上机作业

朱强强
17064001

1. 分析内容：经济发展、物价水平、产业结构、人口规模等因素对失业率的影响；
2. 样本：2017年中国各省、自治区、直辖市国内生产总值（GDP）、失业率、消费者价格指数（CPI）、总人口、第一产业总产值占GDP的比重、第二产业总产值占GDP的比重和第三产业总产值占GDP的比重；
3. 样本数据来源：国家统计局、国泰安统计数据库；

```
/*导入数据*/
data work.project;
input employment gdp pct1 pct2 pct3 cpi population;
cards;
1.4 28014.94    0.4298  19.014  80.5562 101.9097    2170.7
3.5 18549.19    0.9109  40.9376 58.1515 102.1116    1557
3.7 34016.32    9.2014  46.5841 44.2145 101.7292    7519.52
3.4 15528.42    4.6313  43.6547 51.714  101.1124    3702
3.6 16096.21    10.2494 39.7589 49.9916 101.7054    2529
3.8 23409.24    8.1262  39.2999 52.5739 101.3574    4369
3.5 14944.53    7.3295  46.8299 45.8406 101.5655    2717
4.2 15902.6817  18.6462 25.5341 55.8197 101.3498    3788.7
3.9 30632.99    0.3616  30.4595 69.1788 101.6779    2418
3   85869.76    4.7108  45.0157 50.2735 101.7424    8029.3
2.7 51768.26    3.7357  42.9454 53.3189 102.1188    5657
2.9 27018       9.5576  47.5175 42.9249 101.232     6255
3.9 32182.09    6.8831  47.7107 45.4062 101.177     3911
3.3 20006.31    9.1734  48.1247 42.7019 101.9878    4622
3.4 72634.1491  6.6535  45.3545 47.992  101.5179    10005.83
2.8 44552.83    9.2907  47.3719 43.3374 101.3651    9559.13
2.6 35478.09    9.9469  43.5248 46.5284 101.5459    5902
4   33902.96    8.8441  41.7235 49.4325 101.4264    6860.15
2.5 89705.23    4.0259  42.3699 53.6042 101.5077    11169
2.2 18523.26    15.5388 40.2243 44.2369 101.6134    4885
2.3 4462.54     21.5761 22.327  56.097  102.845     926
3.4 19424.73    6.5694  44.1942 49.2364 101.0086    3075.16
4   36980.22    11.526  38.7454 49.7286 101.4121    8302
3.2 13540.8256  15.0085 40.0872 44.9043 100.9102    3580
3.2 16376.34    14.279  37.8898 47.8312 100.9498    4800.5
2.7 1310.92     9.3614  39.1824 51.4562 101.6396    337
3.3 21898.81    7.9523  49.6962 42.3515 101.6374    3835
2.7 7459.8995   11.525  34.3408 54.1342 101.3898    2626
3.1 2624.83     9.0829  44.2852 46.632  101.4978    598
3.9 3443.56     7.2779  45.8993 46.8228 101.5916    682
2.6 10881.96    14.2607 39.7988 45.9405 102.1868    2445
;
run;
/*数据标准化*/
proc standard data=work.project out=work.project mean=0 std=1;
var gdp pct1 pct2 pct3 cpi population;
```

```
41 run;
```

4. 对所选变量间是否存在线性关系进行检验（alpha=10%）；

Pearson 相关系数, N = 31 Prob > |r| under H0: Rho=0							
	employment	gdp	pct1	pct2	pct3	cpi	population
employment	1.00000	-0.07393 0.6926	-0.03102 0.8684	0.30602 0.0941	-0.26869 0.1439	-0.36828 0.0415	0.03317 0.8594
gdp	-0.07393 0.6926	1.00000	-0.41183 0.0213	0.23131 0.2106	0.03841 0.8375	-0.02358 0.8998	0.84697 <.0001
pct1	-0.03102 0.8684	-0.41183 0.0213	1.00000	-0.23258 0.2080	-0.40295 0.0246	0.03817 0.8385	-0.10969 0.5569
pct2	0.30602 0.0941	0.23131 0.2106	-0.23258 0.2080	1.00000	-0.79641 <.0001	-0.32534 0.0741	0.35076 0.0530
pct3	-0.26869 0.1439	0.03841 0.8375	-0.40295 0.0246	-0.79641 <.0001	1.00000	0.28242 0.1237	-0.26187 0.1547
cpi	-0.36828 0.0415	-0.02358 0.8998	0.03817 0.8385	-0.32534 0.0741	0.28242 0.1237	1.00000	-0.23189 0.2094
population	0.03317 0.8594	0.84697 <.0001	-0.10969 0.5569	0.35076 0.0530	-0.26187 0.1547	-0.23189 0.2094	1.00000

