**相关分析与回归分析**

SPSS第一组

张世翔，马家豪，张祎畅，马裕翔，王婧怡，王斯亮。

**1.看起来壮就真的壮吗？（张世翔）**

**抽样调查了 20 个人的生理指标（分别为体重 x1、腰围 x2 和脉搏 x3），同时获得了这 20 人的训练指标（分别为引体向上次数 y1、起坐次数 y2 和跳跃次数 y3），据此针对生理指标和训练指标进行典型相关分析（数据见文件“健康和训练.xlsx”）。**

**问题：**

**（1）列出两组数据所有典型变量的表达式；**

1. **两组指标的相关性有多大；**

**（3）分析所有显著的典型变量所表达的意义。**

**答：**

在SPSS中，通过【分析-相关-典型相关性】工具，进行典型相关分析。得到结果：

**【第1问】**根据计算结果，可以得到两组变量的头三个典型变量以及表达式（标准化之后的变量）

|  |  |
| --- | --- |
| 对于第一组变量而言： | 对于第二组变量而言： |
| V1= 0.480x1-1.376x2+0.039x3 | W1= -0.345y1+1.321y2-0.555y3 |
| V2= -1.961x1+1.364x2-0.311x3 | W2= -0.878y1-0.319y2+1.032y3 |
| V3= 0.338x1-0.625x2-1.031x3 | W3= 0.566y1-0.122y2+0.718y3 |

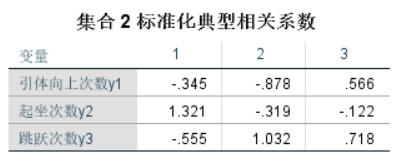
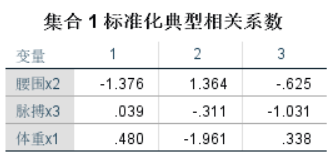


图1 典型变量标准系数

**【第2-3问】**

在回答这两问之前，首先阐述一下典型相关的计算过程与结果：

**第一步**，**计算各个变量之间的相关性（图2）**。可以发现对于第一组变量而言，变量腰围（x2）和体重（x1）相关性较高，说明二者包含的信息有重叠部分；对于第二组变量而言，变量仰卧起坐次数（y2 ）和跳跃次数（y3）相关性较高，且显著，说明二者包含的信息有重叠部分。对于两组变量之间而言，起坐次数（y2）和腰围（x2）、体重（x1）相关性系数都较高，且显著，说明两者之间存在一定相关性。



图2 各个变量之间的相关系数

**第二步，计算典型相关性（图3）。**可以发现，只有第一对典型相关系数通过了0.05的显著性检验。其中第一个典型变量的相关系数为0.805，p=0.033<0.05，显著，**因此只需要对第一个典型相关变量进行解释**。另外两对典型变量的相关系数较小，且不显著，不加以赘述。



图3 典型相关性

**第三步，根据第一问的典型相关系数表（图1）**，可以得到典型变量的表达式V1= 0.480x1-1.376x2+0.039x3 ,W1= -0.345y1+1.321y2-0.555y3,且V1和W1二者的相关系数为0.805。可以看到典型变量V1，主要受腰围影响，且腰围起到负面作用；典型变量W1，主要受仰卧起坐次数影响，且仰卧起坐次数起到正面作用。

**第四步，计算典型载荷系数和交叉载荷系数（图4）**。

典型载荷系数表（左）说明，生理指标维度的第一典型变量与体重的相关系数为-0.732，呈负相关；与腰围的相关系数为-0.972，成负相关；与脉搏的相关系数为0.349，呈正相关。其中，生理指标维度的第一典型变量与腰围的相关系数最强。从训练指标维度来看，其第一典型变量与引体向上次数的相关系数为-0.070，与仰卧起坐次数相关系数为0.821，与跳跃次数相关系数为0.194，说明训练指标维度和它各个变量之间主要还是呈正相关的。

交叉载荷系数表（右）说明，腰围、脉搏、体重和集合2的第一个典型变量的相关系数分别是 -0.782、0.281和-0.589；引体向上次数、仰卧起坐次数和跳跃次数和集合1的第一个典型变量的相关系数分别为-0.056，0.660和0.156。

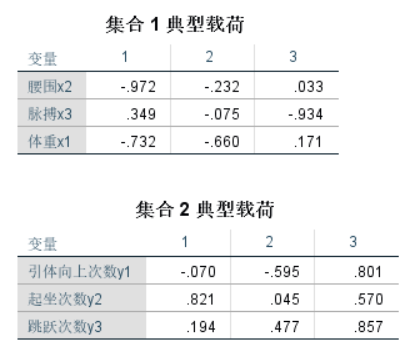
 

图4 典型载荷和交叉载荷系数

**第五步，计算已解释的方差比例（图5）**，说明各典型变量对各变量组方差解释的比例（组内代表比例和交叉解释比例）。集合1（生理指标维度）被自身的第一典型变量解释了53.4%；集合2（训练指标维度）被自身的第一典型变量解释了23.9%；集合1被集合2的第一典型变量解释了34.6%；集合2被集合1的第一典型变量解释了15.5%。



图5 已解释的方差比例

**综上所述：**

**【第二问回答】**从第二步可知，两组变量的典型相关系数为0.805。

**【第三问回答】**显著的典型变量为两组变量的第一典型变量，表达式为：V1= 0.480x1-1.376x2+0.039x3 ,W1= -0.345y1+1.321y2-0.555y3,且V1和W1二者的相关系数为0.805。可以看到典型变量V1，主要受腰围影响，且腰围起到负面作用；典型变量W1，主要受仰卧起坐次数影响，且仰卧起坐次数起到正面作用。

同时，可以发现，集合1（生理指标维度）被自身的第一典型变量解释了53.4%；集合2（训练指标维度）被自身的第一典型变量解释了23.9%；集合1被集合2的第一典型变量解释了34.6%；集合2被集合1的第一典型变量解释了15.5%。

