

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 x_1 、腰围 x_2 和脉搏 x_3 ），同时获得了这 20 人的训练指标（分别为引体向上次数 y_1 、起坐次数 y_2 和跳跃次数 y_3 ），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；
- (2) 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：

$$V_1 = 0.480x_1 - 1.376x_2 + 0.039x_3$$

$$V_2 = -1.961x_1 + 1.364x_2 - 0.311x_3$$

$$V_3 = 0.338x_1 - 0.625x_2 - 1.031x_3$$

对于第二组变量而言：

$$W_1 = -0.345y_1 + 1.321y_2 - 0.555y_3$$

$$W_2 = -0.878y_1 - 0.319y_2 + 1.032y_3$$

$$W_3 = 0.566y_1 - 0.122y_2 + 0.718y_3$$

集合 1 标准化典型相关系数

变量	1	2	3
腰围 x_2	-1.376	1.364	-.625
脉搏 x_3	.039	-.311	-1.031
体重 x_1	.480	-1.961	.338

集合 2 标准化典型相关系数

变量	1	2	3
引体向上次数y1	-.345	-.878	.566
起坐次数y2	1.321	-.319	-.122
跳跃次数y3	-.555	1.032	.718

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。

相关性 ^a						
	腰围(x2)	脉搏(x3)	体重(x1)	引体向上次数(y1)	起坐次数(y2)	跳跃次数(y3)
腰围(x2)	皮尔逊相关性	1	-.353	.870	.105	-.646
	显著性(双尾)		.127	.000	.661	.002
脉搏(x3)	皮尔逊相关性	-.353	1	-.366	-.010	.225
	显著性(双尾)	.127		.113	.967	.340
体重(x1)	皮尔逊相关性	.870	-.366	1	.184	-.493
	显著性(双尾)	.000	.113		.439	.027
引体向上次数(y1)	皮尔逊相关性	.105	-.010	.184	1	.372
	显著性(双尾)	.661	.967	.439		.106
起坐次数(y2)	皮尔逊相关性	-.646	.225	-.493	.372	1
	显著性(双尾)	.002	.340	.027	.106	.001
跳跃次数(y3)	皮尔逊相关性	-.191	.035	-.226	.389	.669
	显著性(双尾)	.419	.884	.337	.090	1

a. 成列 N=20

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。

典型相关性							
相关性	特征值	威尔克统计	F	分子自由度	分母自由度	显著性	
1	.805	1.840	.306	2.379	9.000	34.223	.033
2	.360	.149	.870	.539	4.000	30.000	.708
3	.008	.000	1.000	.001	1.000	16.000	.974

H0 for Wilks 检验是指当前行和后续行中的相关性均为零

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 $V_1 = 0.480x_1 - 1.376x_2 + 0.039x_3$, $W_1 = -0.345y_1 + 1.321y_2 - 0.555y_3$, 且 V_1 和 W_1 二者的相关系数为 0.805。可以看到典型变量 V_1 , 主要受腰围影响, 且腰围起到负面作用; 典型变量 W_1 , 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；引体向上的相关系数分别为-0.056，0.660和0.156。

集合 1 典型载荷				集合 1 交叉载荷			
变量	1	2	3	变量	1	2	3
腰围x2	- .972	- .232	.033	腰围x2	- .782	- .084	.000
脉搏x3	.349	- .075	- .934	脉搏x3	.281	- .027	- .008
体重x1	- .732	- .660	.171	体重x1	- .589	- .238	.001

集合 2 典型载荷				集合 2 交叉载荷			
变量	1	2	3	变量	1	2	3
引体向上次数y1	- .070	- .595	.801	引体向上次数y1	- .056	- .214	.007
起坐次数y2	.821	.045	.570	起坐次数y2	.660	.016	.005
跳跃次数y3	.194	.477	.857	跳跃次数y3	.156	.172	.007

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%。

典型变量	集合 1 * 自身	已解释的方差比例		集合 2 * 集合 1
		集合 1 * 集合 2	集合 2 * 自身	
1	.534	.346	.239	.155
2	.165	.021	.194	.025
3	.301	.000	.567	.000

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： $V_1 = 0.480x_1 - 1.376x_2 + 0.039x_3$, $W_1 = -0.345y_1 + 1.321y_2 - 0.555y_3$, 且 V_1 和 W_1 二者的相关系数为0.805。可以看到典型变量 V_1 ，主要受腰围影响，且腰围起到负面作用；典型变量 W_1 ，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%。

2. 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。

描述统计			
	平均值	标准差	个案数
国内生产总值	1921.0927	1474.80603	30
固定资产投资	528.1750	407.91027	30

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。

相关性			
		国内生产总值	固定资产投资
国内生产总值	皮尔逊相关性	1	.964**
	显著性(双尾)		.000
	平方和与叉积	63076532.02	16815813.03
	协方差	2175052.828	579855.622
	个案数	30	30
固定资产投资	皮尔逊相关性	.964**	1
	显著性(双尾)	.000	
	平方和与叉积	16815813.03	4825332.842
	协方差	579855.622	166390.788
	个案数	30	30

**. 在 0.01 级别(双尾)，相关性显著。

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。

描述统计

	平均值	标准差	个案数
国内生产总值	1921.0927	1474.80603	30
固定资产投资	528.1750	407.91027	30
工业总产值	3069.1243	3011.70662	30

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。

		相关性		
控制变量		国内生产总值	固定资产投资	工业总产值
- 无 - ^a	国内生产总值	相关性	1.000	.964
		显著性（双尾）	.	.000
		自由度	0	28
	固定资产投资	相关性	.964	1.000
		显著性（双尾）	.000	.
		自由度	28	0
	工业总产值	相关性	.958	.963
		显著性（双尾）	.000	.
		自由度	28	28
工业总产值	国内生产总值	相关性	1.000	.533
		显著性（双尾）	.	.003
		自由度	0	27
	固定资产投资	相关性	.533	1.000
		显著性（双尾）	.003	.
		自由度	27	0