每个格子第二行为原假设$\rho = 0$的假设检验得出的$p - vlaue$，若$p - value < \alpha$，则拒绝原假设，说明两个变量之间存在线性关系。例如，GDP与第一产业的比重，$p - value = 0.0213 < \alpha = 0.1$，说明两者存在线性关系。

5. 以调查失业率为因变量、以GDP为自变量构建一元线性回归模型，并检验模型及其系数是否通过检验；

假设y代表失业率，x代表GDP，则回归模型为$y = -2.4235 \times 10^{-16} - 0.07393x$。

方差分析					
源	自由度	平方和	均方	F 值	Pr > F
模型	1	0.06637	0.06637	0.16	0.6926
误差	29	12.07557	0.41640		
校正合计	30	12.14194			

因为$F = 0.1594 < F_\alpha = 0.16$，接受原假设$\beta_1 = 0$，$y$与$x$不存在线性关系，模型没有通过检验。

参数估计					
变量	自由度	参数估计	标准误差	t 值	Pr > |t|
Intercept	1	3.18387	0.11590	27.47	<.0001
gdp	1	-0.04703	0.11781	-0.40	0.6926

对于截距项β_0，因为$p - value < \alpha$，$\beta_0 = 0$假设被拒绝，通过检验。

对于系数项β_1，因为$p - value = 0.6926 > \alpha$，$\beta_1 = 0$假设被接受，无法通过检验。

```
1 proc reg data=work.project;
2     model employment=gdp;
3 run;
```

6. 给定置信水平为99%时构建回归系数估计值的置信区间和失业率的均值置信区间，并绘制出置信区间的图形；

参数估计							
变量	自由度	参数估计	标准误差	t 值	Pr > |t|	99% 置信限	
Intercept	1	3.18387	0.11590	27.47	<.0001	2.86441	3.50333
gdp	1	-0.04703	0.11781	-0.40	0.6926	-0.37177	0.27770

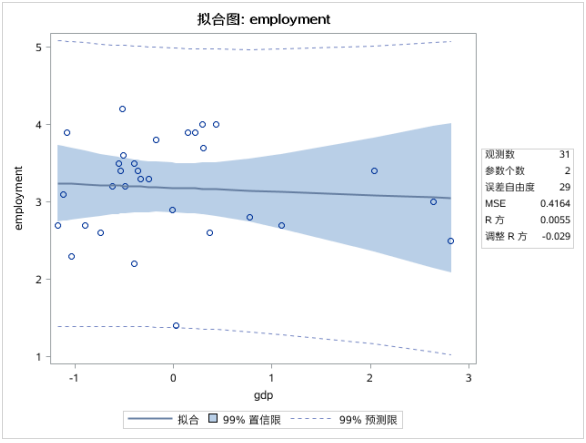
β_0的99%置信区间：[2.86441, 3.50333]

β_1的99%置信区间：[-0.37177, 0.27770]