1. **GDP和投资有关系吗？有多大关系？（马家豪）**

**我国省级行政区相关经济统计数据见文件“chn\_economics.xlsx”之表 2**

**要求：**

**(1)计算国内生产总值（GDP）与固定资产投资的相关系数，并进行显著性检验。**

**答：**

在SPSS中，通过【分析-相关-双变量相关】工具，进行相关分析。得到结果：

首先，从图6可以发现，参与分析的两个变量样本数都为30，GDP的均值为1921.0927亿元，标准差为1474.80603亿元；固定资产投资的平均值为528.1750亿元，标准差为407.91027亿元。



图6 描述统计

然后，计算得出（图7），二者相关系数为0.964，显著性水平为0.000（双尾检验），因此在相关系数旁以两个“\*\*”标识，说明国内生产总值（GDP）与固定资产投资相关性十分显著。



图7 相关系数以及显著性检验

**(2)以工业总产值为控制变量，计算国内生产总值（GDP）和固定资产投资的偏相关系数， 并进行显著性检验。**

答：

在SPSS中，通过【分析-相关-偏相关】工具，进行偏相关分析。得到结果：

首先，从图8可以发现，参与分析的两个变量样本数都为30，GDP的均值为1921.0927亿元，标准差为1474.80603亿元；固定资产投资的平均值为528.1750亿元，标准差为407.91027亿元。而控制变量工业总产值的平均值为3069.1243，标准差为3011.70662。



图8 描述统计

然后，计算得出（图9），在排除工业总产值这一影响后，国内生产总值（GDP）和固定资产投资二者偏相关系数为0.533，相较未排除工业总产值这一因素时有所下降。但由于p-value为0.003（双尾检验），小于0.05，说明国内生产总值（GDP）与固定资产投资二者相关性依旧显著。



图9 偏相关系数以及显著性检验

**3.预测：明年产多少粮食？（张祎畅）**

**逐年的粮食总产量如文件“粮食产量变化 data.xlsx”。要求:**

**(1)判断该地区的粮食生产发展趋势是否接受直线型关系?**

在SPSS中制作散点图，由散点图中的拟合线可以看出年份编号和产量**大致呈线性关系，可以考虑做线性回归预测**。（第二问会有具体的检验方式）

操作步骤：

Step1：将数据导进SPSS

Step2：点击“图形”——“旧对话框”——“散点图”——“简单散点图”画出散点图

Step3：双击散点图激活——点击“元素”——“总计拟合线”作出拟合线

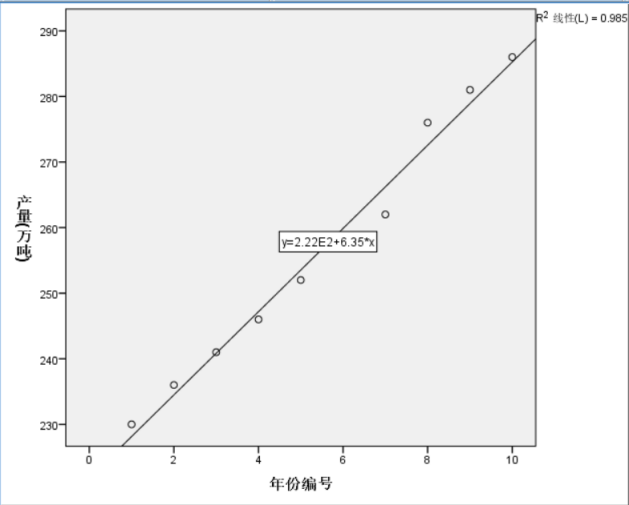


图10 散点图

除此之外，根据皮尔逊相关性可以发现，二者相关系数为0.992，且显著，说明二者高度线性相关，也进一步说明了该地区的粮食生产发展趋势接受直线型关系



图11 相关性

**(2)以粮食产量为因变量，年份为自变量,完成预测年粮食总产量的回归分析,并进行评价。**

在SPSS中进行线性回归分析，将年份作为自变量，产量作为因变量，勾选相关参数，进行分析。

从模型摘要可以看出回归标准差为2.556，决定系数R2为0.985，调整后的R2为0.983，即认为模型你和很好，解释变量具有98.3%的解释能力。



图12 模型摘要

从方差分析表ANOVA中看到，F=508.564,显著性近似为0，说明y对x的线性回归高度显著，模型具有统计意义。



图13 方差分析表

从回归模型的系数及其区间估计和系数显著性检验来看，β0置信水平为95%的区间估计为（217.774，225.826），β1置信水平为95%的区间估计为（5.697，6.994）。同时可以发现，回归系数β1检验的t值为22.551，显著性近似为0，则认为对回归系数β1的估计高度显著，**也说明两变量直线关系显著，解答了第一问的问题**。也可以得到回归方程为y=221.800+6.345x。



图14 回归系数

最后，制作残差散点图。可以发现，回归标准化残差在区间（-2.2）内波动，说明拟合程度好。

**操作步骤：**

Step1：计算残差值保存为“\*ZPRED”（标准化预测值），预测值保存为“\*ZRESID”（标准化残差）

Step2：点击“分析”——“回归”——“线性”