a. 单元格包含零阶（皮尔逊）相关性。

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

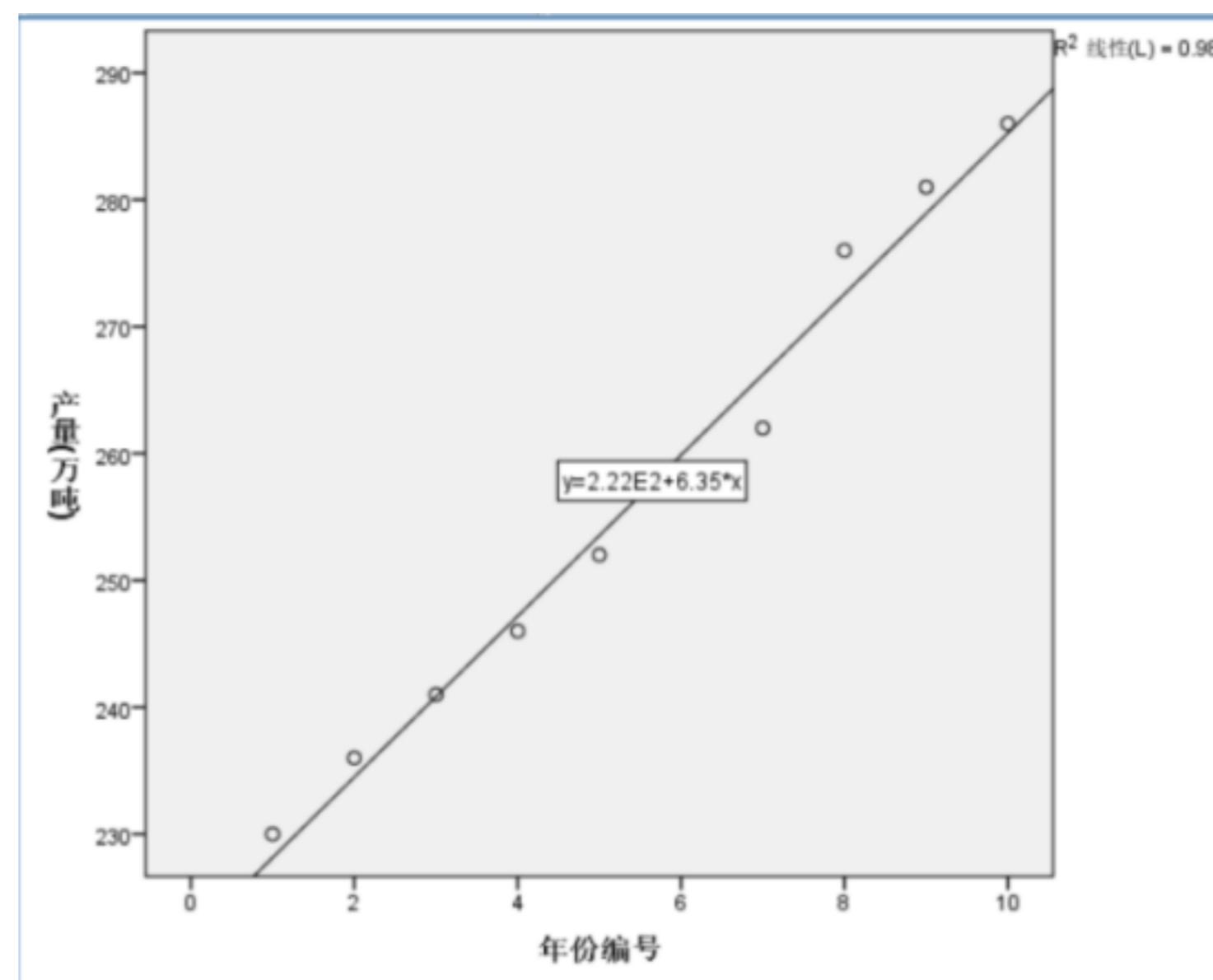


图10 散点图

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

相关性			
	产量(万吨)	年份编号	
皮尔逊相关性	产量(万吨)	1.000	.992
	年份编号	.992	1.000
显著性 (单尾)	产量(万吨)	.	.000
	年份编号	.000	.
个案数	产量(万吨)	10	10
	年份编号	10	10

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。

模型	R	R 方	调整后 R 方	标准估算的误差	更改统计					
					R 方变化量	F 变化量	自由度 1	自由度 2	显著性 F 变化量	德宾-沃森
1	.992 ^a	.985	.983	2.556	.985	508.564	1	8	.000	1.334

a. 预测变量: (常量), 年份编号
b. 因变量: 产量(万吨)

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。

ANOVA ^a					
模型	平方和	自由度	均方	F	显著性
1 回归	3321.845	1	3321.845	508.564	.000 ^b
残差	52.255	8	6.532		
总计	3374.100	9			

a. 因变量: 产量(万吨)
b. 预测变量: (常量), 年份编号

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, β_0 置信水平为95%的区间估计为(217.774, 225.826), β_1 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 β_1 检验的t值为22.551, 显著性近似为0, 则认为对回归系数 β_1 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 $y=221.800+6.345x$ 。

模型	系数 ^a											
	未标准化系数		标准化系数		t	显著性	B 的 95.0% 置信区间		零阶	相关性 偏	部分	共线性统计 容差
1	(常量)	221.800	1.746	Beta	t	显著性	下限	上限				
	年份编号	6.345	.281	.992	22.551	.000	5.697	6.994	.992	.992	.992	1.000

a. 因变量: 产量(万吨)

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

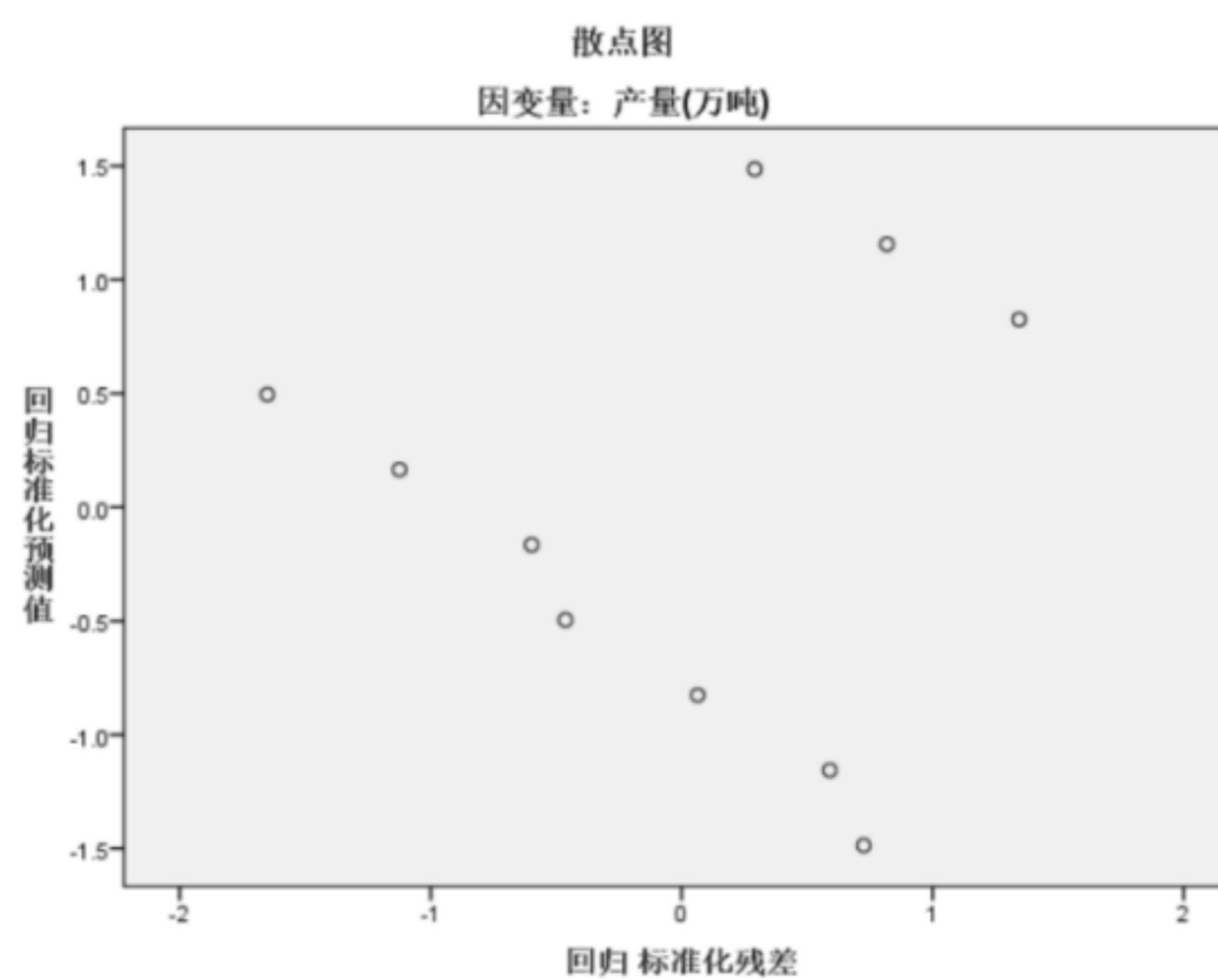


图15 残差散点图

(3)预测第12年的粮食生产水平和预测区间。

以 x_p 预测 y_p , 其预测区间为 $\hat{y}_p \pm t_{\alpha, n-2} \cdot S(y_{p\text{新}})$, 计算方式如下:

其中, $S(y_{p\text{新}})$ 为 $y_{p\text{新}}$ 估计标准差, 对应的方差组成:

$$S^2(y_{p\text{新}}) = S^2(\hat{y}_p) \cdot MSE$$

其中, 第一项为抽样分布方差, 第二项(MSE)为Y分布的方差:

$$S^2(\hat{y}_p) = MSE \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right] \quad MSE = \frac{\sum(y - \hat{y})^2}{n-2}$$

