Hunan University

Mathematical modeling, July 2020

回归分析作业

研究生第 12 组

甘毅辉 (组长)

李峥嵘 (组员)

李俊杰 (组员)

2020 年 7 月

1 问题描述

题目给出 1995-2018 年的中国关于粮食种植方面的数据，认为播种面积可能受六个因素（产量，农业劳动力人口，农民受教育程度，全国小麦进口额，城乡收入差距，家庭负担）的影响，要求用经典最小二乘回归、主成分回归及偏最小二乘回归三种方法进行分析并给出合理解释。

2 经典最小二乘回归

1995-2018 年的中国关于粮食种植方面的数据如表1所示。从表中我们可以发现不同的变量取值范围差异很大，使用 Matlab 程序对数据进行经典最小二乘回归，得到一些统计值如图1所示。从表中可以看到，p 值为 0.0000，小于 0.05，回归方程整体显著。调整的判定系数为 0.9581，大于 0.80，方程对于因变量的解释的信息量符合要求。但对于自变量的 p 值，x_2 与 x_4 的 p 值都大于 0.05 应剔除掉，这里采用的是逐步回归的方法，自动去除对因变量无显著影响的变量。

方差分析表

方差来源	自由度	平方和	均方	F值	p值
回归	6.0000	1346465.0997	224410.8500	84.9150	0.0000
残差	16.0000	42284.3257	2642.7704		
总计	22.0000	1388749.4254			

均方根误差(Root MSE)	51.4079	判定系数(R-Square)	0.9696
因变量均值(Dependent Mean)	5138.4791	调整的判定系数(Adj R-Sq)	0.9581

参数估计

变量	估计值	标准误	t值	p值
常数项	1416.1506	2147.9314	0.6593	0.5191
X1	0.3336	0.1421	2.3476	0.0321
X2	-0.2038	0.3548	-0.5745	0.5736
X3	594.8413	232.9978	2.5530	0.0213
X4	-0.0159	0.0877	-0.1815	0.8582
X5	-341.9781	117.4177	-2.9125	0.0102
X6	-0.1213	0.0395	-3.0686	0.0073

图 1: 经典最小二乘回归的方差分析表和参数估计

剔除结果如图2所示。

变量 x_2 与 x_4 都被剔除掉，最后回归方程为

$$y = 346.947 + 0.356638x_1 + 668.98x_3 - 338.844x_5 - 0.1297x_6$$

表 1: 1995-2018 年的中国关于粮食种植方面的数据

年份	y	x_1	x_2	x_3	x_4	x_5	x_6
1995	4814	1754.2	2814	7.17	1159	2.68	929
1996	4868.2	2056	2822	7.19	812	2.38	1206
1997	4927.31	2372.35	2902	7.2	198	2.36	1270
1998	4963.98	2073.53	2947	7.23	181	2.26	1240
1999	4884.6	2291.46	2961	7.28	69	2.34	1425
2000	4922.3	2236	2989	7.33	96	2.4	1551
2001	4801.6	2299.7	2973	7.39	83	2.52	1731
2002	4855.7	2248.4	3013	7.47	78	2.69	1972
2003	4804.6	2292.5	2982	7.55	67	2.81	2183
2004	4856	2480	2931	7.65	723	2.93	2428
2005	4962.7	2577.7	2873	7.74	361	3.02	2629
2006	5006.7	2822.7	2831	7.86	52	3.05	2871
2007	5213.3	2980.2	2863	7.93	10	3.01	3074
2008	5260	3051	2824	8.07	2	2.99	3418
2009	5263.3	3055	2784	8.13	106	2.97	3701
2010	5280	3082.22	2731	8.19	136	2.85	4093
2011	5323.33	3123	2670	8.27	179	2.76	4320
2012	5340	3177.35	2628	8.31	408	2.72	5032
2013	5367.3	3226	2563	8.45	551	2.64	5628
2014	5407.3	3329	2652	8.48	292	2.59	6438
2015	5426	3501	2569	8.54	385	2.36	7092
2016	5465.3	3466	2542	8.62	337	2.33	7787
2017	5475.3	3550	2516	8.7	430	2.32	8061
2018	5510.2	3660.4	2492	8.78	288	2.3	8669