Step3：点击“图”——将“\*ZPRED ”作为Y，将“ \*ZRESID”作为X，画出标准化残差的散点图

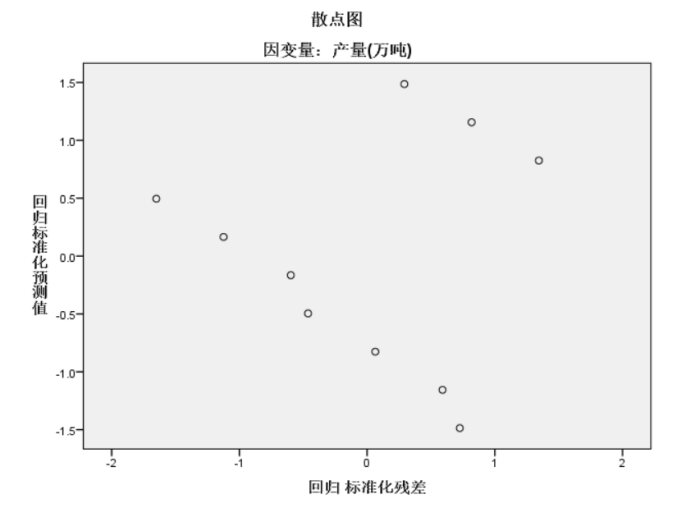


图15 残差散点图

**(3)预测第 12 年的粮食生产水平和预测区间。**

以xp预测yp，其预测区间为，计算方式如下：

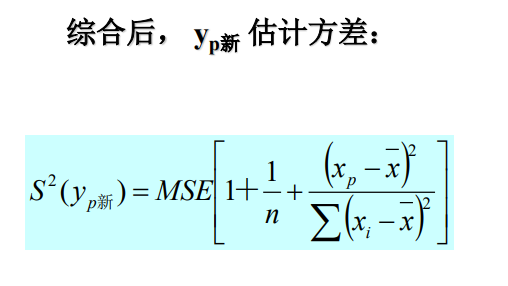
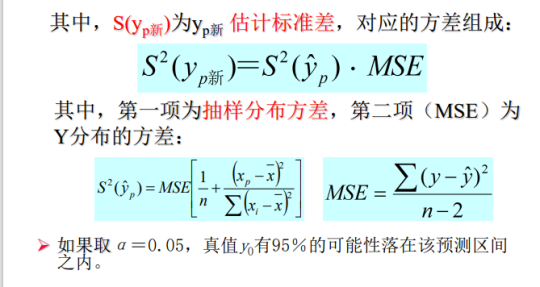
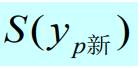


图16 计算方法

由回归方程，y=221.800+6.345x，计算得到，当x=12时，y的估计值为297.94。经过计算，MSE=6.532,n=10，t0.05，8=2.306，=42.25，=82.5，则可以计算出=6.352\*（1+1/10+42.25/82.5）=10.5301，则=3.2450，因此，=2.306\*3.2450=7.483，预测区间为，也就是（290.457，305.423）。

**综上所述，第12年粮食水平的预测值为297.94万吨，预测区间为（290.457，305.423）。**

**4.建立多元最优回归模型（马裕翔）**

**有四个自变量（x1,x2,x3,x4）对因变量 y 产生影响，数据见文件“多自变量的回归分析 data.xls”。建立关于因变量 y 的回归模型，要求：**

在SPSS软件中，采用【分析-回归-线性回归】工具，输入因变量y，自变量x1,x2,x3,x4，采用逐步分析的方法进行线性回归，其他参数设置参考老师的课程ppt。



图17 逐步线性回归

**(1)建立多元最优回归模型(说明理由，包括回归模型各种检验评价，残差分析，共线性诊断等)。**

**【结果1：描述性统计和相关性】**

根据描述统计表可以看到因变量和各个自变量的平均值、标准偏差、个案数等统计指标。



图18 描述统计

根据相关性表格可以得到各变量间的Pearson相关系数和统计检验结果。如果各自间的相关系数过大，提示有多重共线的可能。



图19 相关性表

**【结果2：模型纳入和剔除的变量】**

由于在操作时选用的是Stepwise，共建立过两个回归模型，纳入2各变量（x2和x3），默认纳入标准P≤0.5，剔除标准≤0.1。

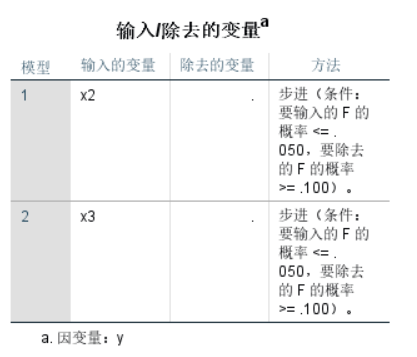


图20 输入和剔除的变量表

**【结果3：模型概要与方差分析】**

结果显示：**最终的模型复相关系数R=0.728,所有自变量于y之间的回归关系比较密切；R2=0.530，说明在y的总变异中，最终模型中2个自变量可以解释的变异占53.0%；与只纳入x2相比，校正后的R方有所增加，标准估算的误差在减小，说明拟合效果越来越好。纳入x3后，R2的改变也有统计学意义，且更加显著。**