均值置信区间如下

输出统计量						
观测	因变量	预测值	标准误差均值预测	99% 置信限均值		残差
1	1.4	3.1824	0.1160	2.8628	3.5020	-1.7824
2	3.5	3.2025	0.1249	2.8582	3.5468	0.2975
3	3.7	3.1697	0.1212	2.8356	3.5038	0.5303
4	3.4	3.2089	0.1317	2.8457	3.5720	0.1911
5	3.6	3.2077	0.1303	2.8484	3.5669	0.3923
6	3.8	3.1922	0.1177	2.8676	3.5167	0.6078
7	3.5	3.2101	0.1332	2.8428	3.5774	0.2899
8	4.2	3.2081	0.1308	2.8475	3.5687	0.9919
9	3.9	3.1769	0.1172	2.8538	3.5000	0.7231
10	3.0	3.0598	0.3318	2.1453	3.9742	-0.0598
11	2.7	3.1321	0.1740	2.6524	3.6117	-0.4321
12	2.9	3.1845	0.1159	2.8650	3.5040	-0.2845
13	3.9	3.1736	0.1187	2.8463	3.5008	0.7264
14	3.3	3.1994	0.1222	2.8624	3.5363	0.1006
15	3.4	3.0878	0.2670	2.3518	3.8239	0.3122
16	2.8	3.1474	0.1476	2.7404	3.5543	-0.3474
17	2.6	3.1666	0.1237	2.8256	3.5076	-0.5666
18	4.0	3.1699	0.1210	2.8363	3.5036	0.8301
19	2.5	3.0516	0.3509	2.0844	4.0189	-0.5516
20	2.2	3.2025	0.1250	2.8581	3.5470	-1.0025
21	2.3	3.2323	0.1678	2.7697	3.6950	-0.9323
22	3.4	3.2006	0.1233	2.8609	3.5404	0.1994
23	4.0	3.1634	0.1267	2.8141	3.5127	0.8366
24	3.2	3.2131	0.1371	2.8352	3.5909	-0.0131
25	3.2	3.2071	0.1297	2.8497	3.5645	-0.007086
26	2.7	3.2390	0.1803	2.7420	3.7361	-0.5390
27	3.3	3.1954	0.1194	2.8662	3.5246	0.1046
28	2.7	3.2260	0.1567	2.7940	3.6580	-0.5260
29	3.1	3.2362	0.1750	2.7538	3.7187	-0.1362
30	3.9	3.2345	0.1718	2.7609	3.7081	0.6655
31	2.6	3.2187	0.1451	2.8187	3.6187	-0.6187

置信区间的图形如下




```
proc reg data=work.project;
    model employment=gdp/clb cli alpha=0.01;
run;
```

7. 了解奥肯定律，并基于判定系数说明GDP对失业率变动的解释效果是否与奥肯定律一致；

   奥肯定律： 当实际GDP增长相对于潜在GDP增长上升2%时，失业率下降大约 1% 。

均方根误差	0.64529	R 方	0.0055
因变量均值	3.18387	调整 R 方	-0.0288
变异系数	20.26746		

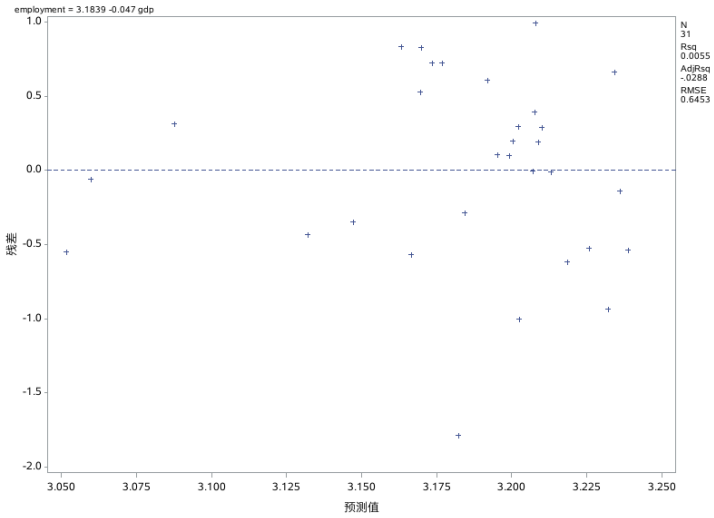
判定系数为0.0055远远小于1，说明GDP与失业率并无显著线性关系，与奥肯定律不一致。

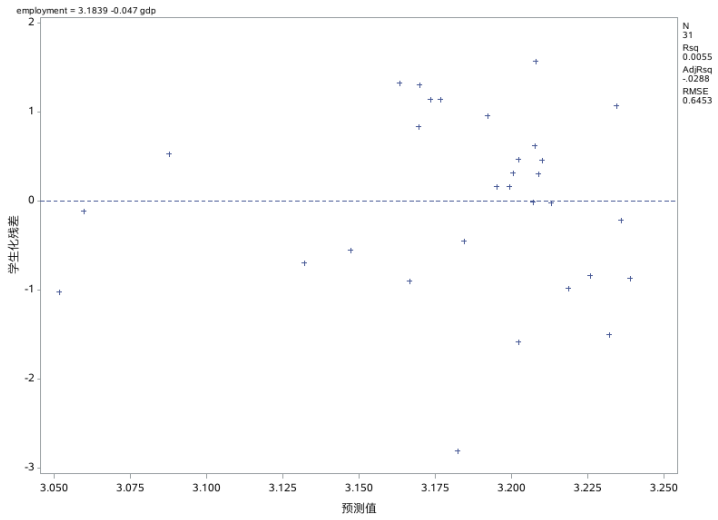
8. 计算该模型的残差、标准化残差、学生化残差、删除残差和R型学生化残差，并利用这些残差绘制残差图以判断是否存在异方差；

