .004c | .061 | .953 | .023 | .533 |
| 居民消费比重（%） | -.076c | -.428 | .681 | -.160 | .080 |
| 4 | 居民消费价格指数增长率（%） | .011d | .171 | .869 | .060 | .553 |
| 居民消费比重（%） | .023d | .285 | .783 | .100 | .353 |
| 少儿抚养系数 | -.099d | -.573 | .582 | -.199 | .076 |
| 5 | 居民消费价格指数增长率（%） | .027e | .393 | .703 | .130 | .569 |
| 居民消费比重（%） | .029e | .334 | .746 | .111 | .354 |
| 少儿抚养系数 | -.011e | -.064 | .950 | -.021 | .084 |
| 老年抚养系数 | .326e | 1.591 | .146 | .469 | .050 |
| a. 因变量: 消费率（%） | | | | | | |
| b. 模型中的预测变量: (常量), 居民消费比重（%）, 总收入（亿元）, 人口增长率（%）, 老年抚养系数, 少儿抚养系数。 | | | | | | |
| c. 模型中的预测变量: (常量), 总收入（亿元）, 人口增长率（%）, 老年抚养系数, 少儿抚养系数。 | | | | | | |
| d. 模型中的预测变量: (常量), 总收入（亿元）, 人口增长率（%）, 老年抚养系数。 | | | | | | |
| e. 模型中的预测变量: (常量), 总收入（亿元）, 人口增长率（%）。 | | | | | | |

**3.使用逐步回归法：**最后也选择了x1，x2，三种方法的结果一致。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **系数a** | | | | | | |
| 模型 | | 非标准化系数 | | 标准系数 | t | Sig. |
| B | 标准 误差 | 试用版 |
| 1 | (常量) | 55.526 | 1.131 |  | 49.109 | .000 |
| 总收入（亿元） | -.004 | .000 | -.965 | -12.235 | .000 |
| 2 | (常量) | 52.497 | .996 |  | 52.686 | .000 |
| 总收入（亿元） | -.004 | .000 | -.902 | -17.599 | .000 |
| 人口增长率（%） | .565 | .132 | .220 | 4.293 | .002 |
| a. 因变量: 消费率（%） | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **已排除的变量a** | | | | | | |
| 模型 | | Beta In | t | Sig. | 偏相关 | 共线性统计量 |
| 容差 |
| 1 | 人口增长率（%） | .220b | 4.293 | .002 | .805 | .917 |
| 居民消费价格指数增长率（%） | .153b | 2.201 | .052 | .571 | .954 |
| 少儿抚养系数 | .337b | 2.722 | .021 | .652 | .257 |
| 老年抚养系数 | -.320b | -1.382 | .197 | -.400 | .107 |
| 居民消费比重（%） | -.163b | -2.116 | .060 | -.556 | .796 |
| 2 | 居民消费价格指数增长率（%） | .027c | .393 | .703 | .130 | .569 |
| 少儿抚养系数 | -.011c | -.064 | .950 | -.021 | .084 |
| 老年抚养系数 | .326c | 1.591 | .146 | .469 | .050 |
| 居民消费比重（%） | .029c | .334 | .746 | .111 | .354 |
| a. 因变量: 消费率（%） | | | | | | |
| b. 模型中的预测变量: (常量), 总收入（亿元）。 | | | | | | |
| c. 模型中的预测变量: (常量), 总收入（亿元）, 人口增长率（%）。 | | | | | | |

**综上，我们选择x1，x2变量作为构建模型中的自变量。**

**二．回归方程**

未标准化的回归方程为

标准化的回归方程为：

1. **回归系数检验：**

由sig均<0.05,此时回归系数t检验均显著

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **系数a** | | | | | | |
| 模型 | | 非标准化系数 | | 标准系数 | t | Sig. |
| B | 标准 误差 | 试用版 |
| 1 | (常量) | 52.497 | .996 |  | 52.686 | .000 |
| 总收入（亿元） | -.004 | .000 | -.902 | -17.599 | .000 |
| 人口增长率（%） | .565 | .132 | .220 | 4.293 | .002 |
| a. 因变量: 消费率（%） | | | | | | |