3 主成分回归

对数据进行主成分分析，分别对前 3 个，前 5 个，前 6 个主成分进行拟合，调整的判定系数分别为 0.9209，0.9436，0.9581。选择所有主成分进行拟合。结果如图3所示。从中发现，x_5 的 p 值超过 0.05，剔除 x_5，结果如图4

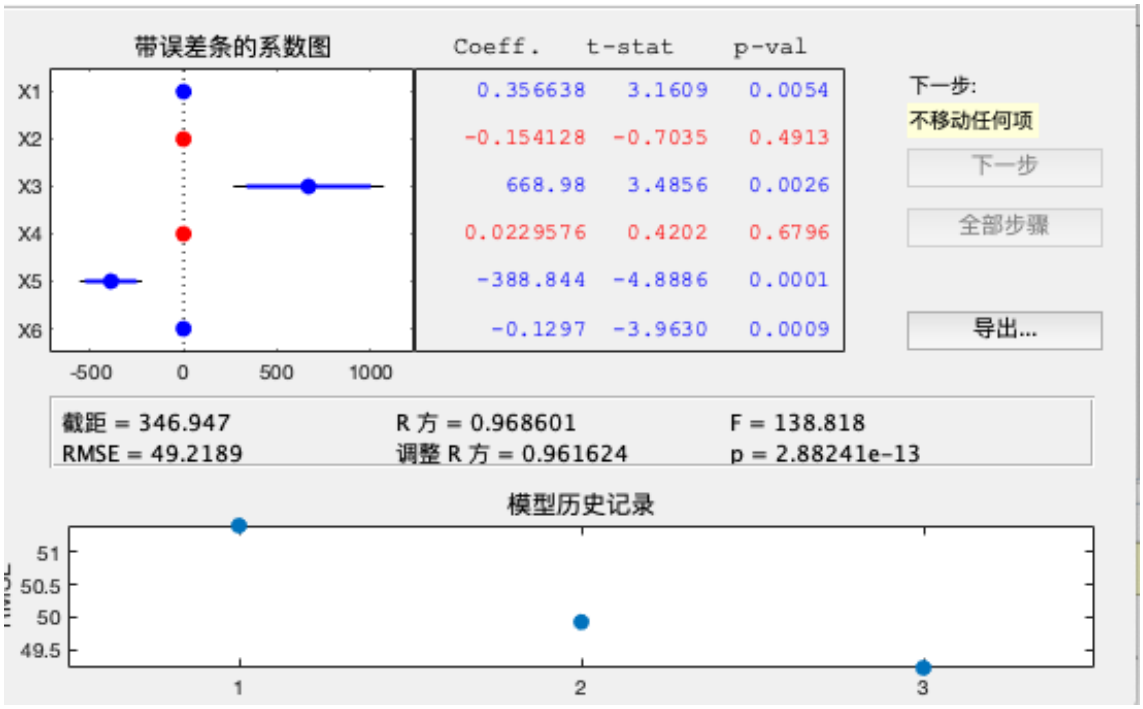

图 2: 使用逐步回归消除无关变量的影响

方差分析表

方差来源	自由度	平方和	均方	F值	p值
回归	6.0000	1346465.0997	224410.8500	84.9150	0.0000
残差	16.0000	42284.3257	2642.7704		
总计	22.0000	1388749.4254			

均方根误差(Root MSE)	51.4079	判定系数(R-Square)	0.9696
因变量均值(Dependent Mean)	5138.4791	调整的判定系数(Adj R-Sq)	0.9581

参数估计

变量	估计值	标准误	t值	p值
常数项	5138.4791	10.7193	479.3677	0.0000
X1	120.4589	5.5386	21.7489	0.0000
X2	31.6720	9.8265	3.2231	0.0053
X3	-31.1830	12.5157	-2.4915	0.0241
X4	-160.2909	47.6467	-3.3642	0.0039
X5	113.7283	88.4447	1.2859	0.2168
X6	412.6801	157.0155	2.6283	0.0183

图 3: 所有主成分回归结果

计算因变量对原始自变量的回归方程为

$$y = 2037.4 + 0.2 * x_1 - 0.5 * x_2 + 654.3 * x_3 - 0.1 * x_4 - 294.4 * x_5 - 0.1 * x_6$$

方差分析表

方差来源	自由度	平方和	均方	F值	p值
回归	5.0000	1342095.3883	268419.0777	97.8077	0.0000
残差	17.0000	46654.0371	2744.3551		
总计	22.0000	1388749.4254			