如果取 $\alpha = 0.05$, 真值 y_0 有95%的可能性落在该预测区间之内。

综合后, $y_{p\text{新}}$ 估计方差:

$$S^2(y_{p\text{新}}) = MSE \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right]$$

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

$$297.94。经过计算, MSE=6.532, n=10, t_{0.05, 8}=2.306, (x_p - \bar{x})^2 = 42.25,$$

$$\sum(x_i - \bar{x})^2 = 82.5, \text{ 则可以计算出 } S^2(y_{p\text{新}}) = 6.352 * (1 + 1/10 + 42.25/82.5)$$

$$= 10.5301, \text{ 则 } S(y_{p\text{新}}) = 3.2450, \text{ 因此, } t_{\alpha, n-2} \cdot S(y_{p\text{新}}) = 2.306 * 3.2450 = 7.483,$$

预测区间为 $\hat{y}_p \pm t_{\alpha, n-2} \cdot S(y_{p\text{新}})$, 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 (x_1, x_2, x_3, x_4) 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 x_1, x_2, x_3, x_4 , 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。

描述统计			
	平均值	标准偏差	个案数
y	11.3175	3.81536	16
x1	27.50	15.275	16
x2	39.00	23.799	16
x3	5.56	1.153	16
x4	7.50	2.582	16

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。

相关性						
	y	x1	x2	x3	x4	
皮尔逊相关性	y	1.000	.297	.535	-.512	.224
	x1	.297	1.000	.000	.085	.000
	x2	.535	.000	1.000	-.036	.000
	x3	-.512	.085	-.036	1.000	-.056
	x4	.224	.000	.000	-.056	1.000
显著性 (单尾)	y		.132	.016	.021	.202
	x1	.132		.500	.377	.500
	x2	.016	.500		.447	.500
	x3	.021	.377	.447		.418
	x4	.202	.500	.500	.418	
个案数	y	16	16	16	16	16
	x1	16	16	16	16	16
	x2	16	16	16	16	16
	x3	16	16	16	16	16
	x4	16	16	16	16	16

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。

输入/除去的变量 ^a			
模型	输入的变量	除去的变量	方法
1	x2	.	步进(条件： 要输入的 F 的 概率 <=. 050, 要除去 的 F 的概率 ≥ .100)。
2	x3	.	步进(条件： 要输入的 F 的 概率 <=. 050, 要除去 的 F 的概率 ≥ .100)。

a. 因变量: y

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。

模型	R	R 方	调整后 R 方	标准估算的误差	更改统计					德宾-沃森
					R 方变化量	F 变化量	自由度 1	自由度 2	显著性 F 变化量	
1	.535 ^a	.287	.236	3.33576	.287	5.623	1	14	.033	
2	.728 ^b	.530	.457	2.81115	.243	6.713	1	13	.022	1.357

a. 预测变量: (常量), x2

b. 预测变量: (常量), x2, x3

c. 因变量: y

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。

ANOVA ^a						
模型		平方和	自由度	均方	F	显著性
1	回归	62.573	1	62.573	5.623	.033 ^b
	残差	155.782	14	11.127		
	总计	218.355	15			
2	回归	115.622	2	57.811	7.315	.007 ^c
	残差	102.733	13	7.903		
	总计	218.355	15			

a. 因变量: y

b. 预测变量: (常量), x2

c. 预测变量: (常量), x2, x3

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 $\alpha=0.05$ 的条件下显著，最终回归模型为： $y=17.162+0.083x_2-1.632x_3$ 。

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 $0.517>0.493$ ， x_2 对y的影响大于 x_3 对y的影响。

对于共线性统计量容差和VIF，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。

模型	系数 ^a										
	未标准化系数		标准化系数		t	显著性	B 的 95.0% 置信区间		相关性		
	B	标准误差	Beta	t			下限	上限	零阶	偏	部分
1	(常量)	7.971	1.639		4.862	.000	4.454	11.487			
	x2	.086	.036	.535	2.371	.033	.008	.163	.535	.535	.535
2	(常量)	17.162	3.807		4.508	.001	8.937	25.387			
	x2	.083	.031	.517	2.718	.018	.017	.149	.535	.602	.517
	x3	-1.632	.630	-.493	-2.591	.022	-2.993	-.271	-.512	-.584	-.493

a. 因变量: y

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（Eigenvalue）、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。

共线性诊断 ^a						
模型	维	特征值	条件指标	方差比例		
				(常量)	x2	x3
1	1	1.861	1.000	.07	.07	
	2	.139	3.658	.93	.93	
2	1	2.789	1.000	.00	.03	.00
	2	.192	3.812	.02	.92	.04
	3	.019	12.218	.97	.05	.96

a. 因变量: y

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。

残差统计 ^a					
	最小值	最大值	平均值	标准偏差	个案数
预测值	6.7318	16.6049	11.3175	2.77635	16
残差	-5.04486	4.83196	.00000	2.61703	16
标准预测值	-1.652	1.904	.000	1.000	16
标准残差	-1.795	1.719	.000	.931	16