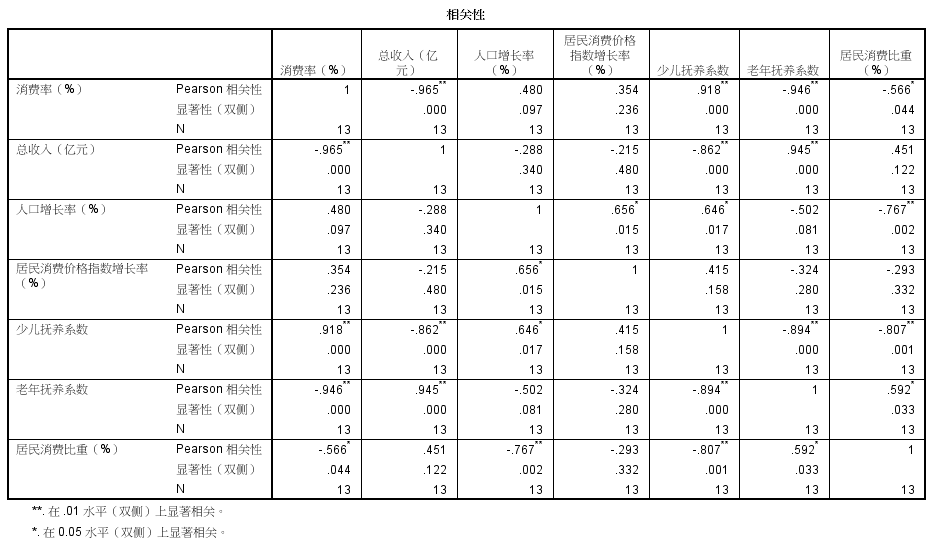
1. **回归方程检验：**

复相关系数R=0.988,决定系数.976,调整的.971，由决定系数看回归方程高度显著。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型汇总b** | | | | |
| 模型 | R | R 方 | 调整 R 方 | 标准 估计的误差 |
| 1 | .988a | .976 | .971 | .97673 |
| a. 预测变量: (常量), 人口增长率（%）, 总收入（亿元）。 | | | | |
| b. 因变量: 消费率（%） | | | | |

方差分析表中，F=202.632，P值=0.000，表明回归方程高度显著。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Anovaa** | | | | | | |
| 模型 | | 平方和 | df | 均方 | F | Sig. |
| 1 | 回归 | 386.623 | 2 | 193.311 | 202.632 | .000b |
| 残差 | 9.540 | 10 | .954 |  |  |
| 总计 | 396.163 | 12 |  |  |  |
| a. 因变量: 消费率（%） | | | | | | |
| b. 预测变量: (常量), 人口增长率（%）, 总收入（亿元）。 | | | | | | |
| **③相关性分析** | | | | | | |



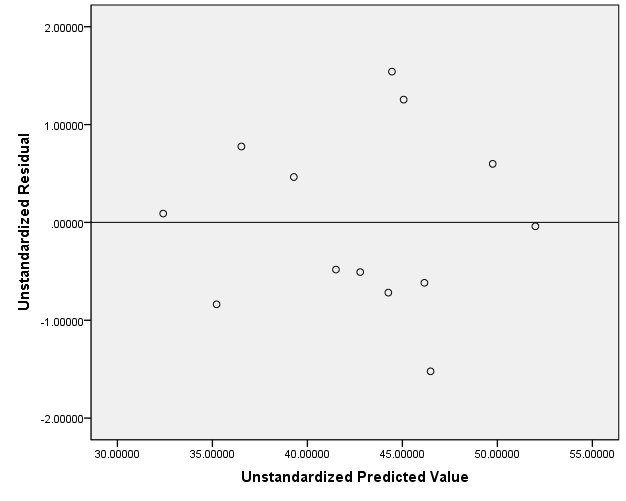
选择皮尔森相关系数，发现，和消费率相关度较大的为总收入、老年抚养系数、少儿抚养系数、居民消费比重；在自变量中，和总收入相关度较大的有老年抚养系数、少儿抚养系数；和人口增长率相关度较大的有居民消费比重、居民消费价格指数增长率、少儿抚养系数；和居民消费价格指数增长率相关度较大的有人口增长率；和少儿抚养系数相关度较大的有老年抚养系数、总收入、居民消费比重、人口增长率；和老年抚养系数相关度较大的有总收入、少儿抚养系数、居民消费比重；和居民消费比重相关度较大的为少年抚养系数、人口增长率、老年抚养系数。这也说明了6个自变量之间可能存在多重共线性。而我们刚才最后用于建模的自变量x1（总收入）与x2（人口增长率）相关性较小。但y与x2的相关性较小。