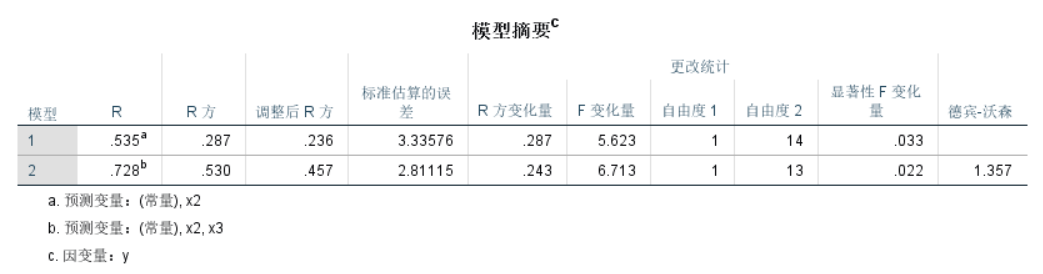


图21 模型概要表

方差分析表ANOVA显示，最终回归模型F=7.315,P<0.01，至少有一个自变量的回归系数不为0，回归模型具有统计学意义。

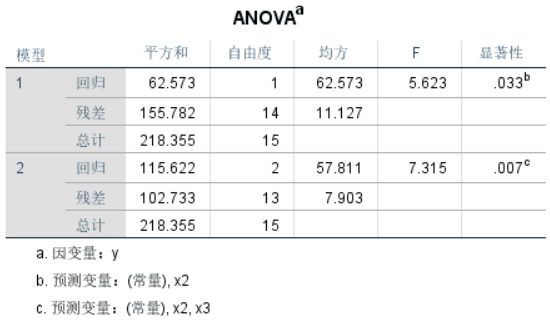


图22 ANOVA表

**【结果4：系数】**

**根据系数表格可以发现，纳入模型的各自变量偏回归系数均不为0，且在α=0.05的条件下显著，最终回归模型为：y=17.162+0.083x2-1.632x3。**

标准化回归系数β去掉了不同自变量单位不同的影响，是利用标准化数据标计算而来，在有统计学意义的前提下，标准化回归系数的绝对值越大，对应自变量对因变量Y的影响越大。其意为固定其他自变量，自变量每改变1个标准差，因变量改变的标准差个数。**由于0.517>0.493，x2对y的影响大于x3对y的影响。**

对于共线性统计量容差和VIF，一般一般容差不小于0.1，VIF（容差的倒数）不大于10可说明自变量不存在共线的情况，本例两个自变量Tolerance=0.999，VIF=1.001，**可以认为不存在共线的情况**。

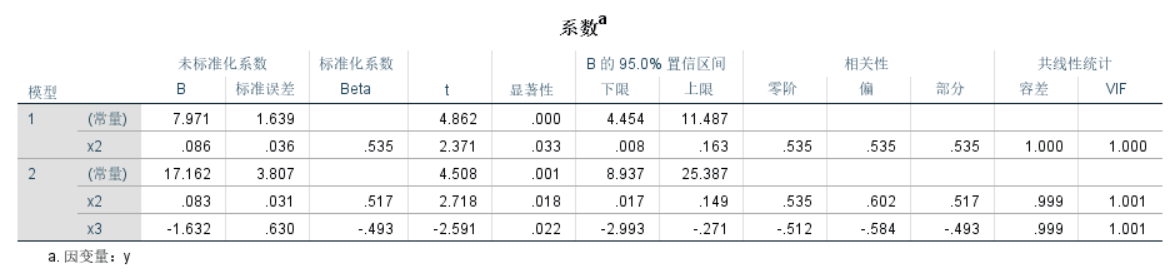


图23 系数表

**【结果5：共线性诊断】**

除了在系数表中输出共线性诊断统计量Tolerance和VIF，共线性诊断还提供了特征根（Elgenvalue）、条件指数（Condition Index）及变异构成（Variance Proportions）。条件指数是最大特征根与每个连续特征根比值的平方根，比值＞15提示可能存在共线性的问题，＞30则表明存在共线性。变异构成（方差比例）是回归模型中各项（包括常数项）变异能被主成分解释的比例，如某主成分对两个或两个以上的自变量贡献均较大（如＞0.5），则提示这几个变量存在一定的共线性。

**结果表明：**最终进入模型的两个自变量x2，x3基本不存在共线性。



图24 共线性诊断

**【结果6：残差统计量和残差正态分布考察】**

根据残差统计量和残差正态分布的直方图和pp图可以发现，残差基本上符合正态分布。



图25 残差统计

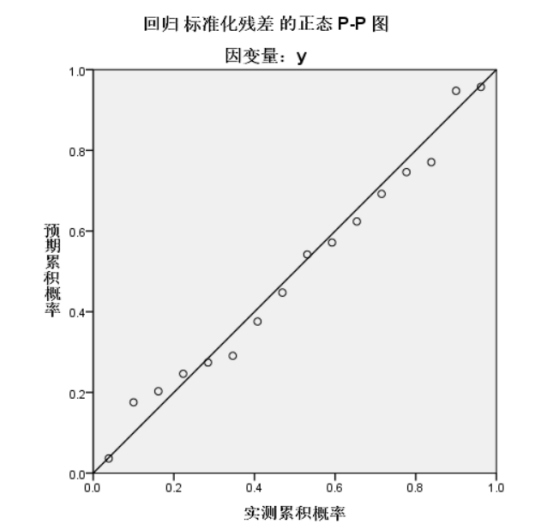
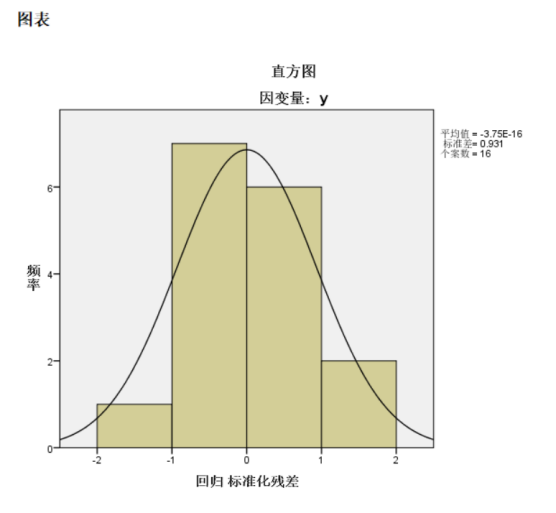
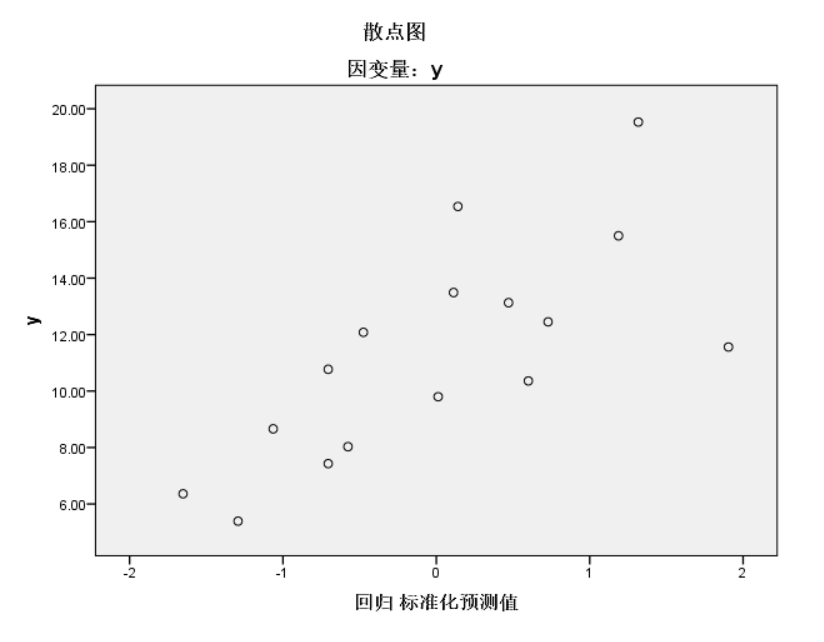