a. 因变量: y

图25 残差统计

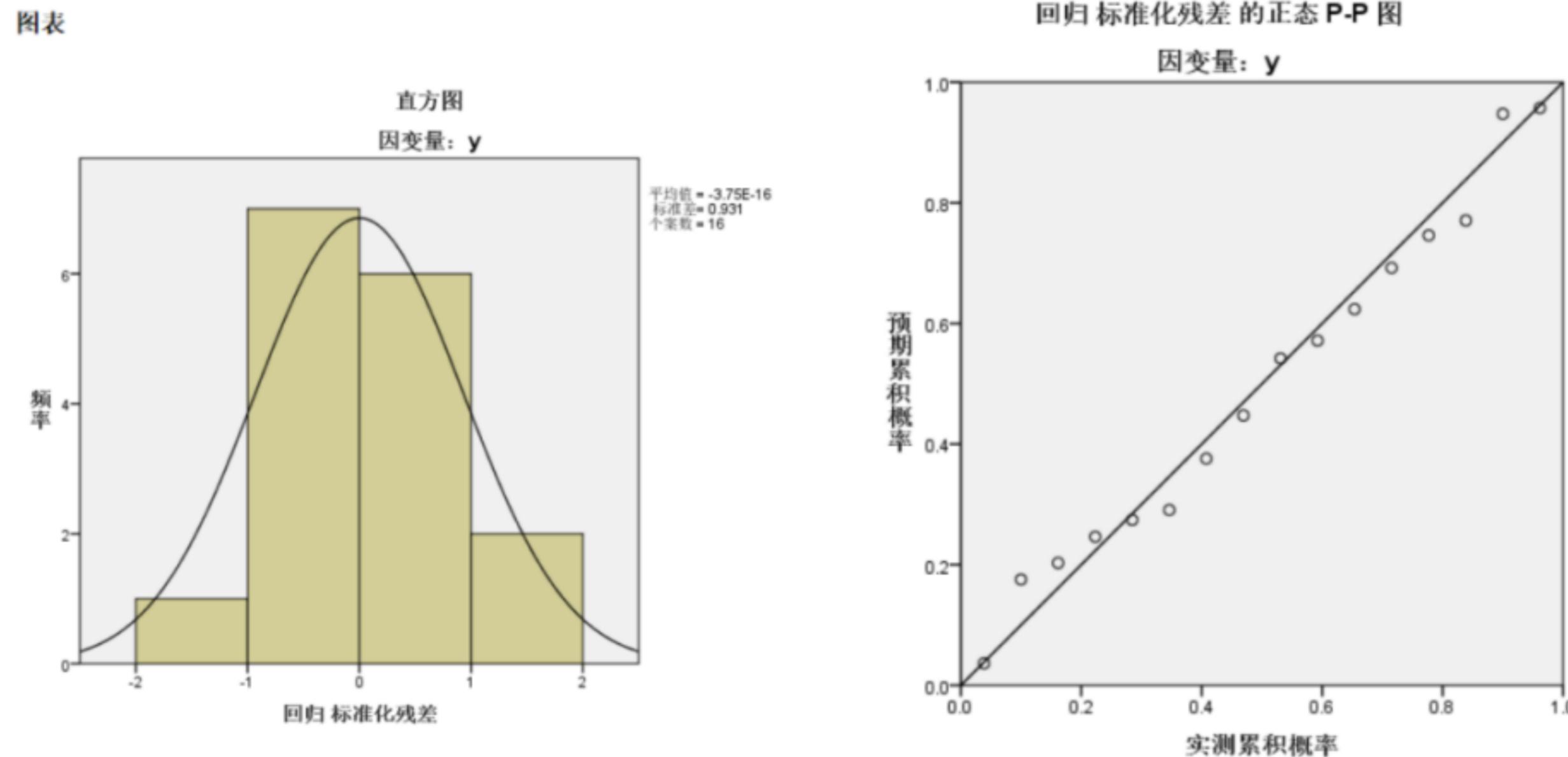
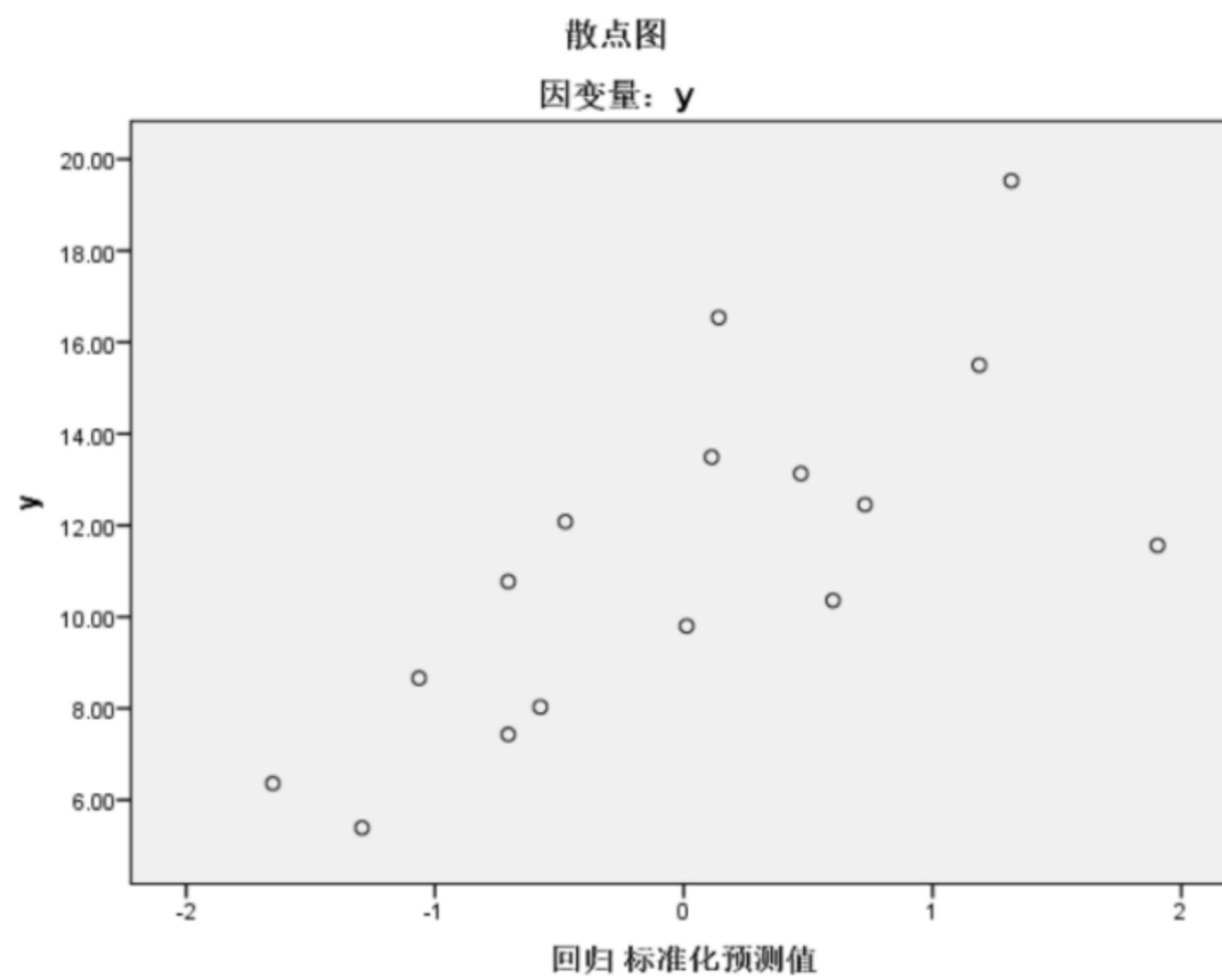


图26-27 直方图和pp图

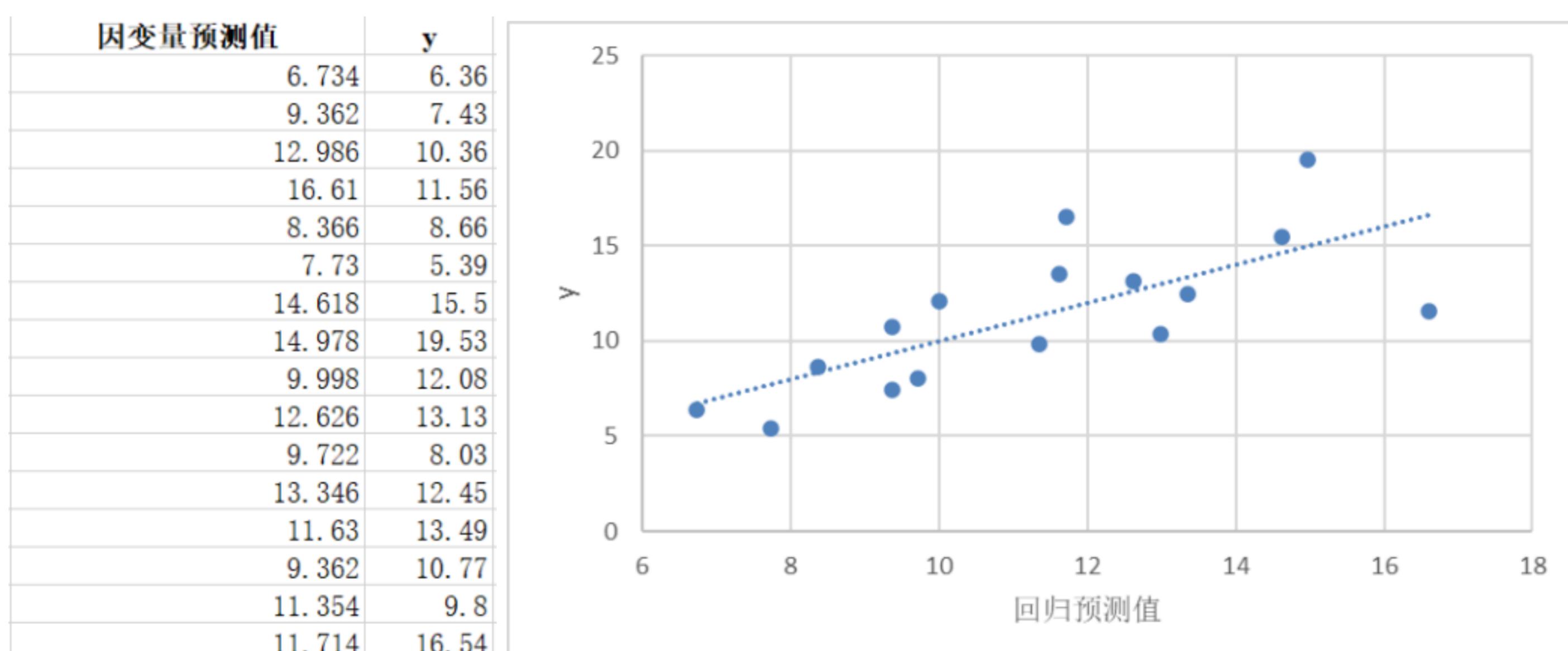
综上所述，最终回归模型为： $y=17.162+0.083x_2-1.632x_3$ 。