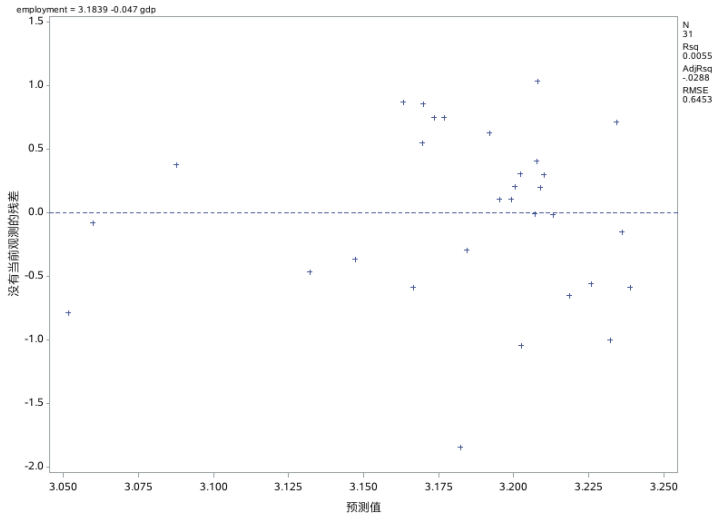
计算结果如下

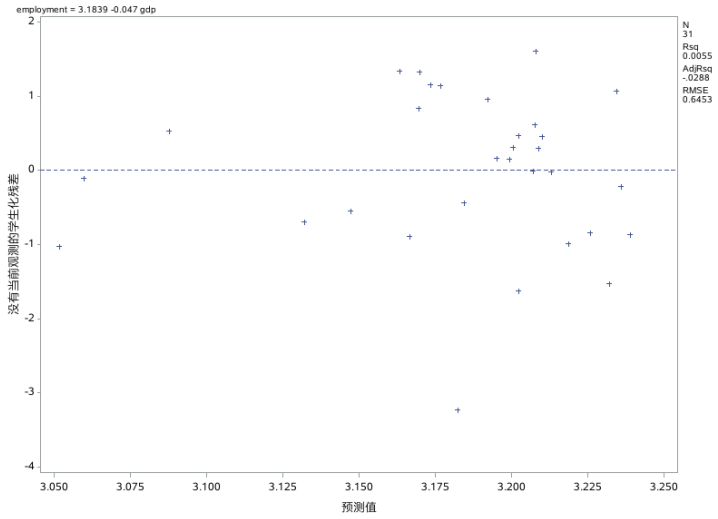
Obs	employment	gdp	pct1	pct2	pct3	cpi	population	res	stu_res	h	press_res	rstu_res	std_res
1	1.4	0.03100	-1.72419	-2.90286	3.80380	0.83474	-0.80500	-1.78241	-2.80790	0.03229	-1.84189	-3.23338	-2.76219
2	3.5	-0.39563	-1.62654	0.03748	0.97614	1.33813	-1.01907	0.29752	0.46996	0.03748	0.30910	0.46355	0.46107
3	3.7	0.30149	0.05616	0.79478	-0.78283	0.38470	1.06077	0.53031	0.83671	0.03529	0.54971	0.83227	0.82182
4	3.4	-0.53179	-0.87142	0.40189	0.16367	-1.15315	-0.27085	0.19112	0.30254	0.04168	0.19943	0.29775	0.29617
5	3.6	-0.50619	0.26887	-0.12060	-0.05371	0.32536	-0.68002	0.39232	0.62077	0.04080	0.40901	0.61407	0.60798
6	3.8	-0.17658	-0.16207	-0.18216	0.27220	-0.54230	-0.03819	0.60782	0.95802	0.03330	0.62876	0.95662	0.94194
7	3.5	-0.55810	-0.32378	0.82775	-0.57760	-0.02344	-0.61444	0.28988	0.45912	0.04264	0.30279	0.45278	0.44922
8	4.2	-0.51492	1.97315	-2.02840	0.68185	-0.56124	-0.24061	0.99191	1.56975	0.04110	1.03442	1.61247	1.53715
9	3.9	0.14900	-1.73803	-1.36782	2.36788	0.25680	-0.71873	0.72314	1.13960	0.03300	0.74781	1.14573	1.12064
10	3.0	2.63861	-0.85529	0.58443	-0.01813	0.41762	1.23859	-0.05976	-0.10798	0.26433	-0.08124	-0.10613	-0.09262
11	2.7	1.10160	-1.05320	0.30676	0.36622	1.35609	0.41109	-0.43206	-0.69531	0.07271	-0.46594	-0.68898	-0.66956
12	2.9	-0.01393	0.12846	0.91996	-0.94559	-0.85495	0.61968	-0.28453	-0.44822	0.03226	-0.29401	-0.44196	-0.44093
13	3.9	0.21882	-0.41438	0.94588	-0.63243	-0.99208	-0.19795	0.72642	1.14528	0.03385	0.75188	1.15171	1.12573
14	3.3	-0.32996	0.05048	1.00140	-0.97373	1.02947	0.05006	0.10061	0.15879	0.03589	0.10435	0.15610	0.15591
15	3.4	2.04206	-0.46098	0.62987	-0.30608	-0.14212	1.92804	0.31218	0.53142	0.17126	0.37669	0.52474	0.48378