图26-27 直方图和pp图

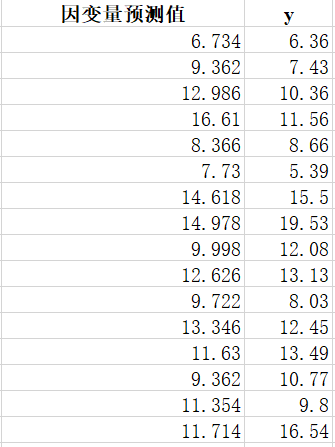
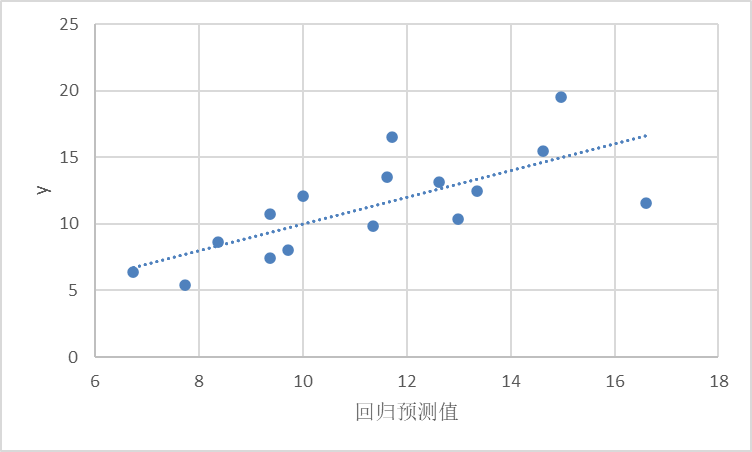
**综上所述，最终回归模型为：y=17.162+0.083x2-1.632x3。**

**(2)针对选出的最佳回归模型，画出因变量预测值和观测值得散点图。**

在spss中可以直接输出回归标准化预测值和实际观测值的散点图（图28），也可以在excel中利用公式计算得到预测值和实际观测值制作散点图（图29）。



（图28 spss散点图）

（图29 数据和excel 散点图）

**5.建立最佳非线性回归模型（王婧怡）**

**以 x1 和 x2 为自变量，y 为因变量，建立非线性回归模型，数据文件见“非线性拟合Data.xlsx”。要求：**

1. **确定最佳非线性回归模型的结构；**

首先，先在Excel中制作x1-y和x2-y的散点图，可以看出因变量y和自变量x1、x2均不存在线性关系，一个自变量也不对应唯一因变量，因此需要对x1和x2进行合并运算。

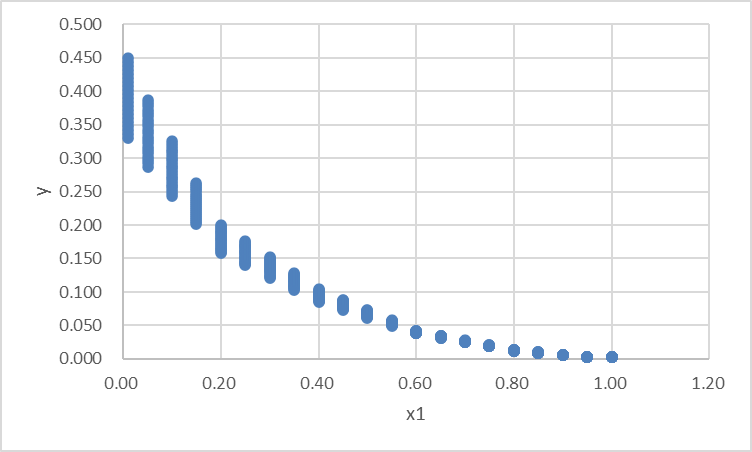
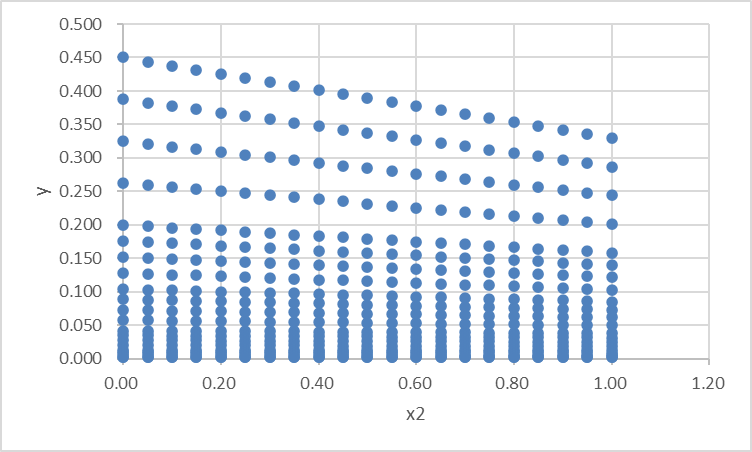