(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 x_1 和 x_2 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 x_1-y 和 x_2-y 的散点图，可以看出因变量 y 和自变量 x_1 、 x_2 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 x_1 和 x_2 进行合并运算。

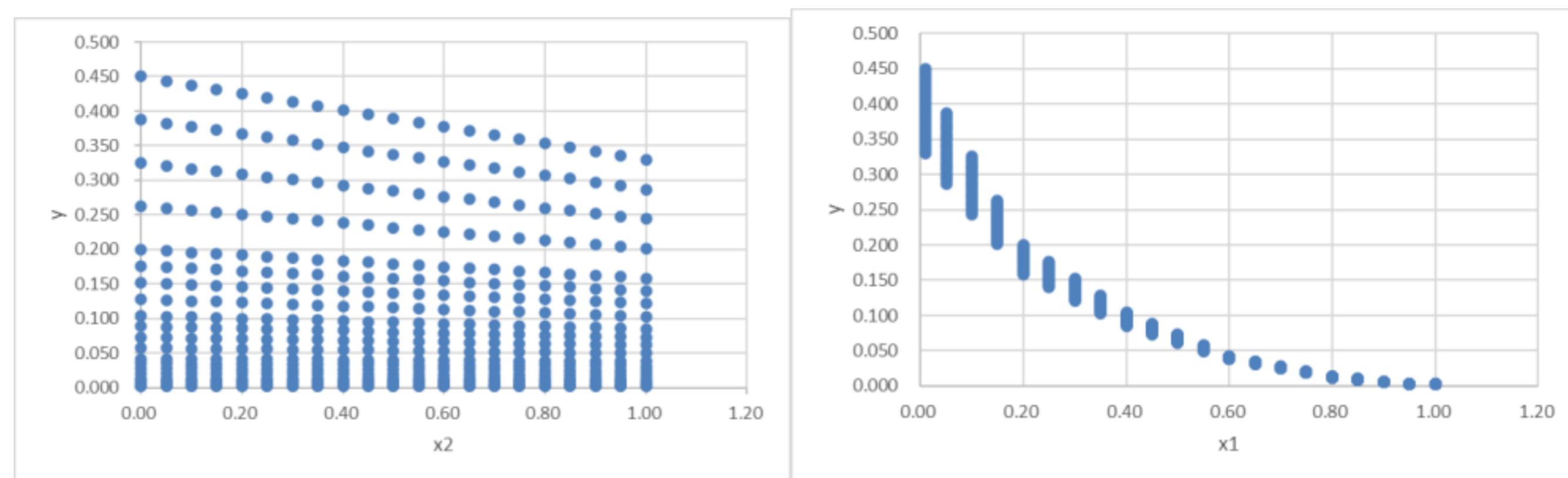


图30-31 x_1-y 和 x_2-y 的散点图

对 x_1 和 x_2 尝试合并运算，可以初步判断 x_1 和 x_2 的组合形式应该为 $\theta = a * x_1 + b * x_2$ 。

表1 变量及散点图

变量	x_1+x_2	$x_1 * x_2$	$x_1+x_2-x_1 * x_2$
散点图			
变量	$\exp(x_1) - x_2$	$x_1 + 0.1 * x_2$	$x_1 + 0.05 * x_2$
散点图			

取 $\theta = x_1 + 0.05 * x_2$ 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：

$$y = b_0 + b_1 * (m * x_1 + n * x_2) + b_2 * (m * x_1 + n * x_2)^2 + b_3 * (m * x_1 + n * x_2)^3.$$