16	2.8	0.77639	0.07428	0.90044	-0.89353	-0.52310	1.77222	-0.34735	-0.55296	0.05235	-0.36654	-0.54623	-0.53829
17	2.6	0.36738	0.20747	0.38447	-0.49080	-0.07231	0.49655	-0.56659	-0.89464	0.03676	-0.58821	-0.89147	-0.87804
18	4.0	0.29638	-0.01636	0.14289	-0.12427	-0.37026	0.83077	0.83007	1.30960	0.03519	0.86034	1.32665	1.28635
19	2.5	2.81148	-0.99430	0.22958	0.40223	-0.16756	2.33378	-0.55163	-1.01866	0.29574	-0.78328	-1.01935	-0.85486
20	2.2	-0.39680	1.34245	-0.05818	-0.78000	0.09598	0.14180	-1.00253	-1.58360	0.03751	-1.04160	-1.62804	-1.55362
21	2.3	-1.03054	2.56782	-2.45853	0.71684	3.16670	-1.23917	-0.93234	-1.49635	0.06766	-1.00000	-1.53060	-1.44484
22	3.4	-0.35617	-0.47805	0.47425	-0.14902	-1.41195	-0.48951	0.19938	0.31477	0.03649	0.20693	0.30982	0.30897
23	4.0	0.43508	0.52798	-0.25653	-0.08690	-0.40591	1.33371	0.83659	1.32221	0.03857	0.87015	1.34024	1.29646
24	3.2	-0.62137	1.23481	-0.07657	-0.69577	-1.65729	-0.31341	-0.01310	-0.02077	0.04513	-0.01372	-0.02041	-0.02030
25	3.2	-0.49357	1.08675	-0.37128	-0.32637	-1.55856	0.11233	-0.00709	-0.01121	0.04038	-0.00738	-0.01101	-0.01098
26	2.7	-1.17259	0.08863	-0.19792	0.13113	0.16131	-1.44463	-0.53902	-0.86998	0.07809	-0.58468	-0.86623	-0.83532
27	3.3	-0.24466	-0.19737	1.21217	-1.01796	0.15582	-0.22446	0.10462	0.16498	0.03425	0.10833	0.16219	0.16213
28	2.7	-0.89545	0.52777	-0.84727	0.46912	-0.46151	-0.64618	-0.52599	-0.84028	0.05899	-0.55896	-0.83590	-0.81512
29	3.1	-1.11337	0.03211	0.48646	-0.47772	-0.19224	-1.35358	-0.13624	-0.21935	0.07358	-0.14706	-0.21571	-0.21113
30	3.9	-1.07647	-0.33425	0.70293	-0.45364	0.04163	-1.32428	0.66550	1.06993	0.07088	0.71627	1.07271	1.03132
31	2.6	-0.74121	1.08303	-0.11525	-0.56499	1.52563	-0.70932	-0.61873	-0.98405	0.05057	-0.65169	-0.98350	-0.95885