回归过程中发现，尽管在刚开始选取多个变量时回归系数不显著，但是较大，这可能是由于样本量n=13较小，变量个数较大，导致F检验或t检验的自由度太大，造成的虚假现象，但参数估计的效果非常不稳定。

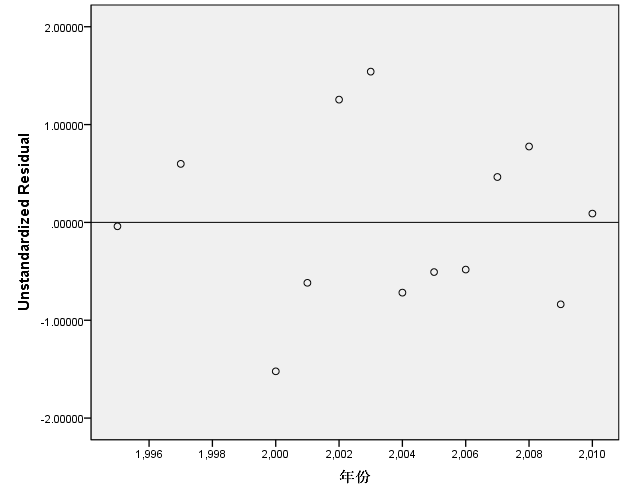
**三．异方差性检验**

**1.残差图分析**

①以拟合值为横坐标，残差为纵坐标



②以观测时间为横坐标，纵坐标为残差



由两种残差图都可以看出，误差项没有明显的异方差性。

**2.等级相关系数**

计算残差的绝对值后，对，x1，x2计算等级相关系数，残差绝对值和x2人口增长率的相关系数为-0.571，和x1总收入的相关系数为-0.082，出现了异方差，所以选择x2构造权函数。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **相关系数** | | | | | |
|  | | | 总收入（亿元） | 人口增长率（%） | cancha |
| Spearman 的 rho | 总收入（亿元） | 相关系数 | 1.000 | -.071 | -.082 |
| Sig.（双侧） | . | .817 | .789 |
| N | 13 | 13 | 13 |
| 人口增长率（%） | 相关系数 | -.071 | 1.000 | -.571\* |
| Sig.（双侧） | .817 | . | .041 |
| N | 13 | 13 | 13 |
| cancha | 相关系数 | -.082 | -.571\* | 1.000 |
| Sig.（双侧） | .789 | .041 | . |
| N | 13 | 13 | 13 |
| \*. 在置信度（双测）为 0.05 时，相关性是显著的。 | | | | | |

