What influences the regional development in Taiwan?

Motivation

I was born in Nantou County, a beautiful place with comfortable environment. I am proud of my hometown; however, sometimes I feel worried about her poverty and her being left-behind.

Due to the anxiety, I would like to find out the factors that influence the regional development. Moreover, a regression model is given in order to check if those factors are able to make a satisfactory explanation.

Analysis Process

Stage1: Identifying structure through data summarization and data reduction

The data is included of eight indices, which respectively refer to:

X1: The ratio of people aged 15 and over receiving advanced education (%) (college and over)
X2: The ratio of literate people aged 15 and over (%)
X3: Average expenditure on public safety ($ per capita)
X4: The clearance rate of criminal cases (%)
X5: Road density (km / km^2)
X6: Number of telephone subscribers per thousand population
X7: Density of suspended particulate (micrograms / m^3)
X8: The ratio of well-disposed refuse (%)

The original data:

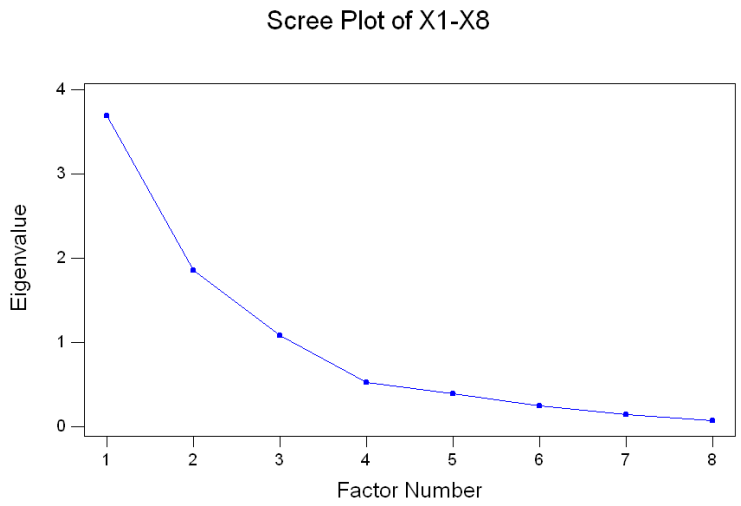
	X1	X2	X3	X4	X5	X6	X7	X8
Taipei County	25.82	97.22	2113.9	39.86	1.16	651.81	77.94	98.25
Yilan County	16.56	94.71	3459.1	70.66	0.71	450.79	66.25	74.37
Taoyuan County	22.87	96.44	2205.1	55.89	1.74	571.72	96	85.45
Hsinchu County	23.39	96.97	2684.4	68.51	0.75	463.39	63.26	90.24
Miaoli County	16.48	96.98	2804.1	72.53	0.87	437.21	75.6	96.31
Taichung County	20.5	95.68	2027.5	77.39	1.26	459.86	97.46	97.48
Chunghwa County	17.3	92.53	2408.6	72.31	1.93	434.72	94.45	99.63
Nantou County	17.9	95.73	3404.8	68.47	0.52	450.78	69.83	94.71
Yunlin County	15.32	91.84	2626.7	60.83	1.71	399.86	116.3	90.64
Chiayi County	13.87	92.64	3075.6	71.92	1.08	295.31	93.44	93.43
Tainan County	19.08	94.93	2625.5	57.79	1.5	299.1	88.08	95.84
Kaohsiung County	18.48	94.28	2444.7	55.06	0.99	433.53	138.85	99.8
Pingtong County	17.55	94.36	2907.6	56.85	0.82	388.72	94.01	96.83
Taitung County	9.69	96.01	5752.2	76.58	0.37	404.01	57.62	83.95
Hualien County	17.69	97.6	4762.4	59.77	0.3	457.24	61.15	88.97
Penghu County	21.62	96.83	12570	87.2	2.01	443.16	66.9	98.7
Keelung City	22.63	95.87	3853.3	73.01	2.38	549.01	114.8	100
Hsinchu City	31.59	96.72	3052.3	63.86	3.74	733.62	67.17	100
Taichung City	38.45	98.14	2775.9	49.85	8.52	845.08	87.73	100
Chiayi City	37.07	96.1	4001.4	45.02	8.31	783.28	113.39	100
Tainan City	32.84	96.77	3073.7	47.94	6.65	873.02	101.89	99.99
Taipei City	47.61	97.99	5005.2	73.57	4.64	890.71	83.4	100
Kaohsiung City	31.93	97.11	3899.7	52.63	10.1	638.13	141.92	100

The correlation matrix:

	X1	X2	X3	X4	X5	X6	X7	X8
X1	1							
X2	0.598	1						
X3	0.026	0.251	1					
X4	-0.373	-0.160	0.491	1				
X5	0.782	0.376	0.019	-0.470	1			
X6	0.921	0.624	-0.025	-0.442	0.757	1		
X7	0.192	-0.309	-0.292	-0.400	0.473	0.128	1	
X8	0.515	0.174	0.024	-0.216	0.481	0.412	0.438	1

From the matrix, we find that several variables are highly interrelated (correlation coefficient $\geqq$ 0.6).