表2 回归模型

	对数函数	指数函数	幂函数	三次多项式
R^2	0.991	0.983	0.845	0.997
显著性	0.000	0.000	0.000	0.000

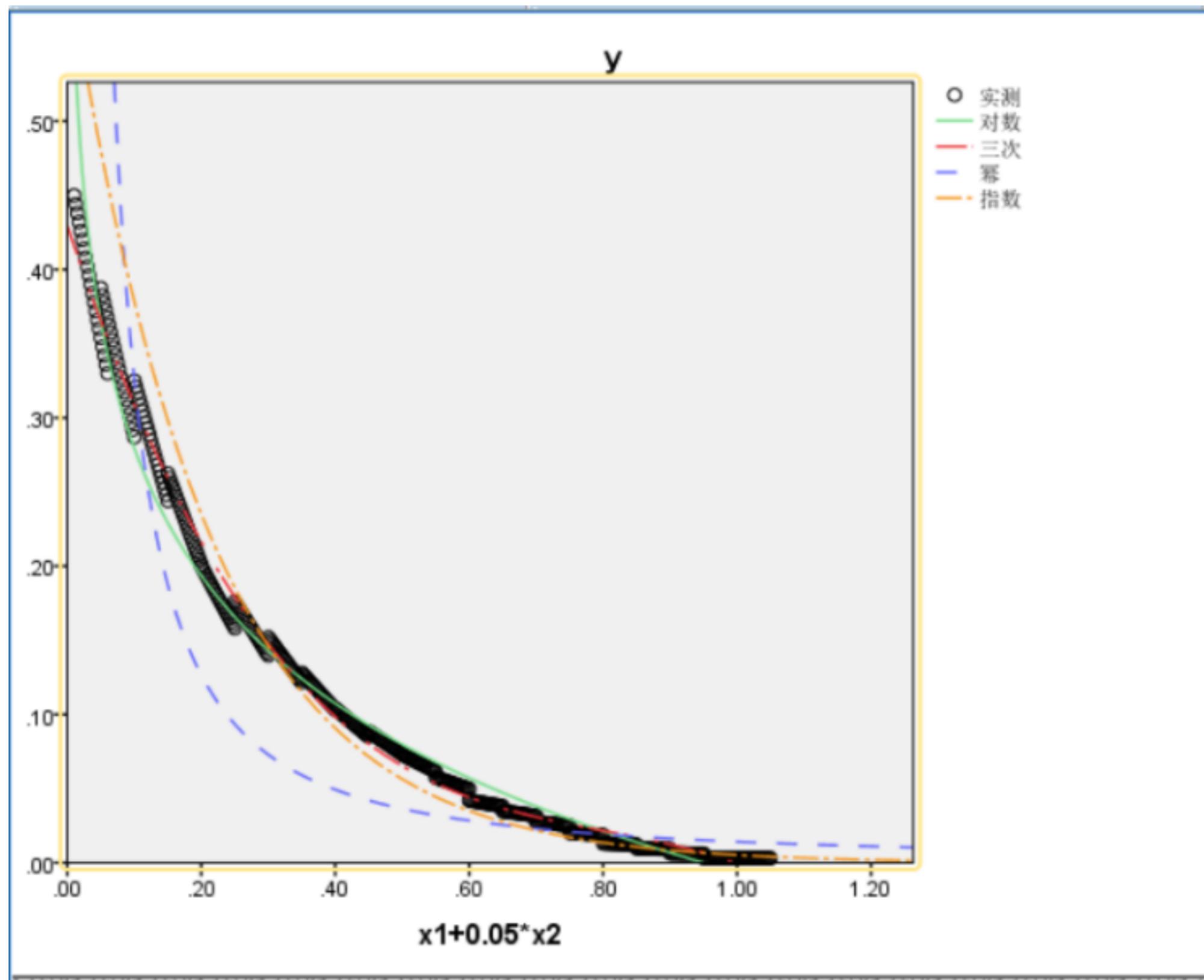


图32 拟合曲线

(2) 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。

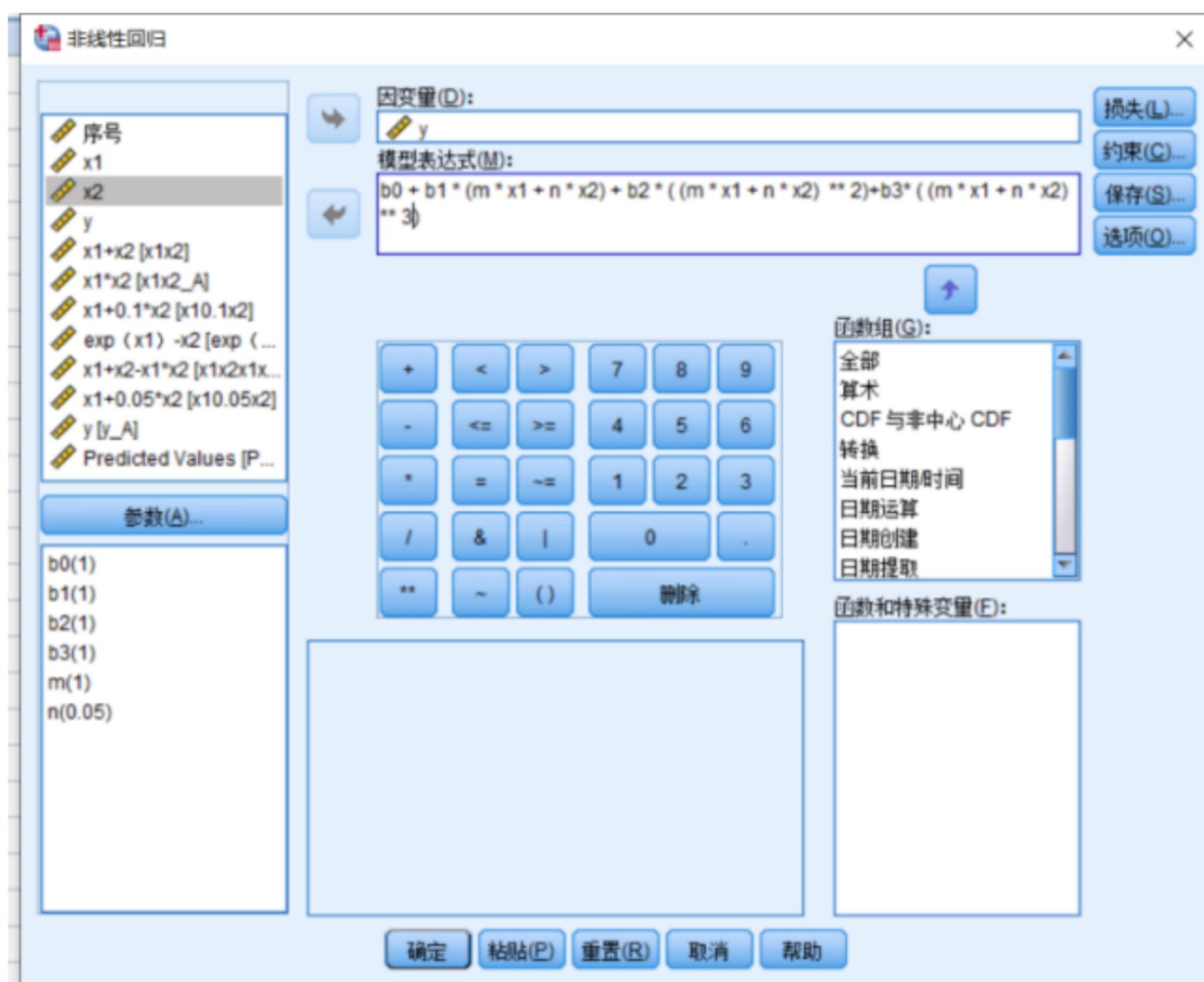


图33-34 非线性回归模型以及参数设置

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： $b_0=0.445$, $b_1=-2.658$, $b_2=5.839$, $b_3=-4.557$, $m=0.520$, $n=0.040$ ，最终回归模型为：

$$y = 0.445 - 2.658 * (0.520 * x_1 + 0.040 * x_2) + 5.839 * (0.520 * x_1 + 0.040 * x_2)^2 - 4.557 * (0.520 * x_1 + 0.040 * x_2)^3.$$

除此之外，根据ANOVA表中的 $R^2=0.996$ ，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。

迭代历史记录 ^b							
迭代编号 ^a	残差平方和	参数					
		b0	b1	b2	b3	m	n
1.0	2373.329	1.000	1.000	1.000	1.000	1.000	.050
1.1	4.304E+18	.430	92.772	189.844	281.725	-93.128	-4.656
1.2	2.117E+10	.430	8.184	20.668	27.961	-8.541	-.427
1.3	48.472	.430	-.415	3.471	2.166	.058	.003
15.0	.022	.445	-2.658	5.839	-4.557	.520	.040
15.1	.022	.445	-2.660	5.849	-4.568	.519	.040

将通过数字计算来确定导数。

a. 主迭代号在小数点左侧显示，次迭代号在小数点右侧显示。
b. 由于连续残差平方和之间的相对减小量最多为 $SSCON = 1.000E-8$ ，因此运行在 39 次模型评估和 15 次导数评估后停止。

图35 迭代历史记录

参数估算值					
参数	估算	标准误差	95% 置信区间		
			下限	上限	
b0	.445	.002	.442	.449	
b1	-2.658	60076.925	-118079.793	118074.478	
b2	5.839	264015.753	-518899.284	518910.963	
b3	-4.557	309041.757	-607405.248	607396.134	
m	.520	11749.842	-23093.000	23094.040	
n	.040	908.079	-1784.728	1784.808	

参数估算值相关性						
	b0	b1	b2	b3	m	n
b0	1.000	-.016	.016	-.016	-.016	-.016
b1	-.016	1.000	-1.000	1.000	1.000	1.000
b2	.016	-1.000	1.000	-1.000	-1.000	-1.000
b3	-.016	1.000	-1.000	1.000	1.000	1.000
m	-.016	1.000	-1.000	1.000	1.000	1.000
n	-.016	1.000	-1.000	1.000	1.000	1.000

图36 参数估计值及其相关性

ANOVA ^a			
源	平方和	自由度	均方
回归	10.901	6	1.817
残差	.022	435	.000
修正前总计	10.923	441	
修正后总计	5.715	440	