利用SPSS软件进行权函数估计，幂函数范围为[-5,5]之间的整数，结果如下：

|  |  |  |
| --- | --- | --- |
| **对数似然值b** | | |
| 幂 | -5.000 | -22.745 |
| -4.500 | -20.857 |
| -4.000 | -19.057 |
| -3.500 | -17.511 |
| -3.000 | -16.349 |
| -2.500 | -15.628 |
| -2.000 | -15.324a |
| -1.500 | -15.357 |
| -1.000 | -15.636 |
| -.500 | -16.083 |
| .000 | -16.640 |
| .500 | -17.266 |
| 1.000 | -17.933 |
| 1.500 | -18.618 |
| 2.000 | -19.300 |
| 2.500 | -19.967 |
| 3.000 | -20.613 |
| 3.500 | -21.250 |
| 4.000 | -21.905 |
| 4.500 | -22.614 |
| 5.000 | -23.402 |
| a. 选择对应幂以用于进一步分析，因为它可以使对数似然函数最大化。 | | |
| b. 因变量: y，源变量: x2 | | |

|  |  |
| --- | --- |
| **模型摘要** | |
| 复相关系数 | .996 |
| R 方 | .991 |
| 调整 R 方 | .989 |
| 估计的标准误 | 3.062 |
| 对数似然函数值 | -15.324 |

在m=-2时对数似然函数达到极大，因而幂指数m的最优取值为m=-2，加权最小二乘法的=0.991，F值=562.966，而普通最小二乘法的 =.976 ,F值=202.632。这说明加权最小二乘估计的效果好于普通最小二乘估计的效果。

加权后的回归结果是：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **系数a,b** | | | | | | |
| 模型 | | 非标准化系数 | | 标准系数 | t | Sig. |
| B | 标准 误差 | 试用版 |
| 1 | (常量) | 51.917 | 1.027 |  | 50.528 | .000 |
| 总收入（亿元） | -.004 | .000 | -.789 | -19.554 | .000 |
| 人口增长率（%） | .639 | .094 | .274 | 6.785 | .000 |
| a. 因变量: 消费率（%） | | | | | | |
| b. 加权的最小二乘回归 - 按 x 进行加权 | | | | | | |
|  | | | | | | |
| 图3：加权最小二乘残差图（左图），普通最小二乘残差图（右图） | | | | | | |

做出加权最小二乘残差图（左图），并与普通最小二乘残差图（右图）进相比较，差异不是很明显。

所以加权最小二乘的回归方程是：

普通最小二乘的回归方程是：

**四．自相关性**

①加权后的DW统计量值=2.335，误差项呈现负自相关

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **模型汇总b,c** | | | | | |
| 模型 | R | R 方 | 调整 R 方 | 标准 估计的误差 | Durbin-Watson |
| 1 | .996a | .991 | .989 | 3.06237 | 2.335 |
| a. 预测变量: (常量), 人口增长率（%）, 总收入（亿元）。 | | | | | |
| b. 因变量: 消费率（%） | | | | | |
| c. 加权的最小二乘回归 - 按 x 进行加权 | | | | | |
|  | | | | | |

②普通最小二乘回归DW统计量值=1.983，误差项呈现正自相关

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **模型汇总b** | | | | | |
| 模型 | R | R 方 | 调整 R 方 | 标准 估计的误差 | Durbin-Watson |
| 1 | .988a | .976 | .971 | .97673 | 1.983 |
| a. 预测变量: (常量), 人口增长率（%）, 总收入（亿元）。 | | | | | |
| b. 因变量: 消费率（%） | | | | | |
| 由于本题n=13，数据量太少，无法使用DW表 | | | | | |

**五．多重共线性检验**

**①普通最小二乘：**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **系数a** | | | | | | | | |
| 模型 | | 非标准化系数 | | 标准系数 | t | Sig. | 共线性统计量 | |
| B | 标准 误差 | 试用版 | 容差 | VIF |
| 1 | (常量) | 52.497 | .996 |  | 52.686 | .000 |  |  |
| 总收入（亿元） | -.004 | .000 | -.902 | -17.599 | .000 | .917 | 1.090 |
| 人口增长率（%） | .565 | .132 | .220 | 4.293 | .002 | .917 | 1.090 |
| a. 因变量: 消费率（%） | | | | | | | | |