图30-31 x1-y和x2-y的散点图

对x1和x2尝试合并运算，可以初步判断x1和x2的组合形式应该为θ=a\*x1+b\*x2。

表1 变量及散点图

|  |  |  |  |
| --- | --- | --- | --- |
| 变量 | x1+x2 | x1\*x2 | x1+x2-x1\*x2 |
| 散点图 |  |  |  |
| 变量 | exp（x1）-x2 | x1+0.1\*x2 | x1+0.05\*x2 |
| 散点图 |  |  |  |

取θ=x1+0.05\*x2和y进行非线性回归，结果如下，可以发现三次多项式效果最优，回归模型的结构如下：

y=b0+b1\*(m\*x1+n\*x2)+b2\*(m\*x1+n\*x2)2+b3\*(m\*x1+n\*x2)3。

表2 回归模型

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 对数函数 | 指数函数 | 幂函数 | 三次多项式 |
| R2 | 0.991 | 0.983 | 0.845 | 0.997 |
| 显著性 | 0.000 | 0.000 | 0.000 | 0.000 |

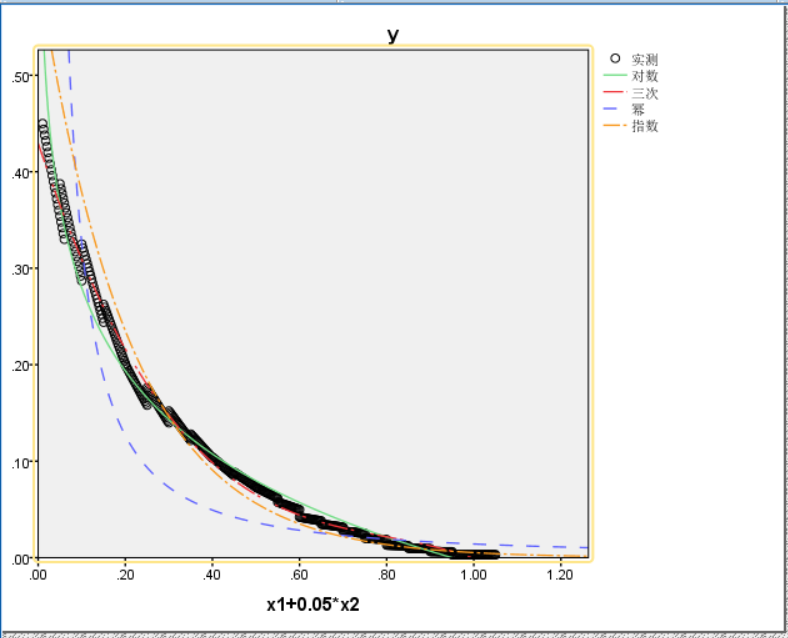


图32 拟合曲线

1. **以迭代法求解，给出回归模型和回归拟合程度；**

在SPSS使用【分析-回归-非线性回归】工具，进行迭代法求解，设置最大迭代次数为30次 。

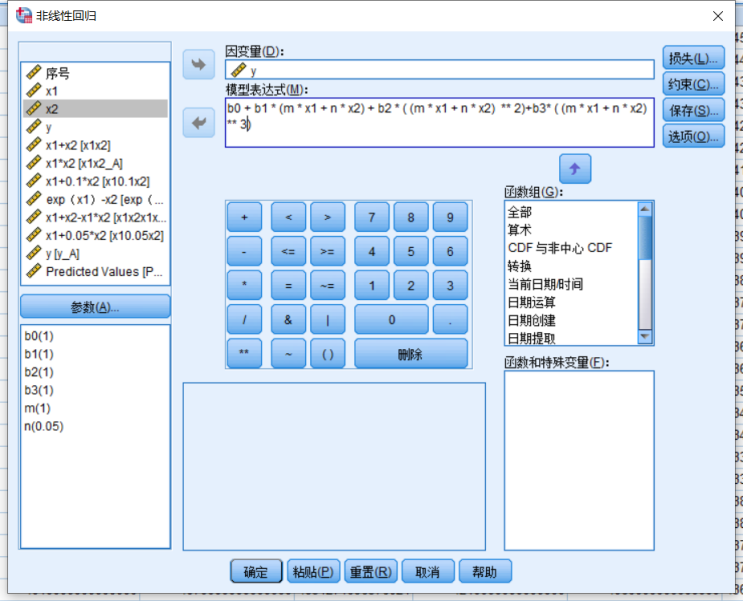


图33-34 非线性回归模型以及参数设置

计算结果如下，可以得到迭代历史记录、参数估算值、参数估算值相关性以及ANOVA表。根据迭代历史记录可以发现，进行了39次模型评估和15次导数评估后才停止迭代。可以得到参数的估计值：b0=0.445，b1=-2.658，b2=5.839，b3=-4.557，m=0.520，n=0.040，最终回归模型为：

y=0.445-2.658\*(0.520\*x1+0.040\*x2)+5.839\*(0.520\*x1+0.040\*x2)2-4.557\*(0.520\*x1+0.040\*x2)3。