因变量: y

a. R 方 = 1 - (残差平方和) / (修正平方和) = .996。

图38 ANOVA表

(3) 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。

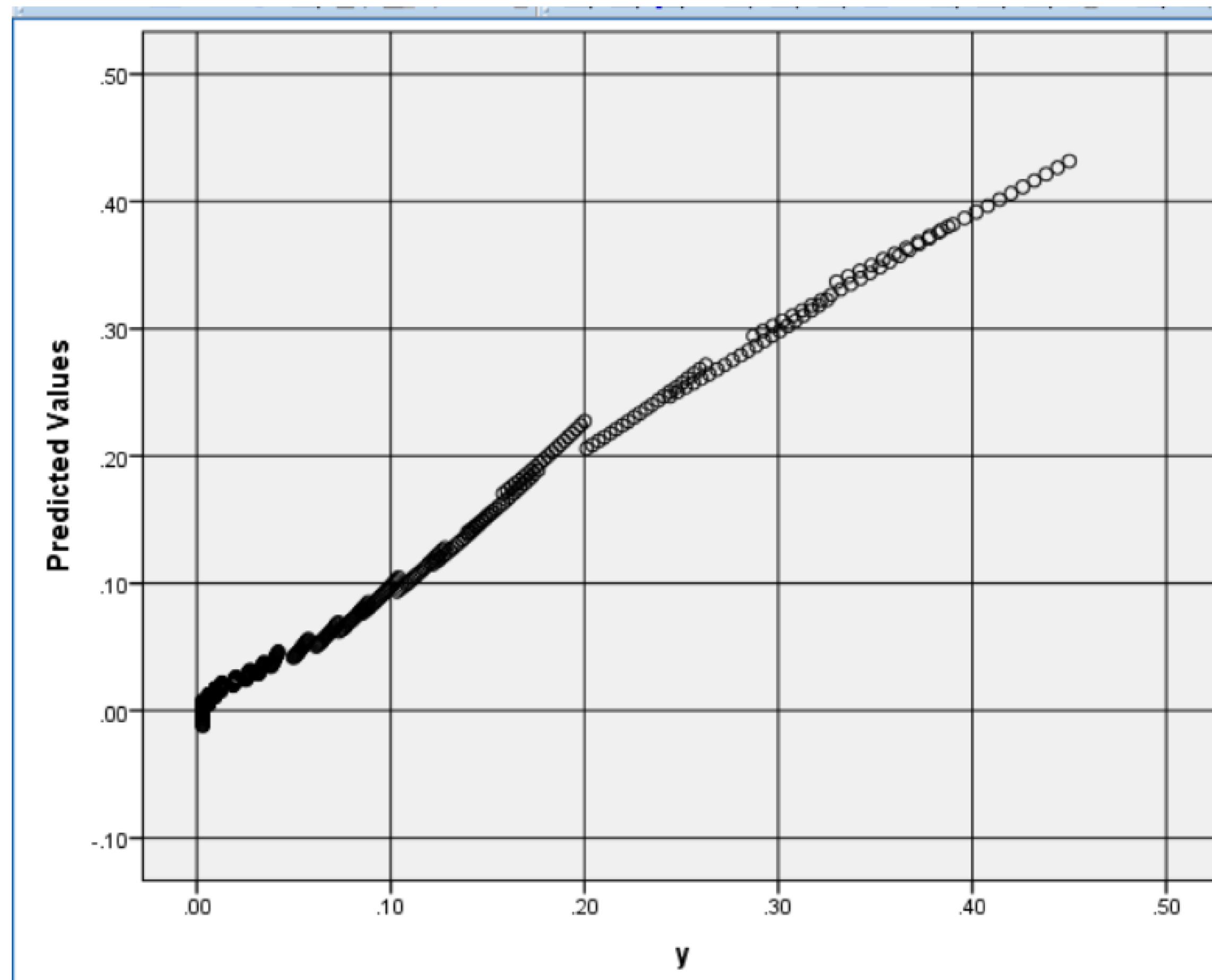


图 39 预测值与观测值 y 的散点图

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。）

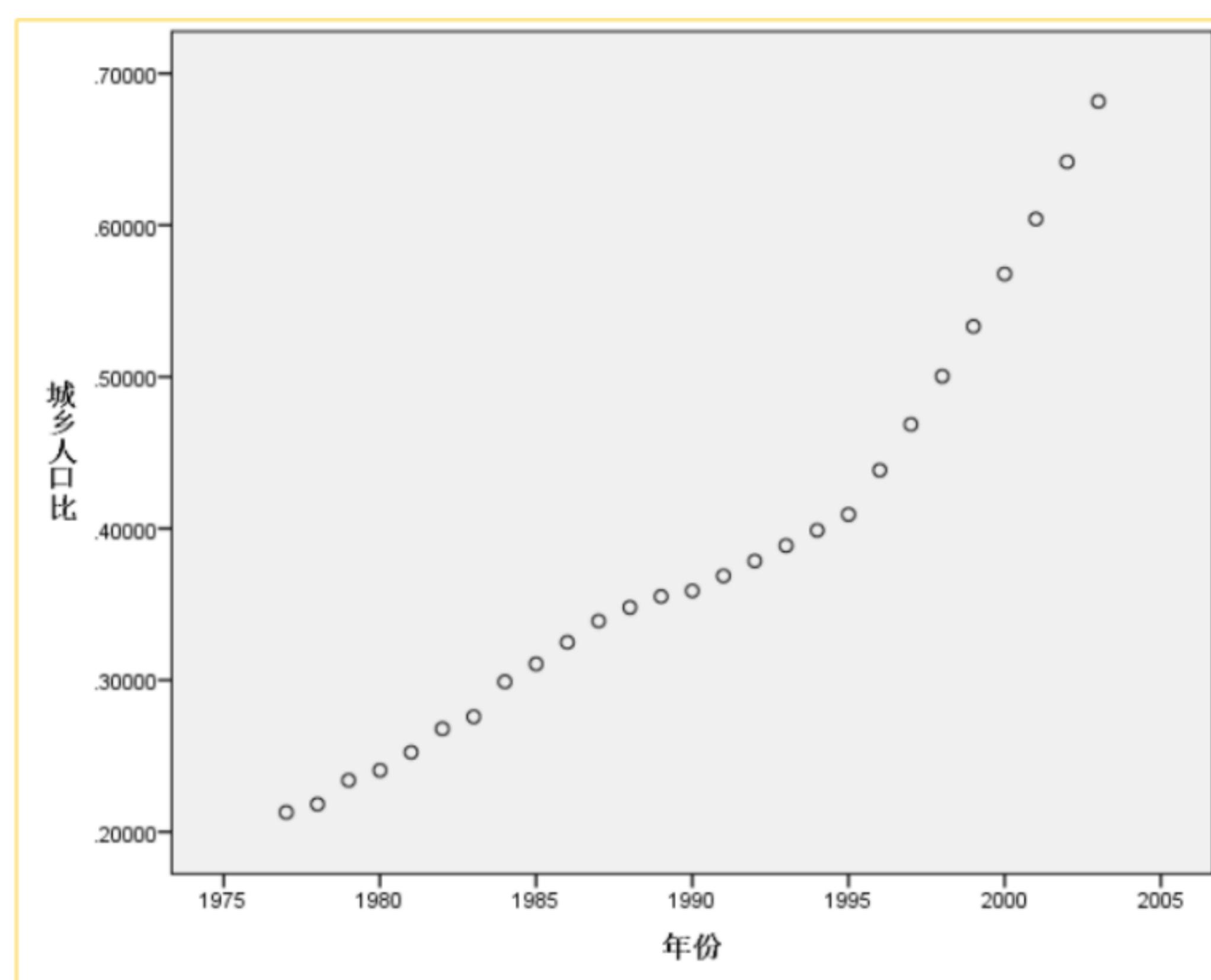


图40 散点图

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 R^2 以及显著性。可以发现 S 型方程效果最优。其表达形式为：
$$Y = e^{(b_0+b_1/X)}$$
。

表3 回归模型

	对数方程	倒数方程	S型方程
R^2	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 $Y=\ln y$, $X=1/x$, 求解 Y 和 X 的线性模型 $Y=b_0+b_1X$ 。根据返回的结果可以发现， $b_0=81.153$, $b_1=-163493.516$ ，则表达式为 $Y=81.153+ -163493.516*X$ ，对于 x 和 y 而言，最终表达式为 $y=\exp(81.153-163493.516/x)$, R^2 为 0.980。