看到VIF均小于10，说明自变量之间多重共线性不显著。

1. **加权最小二乘回归**

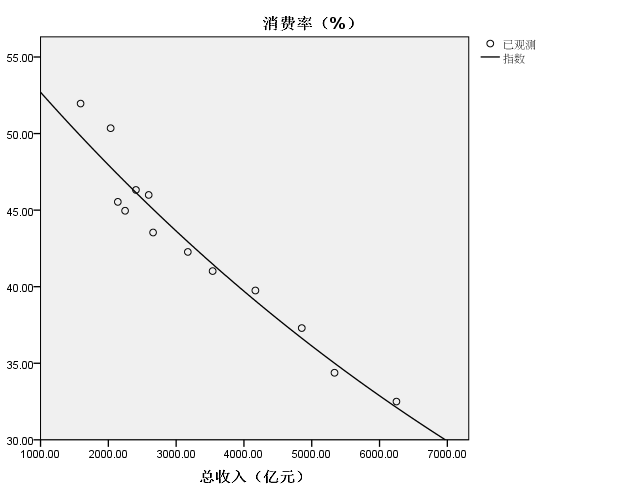
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **系数a,b** | | | | | | | | |
| 模型 | | 非标准化系数 | | 标准系数 | t | Sig. | 共线性统计量 | |
| B | 标准 误差 | 试用版 | 容差 | VIF |
| 1 | (常量) | 51.917 | 1.027 |  | 50.528 | .000 |  |  |
| 总收入（亿元） | -.004 | .000 | -.789 | -19.554 | .000 | .540 | 1.851 |
| 人口增长率（%） | .639 | .094 | .274 | 6.785 | .000 | .540 | 1.851 |
| a. 因变量: 消费率（%） | | | | | | | | |
| b. 加权的最小二乘回归 - 按 x 进行加权 | | | | | | | | |

同样，VIF均小于10，说明自变量之间多重共线性不显著。

**六．非线性回归**

1. 用SPSS对数据y对x1进行了曲线拟合

曲线回归中，调整后的R方最大为指数、增长、复合、对数模型时，其值为0.951，没有上面建立的二元线性回归拟合效果好。下面仅选择指数模型进行呈现。



1. **尝试多项式回归方法，**

分别构造

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **模型汇总c** | | | | | | | | | | |
| 模型 | R | R 方 | 调整 R 方 | 标准 估计的误差 | 更改统计量 | | | | | Durbin-Watson |
| R 方更改 | F 更改 | df1 | df2 | Sig. F 更改 |
| 1 | .965a | .932 | .925 | 1.57016 | .932 | 149.689 | 1 | 11 | .000 |  |
| 2 | .990b | .979 | .975 | .90841 | .048 | 22.863 | 1 | 10 | .001 | 2.092 |
| a. 预测变量: (常量), 总收入（亿元）。 | | | | | | | | | | |
| b. 预测变量: (常量), 总收入（亿元）, x22。 | | | | | | | | | | |
| c. 因变量: 消费率（%） | | | | | | | | | | |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **系数a** | | | | | | | | | | 模型 | | 非标准化系数 | | 标准系数 | t | Sig. | 共线性统计量 | | | B | 标准 误差 | 试用版 | 容差 | VIF | | 1 | (常量) | 55.526 | 1.131 |  | 49.109 | .000 |  |  | | 总收入（亿元） | -.004 | .000 | -.965 | -12.235 | .000 | 1.000 | 1.000 | | 2 | (常量) | 53.340 | .798 |  | 66.836 | .000 |  |  | | 总收入（亿元） | -.004 | .000 | -.878 | -17.873 | .000 | .863 | 1.159 | | x22 | .053 | .011 | .235 | 4.782 | .001 | .863 | 1.159 | | a. 因变量: 消费率（%） | | | | | | | | | | | | | | | | | | | |

调整后的R方仍然没有加权时的大，但是比普通最小二乘回归的调整后R方大，所以也是一个较好的拟合，回归方程为：

**综上：**

|  |  |  |
| --- | --- | --- |
|  | 回归方程 | 调整后R方 |
| 普通最小二乘线性回归 |  | 0.971 |
| 加权最小二乘线性回归 |  | 0.989 |
| 多项式回归 |  | 0.975 |