残差图









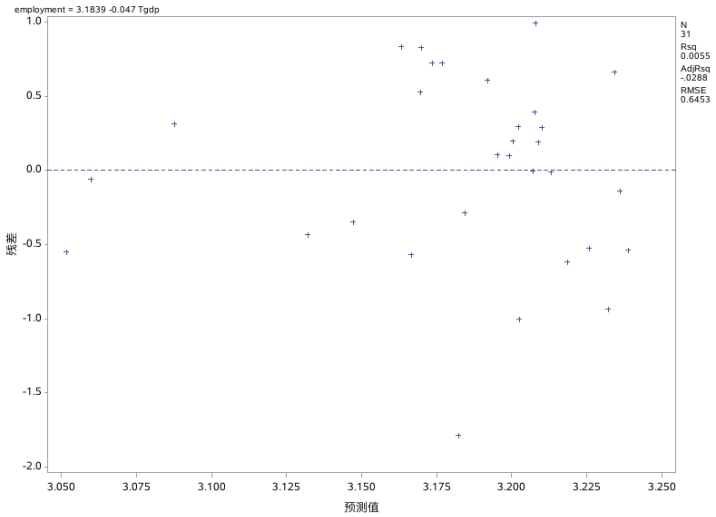
由残差图可观察残差呈放射状，残差与自变量相关，存在异方差。

```
proc reg data=work.project;
    model employment=gdp/r partial;
    output out=res_data residual=res student=stu_res press=press_res
rstudent=rstu_res h=h;
run;
data res_data;
    set res_data;
    std_res=stu_res*sqrt(1-h);
```

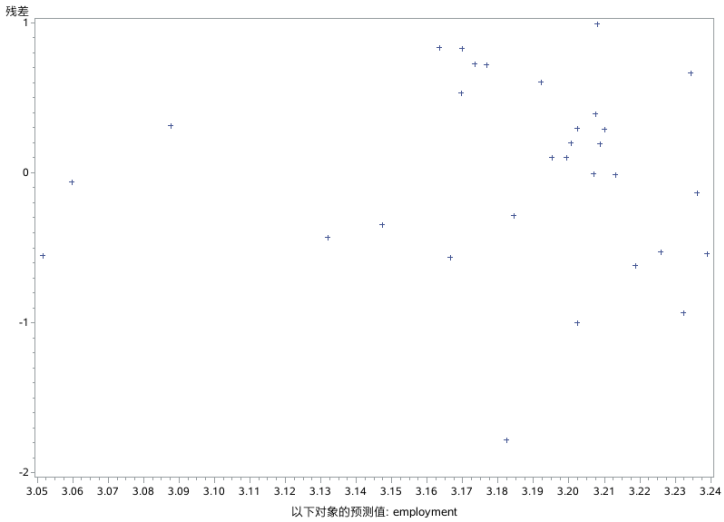
```
run;
proc print data=res_data;
run;
proc reg data=work.project;
    model employment=gdp/r partial;
    plot r.*p.;
    plot student.*p.;
    plot press.*p.;
    plot rstudent.*p.;
run;
```

9. 分别采用box-cox法和加权最小二乘法重新估计模型，并对新估计的模型残差再次进行残差分析；

box-cox法估计的模型的残差图



加权最小二乘法估计的模型的残差图




```
proc transreg data=work.project detail ss2;
    model boxcox(employment)=identity(gdp);
    output out=work.box_cox;
run;
proc reg data=work.box_cox;
    model employment=tgdp/r p;
    plot r.*p.;
run;
```

```
data work.project_w;
    set work.project;
    array row{9} w1-w9; /*w1-w9为不同m时的权数值*/
    array p{9} (-2,-1.5,-1,-0.5,0,0.5,1,1.5,2);
    do i=1 to 9;
        row(i)=1/gdp**p{i};
end;
run;
proc print data=work.project_w;
run;
proc reg data=work.project_w;
    model employment=gdp/r p;
    weight wwork.project;
    output out=work.wls residual=r predicted=p;
run;
proc gplot data=work.wls;
    plot r*p;
run;
```