均方根误差(Root MSE)	52.3866	判定系数(R-Square)	0.9664
因变量均值(Dependent Mean)	5138.4791	调整的判定系数(Adj R-Sq)	0.9565

参数估计

变量	估计值	标准误	t值	p值
常数项	5138.4791	10.9234	470.4120	0.0000
X1	120.4589	5.6441	21.3426	0.0000
X2	31.6720	10.0136	3.1629	0.0057
X3	-31.1830	12.7539	-2.4450	0.0257
X4	-160.2909	48.5538	-3.3013	0.0042
X5	412.6801	160.0048	2.5792	0.0195

图 4: 5 个主成分回归结果

4 偏最小二乘回归

先观察所有成分对对整体的情况，如图5所示。因变量的误差不断变小，可以选取所有成分，成分选取最多 6 个。

```
MSE =

    5.7391    2.0006    0.9163    0.1317    0.0202    0.0089    0.0000
    0.9565    0.0808    0.0633    0.0585    0.0405    0.0319    0.0291
```

图 5: 整体所有成分对的情况

选取成分 6 对，求得回归方程为

$$1416.2 + 0.3 * x_1 - 0.2 * x_2 + 594.8 * x_3 - 342.0 * x_5 - 0.1 * x_6$$

附录 1

```
clc, clear
origin_xdata=xlsread('粮食数据1.xlsx');
origin_xdata=origin_xdata(2:end,2:end);
xdata(:, 1:6) = origin_xdata(:, 2:7)
```

```
xdata(:, 7) = origin_xdata(:, 1)
% x=zscore(xdata(:,2:7));%读取自变量数据矩阵
% y1=zscore(xdata(:,1));%读取因变量数据矩阵
x = xdata(:,1:6);
y1 = xdata(:, 7);
reglm(y1,x)


%逐步回归
inmodel=1:6;        %进入模型的变量为前6个
stepwise(x,y1,inmodel)    %逐步回归

%%%%主成分回归

xz = zscore(x);%数据标准化
[coeff,score,latent,tsquare,explained]=pca(xz)    %由观测数据矩阵作分析
z1=score(:,[1:6]);
reglm(y1,z1)%发现第五个主成分得分量对因变量影响不显著（p>0.05），因而删除它!
z1=score(:,[1 2 3 4 6]);
reglm(y1,z1)

%若只考虑前三个主成分，则拟合优度大大降低
z1=score(:,[1:3]);
reglm(y1,z1)

%计算因变量对原始自变量的回归方程系数
xn=zscore(x);
yn=zscore(y1);
d=xn*coeff;
st=coeff(:,[1 2 3 4 6])*(d(:,[1 2 3 4 6])\yn);
```

```
35 st2=[mean(y1)-std(y1)*mean(x)./std(x)*st,std(y1)*st'./std(x
       )],
36
37 %直接考虑前三个主成分时的回归方程式
38 st=coeff(:,[1:3])*(d(:,[1:3])\yn);
39 st3=[mean(y1)-std(y1)*mean(x)./std(x)*st,std(y1)*st'./std(x
       )],
40
41
42 %只考虑一个因变量时
43 mu=mean(xdata(:,1:7));sig=std(xdata(:,1:7)); %求均值和标准
       差
44 ab=zscore(xdata(:,1:7)); %数据标准化
45 a=ab(:,1:6);b1=ab(:,7);
46 [XL,YL,XS,YS,BETA,PCTVAR,MSE,stats]=plsregress(a,b1)%观测整
       体所有成分对的情况
47 ncomp=6; %根据整体情况, 选择成分的对数
48 [XL,YL,XS,YS,BETA,PCTVAR,MSE,stats]=plsregress(a,b1,ncomp)
49 contr=cumsum(PCTVAR,2) %求累积贡献率
50 n=size(a,2); m=size(b1,2); %n是自变量的个数,m是因变量的个数
51 BETA2(1,:)=mu(n+1:end)-mu(1:n)./sig(1:n)*BETA([2:end],:).*
       sig(n+1:end); %原始数据回归方程的常数项
52 BETA2([2:n+1],:)=(1./sig(1:n))'*sig(n+1:end).*BETA([2:end
       ],:) %计算原始自变量x1,...,xn的系数, 每一列是一个回归方
       程
```