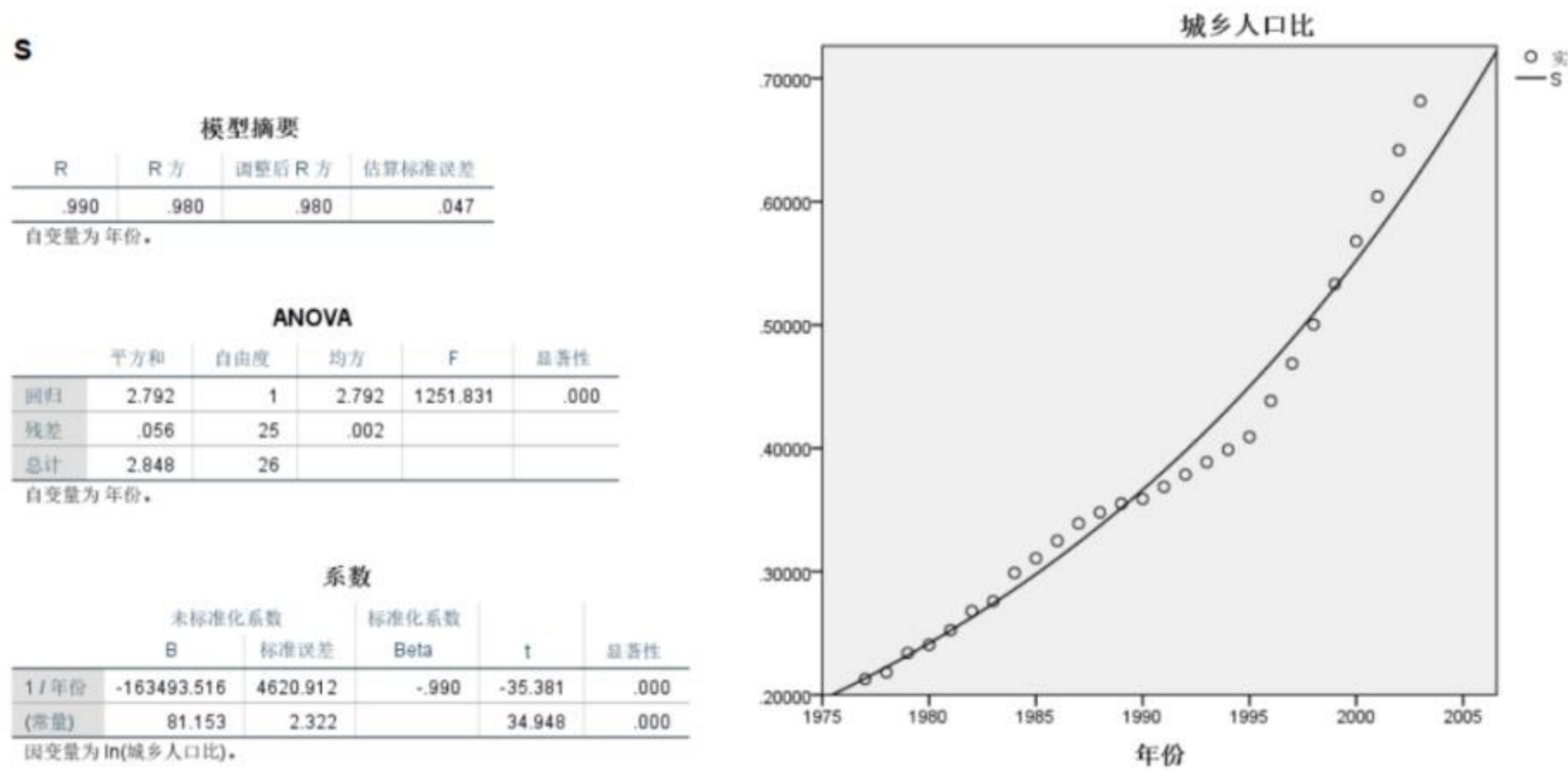


图42 S型方程模型结果

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

参数估算值

参数	估算	95% 置信区间		
		标准误差	下限	上限
b0	85.642	2.935	79.597	91.687
b1	-172436.716	5855.249	-184495.827	-160377.606

参数估算值相关性

b0	b1
1.000	-1.000
-1.000	1.000

ANOVA^a

源	平方和	自由度	均方
回归	4.456	2	2.228
残差	.011	25	.000
修正前总计	4.467	27	
修正后总计	.447	26	

因变量：城乡人口比

a. R 方 = 1 - (残差平方和) / (修正平方和) = .975.

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 R^2 来进行判断。

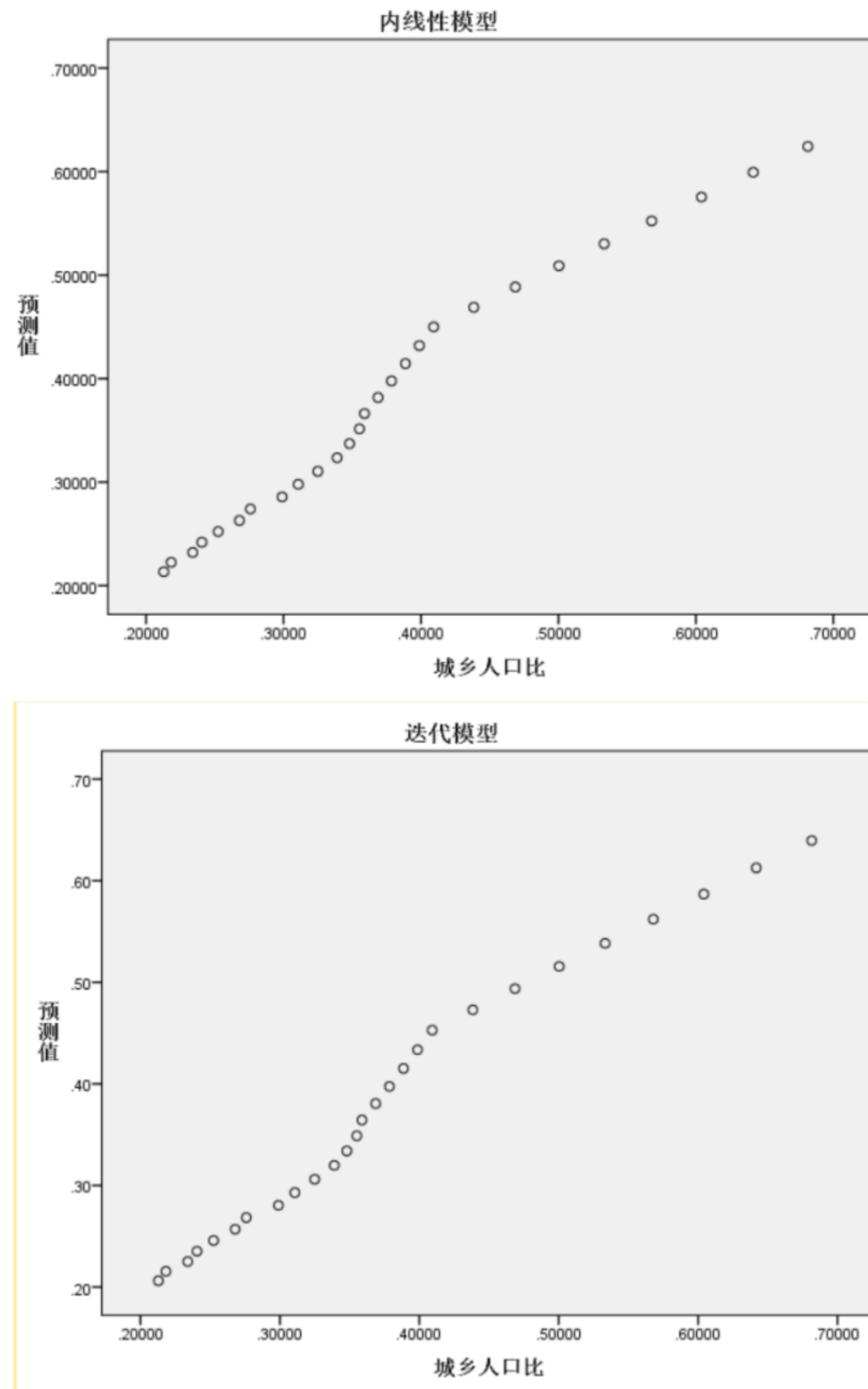


图45-46 两种方法散点图

综上所述，根据内线性方法和迭代方法的决定系数 R^2 知（ $0.980 > 0.975$ ），内线性模型的拟合效果更好。