10. 对比第（4）步和第（7）步估计的模型残差，以判断box-cox法和加权最小二乘法是否更有效；

    由残差图我们可以看到，box-cox法和加权最小二乘法得出的残差，大多在[-1, 1]区间内，所以更加有效。

11. 构建包含经济发展、物价水平、产业结构、人口规模等因素的多元线性回归模型，并对模型进行拟合、检验；

参数估计

变量	自由度	参数估计	标准误差	t 值	Pr > |t|
Intercept	1	3.18387	0.11247	28.31	<.0001
gdp	B	-0.10339	0.30811	-0.34	0.7400
pct1	B	-0.01911	0.15880	-0.12	0.9052
pct2	B	0.15023	0.13382	1.12	0.2723
pct3	0	0	.	.	.
cpi	B	-0.18456	0.13165	-1.40	0.1732
population	B	0.01107	0.29886	0.04	0.9707

    由上图我们可以构建多元线性模型：

    $employment = 3.18387 - 0.10339 \times gdp - 0.01911 \times pct1 + 0.15023 \times pct2 - 0.18456 \times cpi + 0.01107 \times population$

    。

方差分析

源	自由度	平方和	均方	F 值	Pr > F
模型	5	2.33891	0.46778	1.19	0.3410
误差	25	9.80302	0.39212		
校正合计	30	12.14194			

    因为$p - value = 0.3410 > \alpha = 0.05$，所以模型不成立，无法通过检验。

```
proc reg data=work.project;
    model employment=gdp pct1 pct2 pct3 cpi population;
    output out=res_multi r=res_multi student=stu_res press=press_res
rstudent=rstu_res h=h;
run;
```

12. 判断该样本中是否有单位为杠杆点或强影响点，若存在的话是分析删除该单位对模型系数估计的影响和对模型拟合效果的影响；

Obs	employment	gdp	pct1	pct2	pct3	cpi	population	res_multi	stu_res	h	press_res	rstu_res
1	1.4	0.03100	-1.72419	-2.90286	3.80380	0.83474	-0.80500	-1.21454	-2.85861	0.53964	-2.63826	-3.41381
2	4.2	-0.51492	1.97315	-2.02840	0.68185	-0.56124	-0.24061	1.20441	2.33193	0.31971	1.77043	2.58294

有两个强影响点，分别是$employment = 1.4$和$employment = 4.2$。

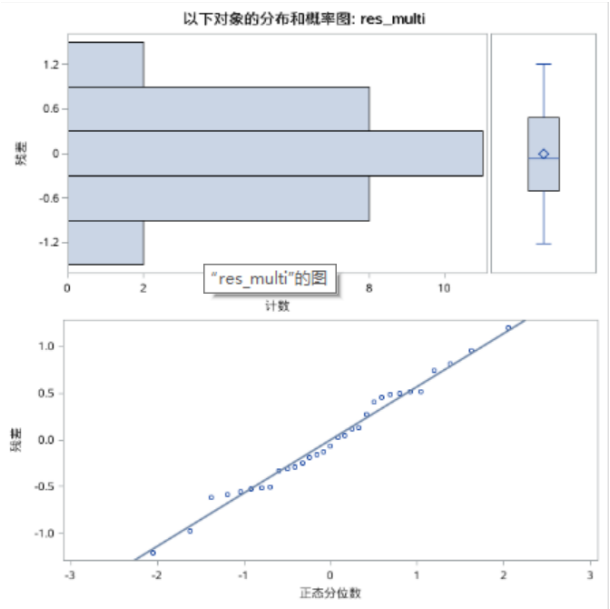
均方根误差	0.62620	R 方	0.1926
因变量均值	3.18387	调整 R 方	0.0312
变异系数	19.66774		

均方根误差	0.47612	R 方	0.3406
因变量均值	3.21034	调整 R 方	0.1972
变异系数	14.83091		

去除两个强影响点之后模型判定系数$R^2 = 0.3406 > 0.1926$，说明去除强影响点之后的模型拟合优度更好。

```
data work.res_multi_new(where=(stu_res>2 or stu_res<-2));
     set res_multi;
run;
proc print data=work.res_multi_new;
run;
*删除异常值;
data work.res_multi_new;
    set res_multi;
    if employment=1.4 then delete;
    if employment=4.2 then delete;
run;
proc print data=work.res_multi_new;
run;
*重新进行拟合;
proc reg data=work.res_multi_new;
    model employment=gdp pct1 pct2 pct3 cpi population;
run;
```

13. 检验所构建的多元线性模型的残差是否服从正态分布；



由该图我们知道残差服从正态分布。

```
proc univariate data=res_multi plot;
    var res_multi;
run;
```