除此之外，根据ANOVA表中的R2=0.996，拟合度为0.996，说明此模型能够解释99.6%的变异，拟合度非常高。

|  |
| --- |
|  |
| 图35 迭代历史记录 |



图36 参数估计值及其相关性

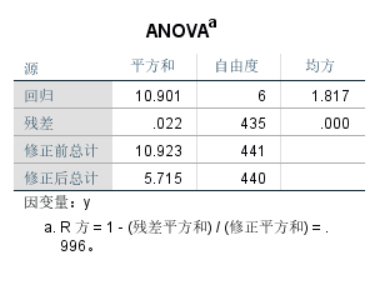


图38 ANOVA表

1. **画出预测值与观测值 y 的散点图。**

根据计算得出的预测值，在spss中将其与观测值y制作散点图，如下，可以发现其分布大致沿y=x直线，说明拟合程度良好**。**

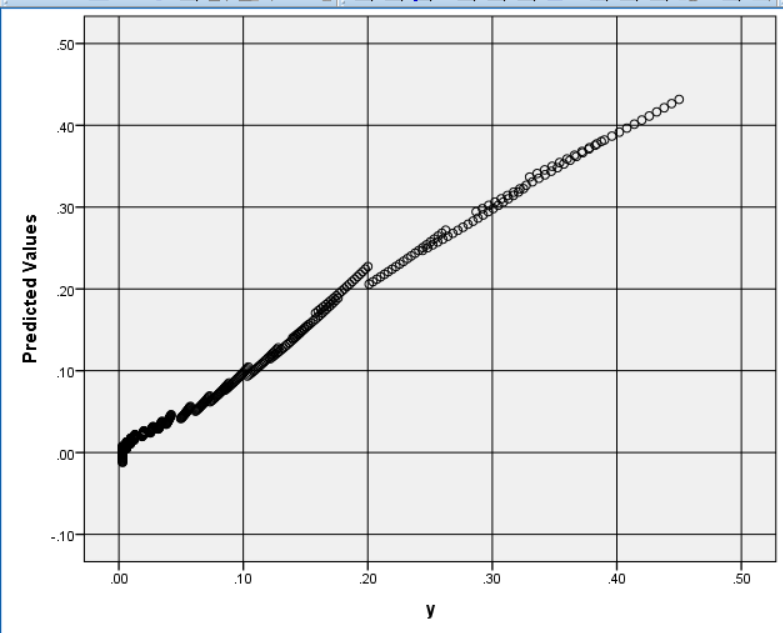


图 39 预测值与观测值 y 的散点图

**6.27年城市化进程：建立模型（王斯亮）**

**市镇人口与乡村人口的比值称为城乡人口比，现有我国 1977~2003 年 27 年的城乡人口比数据，试以时间为自变量，以城乡人口比为因变量，建立一个预测模型（见数据文件“城乡人口变化”）并进行拟合评价。如果所建模型为内线性模型，要求分别采用线性化和迭代求解方法求解，并进一步比较两种方法拟合结果的优劣。**

第一步，先做出年份和城乡人口比的散点图，观察其大致关系。可以发现，二者呈非线性关系。由于年份的数量级和城乡人口比的数量级差距较大，应当考虑对年份进行对数、倒数等处理，因此在spss中选择了倒数方程、S形曲线以及对数方程三种方式。（其实我勾选了所有的非线性函数类型，但是回过头来想想，这三种更考虑了数据本身的特征，因此下边的叙述仅考虑这三种模型。）

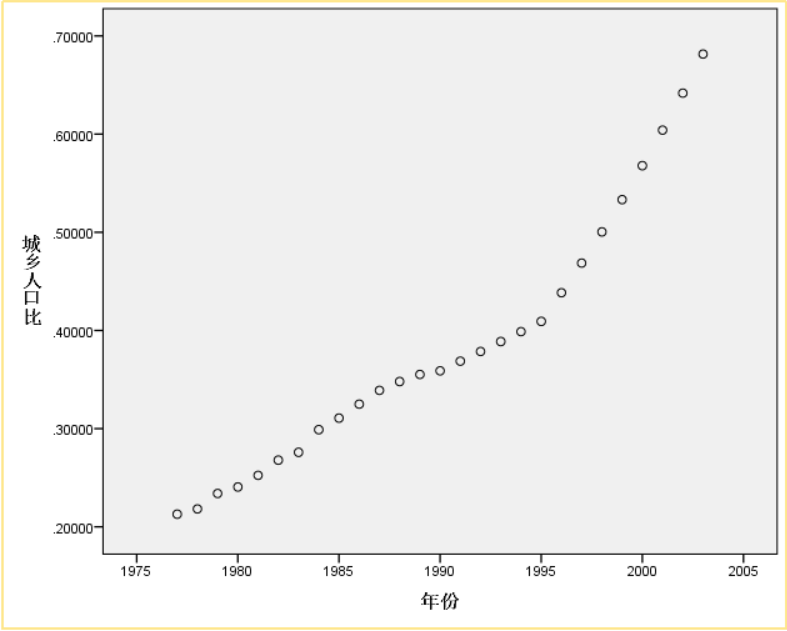


图40 散点图

第二步，采用【分析-回归-曲线估算】工具，进行计算。可以得到各个模型的R2以及显著性。可以发现S型方程效果最优。其表达形式为： 。

表3 回归模型

|  |  |  |  |
| --- | --- | --- | --- |
|  | 对数方程 | 倒数方程 | S型方程 |
| R2 | 0.929 | 0.928 | 0.980 |
| 显著性 | 0.000 | 0.000 | 0.000 |

由上述表达式可知，该模型是内线性模型，在spss中，对其的处理是将其转化为线性模型，令Y=lny，X=1/x，求解Y和X的线性模型Y=b0+b1X。根据返回的结果可以发现，b0=81.153，b1=-163493.516，则表达式为Y=81.153+-163493.516\*X，对于x和y而言，最终表达式为y=exp(81.153-163493.516/x)，R2为0.980。

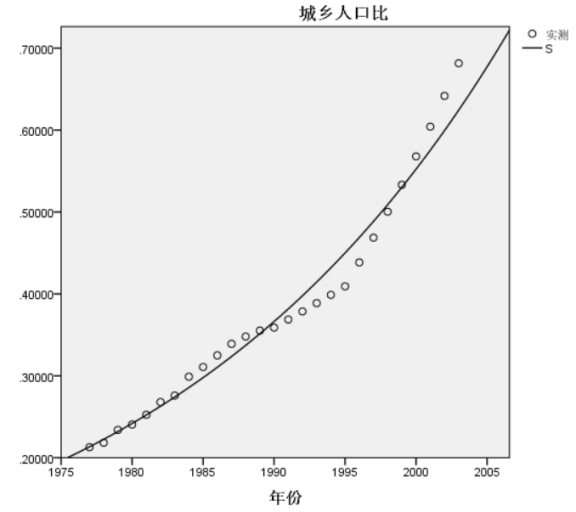
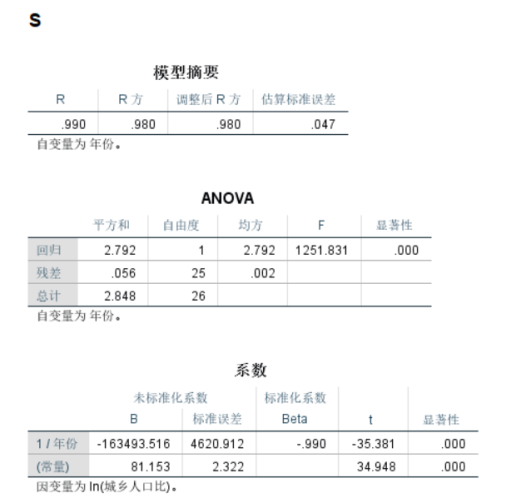


图42 S型方程模型结果

第三步，进行迭代法计算。采用【分析-回归-非线性模型】工具，进行计算，在运行16次模型评估和8次倒数评估后停止。可以得到模型的各个参数以及R2。最终表达式为y=exp(85.642-172436.716/x)，R2为0.975。



图43 非线性模型参数设置

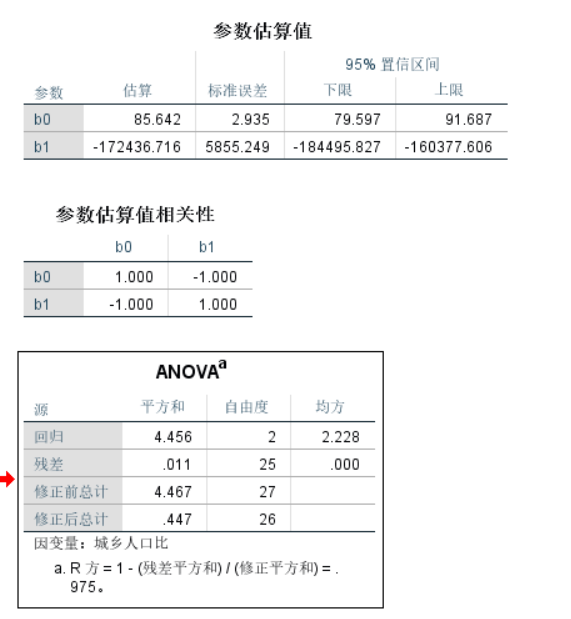
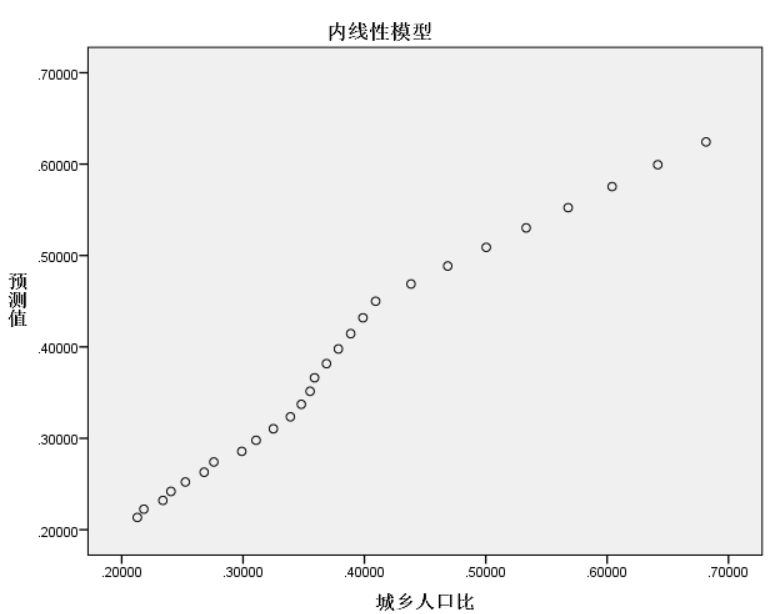


图44 迭代法模型结果

结合散点图可以发现，二者拟合程度大体一致，基本看不出什么区别，因此可以比较R2来进行判断。



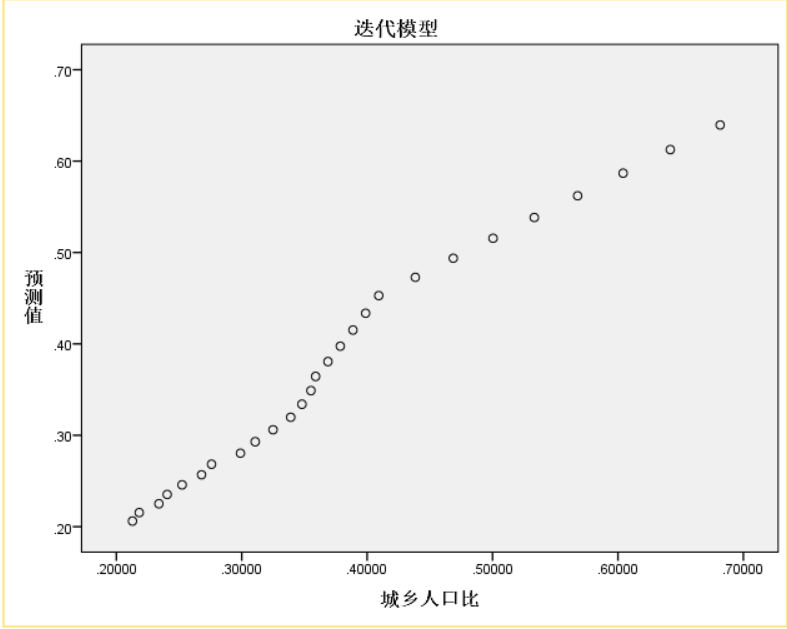


图45-46 两种方法散点图

综上所述，根据内线性方法和迭代方法的决定系数R2知（0.980>0.975），内线性模型的拟合效果更好。