The pairs are (X1 , X2), (X1 , X5), (X1 , X6), (X2 , X6), and (X5 , X6)
There may be the problem of multicollinearity, so it is reasonable to employ the factor analysis to do the data reduction in this case.



	Eigenvalue	Proportion	Cumulative
1	3.68880829	0.4611	0.4611
2	1.85002757	0.2313	0.6924
3	1.08401443	0.1355	0.8279
4	0.52758997	0.0659	0.8938
5	0.39370874	0.0492	0.9430
6	0.24219216	0.0303	0.9733
7	0.14727359	0.0184	0.9917
8	0.06638525	0.0083	1.0000

Since $\lambda_1 = 3.68880829$, $\lambda_2 = 1.85002757$, $\lambda_3 = 1.08401443$, by the latent root criterion (eigenvalue > 1) we extract three common factors. And from the cumulative percent of variance, we find that the explanatory power to the variance reaches to 0.828, which is satisfactory to us.

The unrotated factor matrix is:

Variable	Factor1	Factor2	Factor3	Communality
X1	-0.924	0.218	0.014	0.901
X2	-0.562	0.656	0.286	0.828
X3	0.087	0.725	-0.541	0.825
X4	0.591	0.490	-0.446	0.788
X5	-0.888	-0.049	-0.150	0.813
X6	-0.909	0.222	0.165	0.903
X7	-0.400	-0.732	-0.414	0.867

X8	-0.623	-0.135	-0.538	0.697

After examining the unrotated loadings, we find that there are merely two factors retained, which differs from our initial extraction. Owing to the contradiction, we need a factor matrix rotation. The VARIMAX rotation is applied here.

The VARIMAX rotated factor matrix is:

Variable	Factor1	Factor2	Factor3	Communality
X1	0.887	0.332	0.066	0.901
X2	0.859	-0.256	-0.158	0.828
X3	0.149	-0.010	-0.896	0.825
X4	-0.366	-0.209	-0.782	0.788
X5	0.688	0.564	0.144	0.813
X6	0.912	0.214	0.158	0.903
X7	-0.117	0.854	0.353	0.867
X8	0.332	0.758	-0.108	0.697

Variance Explained by Each Factor			
	Factor 1	Factor 2	Factor 3
Before rotation	3.6888083	1.8500276	1.0840144
After rotation	3.1102998	1.8883292	1.6242213

The above two tables show that the problem of inconsistency is solved after rotation. Next we are going to name these factors and interpret them.

Stage2: Interpreting the factors

	Factor1	Factor2	Factor3
Variables	The ratio of people aged 15 and over receiving advanced education The ratio of literate people aged 15 and over Road density Number of subscribers per thousand population	Density of suspended particulate The ratio of well-disposed refuse	Average expenditure on public safety The clearance rate of criminal cases

Relations among Variables	The first two variables refer to the education level, while the rest two refer to the level of infrastructure.	That the county is able to dispose all kinds of waste better than others means that it reaches to a relatively high degree of industrialization. And a higher density of suspended particulate reflects a higher degree of industrialization.	These two factors are both about public safety.
Factor Name	Government Expenditure	Degree of Industrialization	Public Safety

(Table 1)

Regression Model

To see if these extracted factors really have influence on regional development, we are going to fit a regression model. Intuitively, people living in the better-developed places will have higher income level. From the viewpoint, we choose the average income per capita as the regressand and the factor scores as the regressors to fit a regression model.

The regression model is

$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3$, where

Y = the average income per capita

Z_1 = score of factor1, Z_2 = score of factor2, Z_3 = score of factor3

The software output suggests that there is one outlier and one influential case, which lead a contradiction to our basic assumptions.

```
Unusual Observations
Obs        C2        C1        Fit     SE Fit     Residual      St Resid
 16      0.15    203920     221262      22646     -17343      -1.16 X
 22      1.86    357214     277220      13518      79994       3.40 R
R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.
```

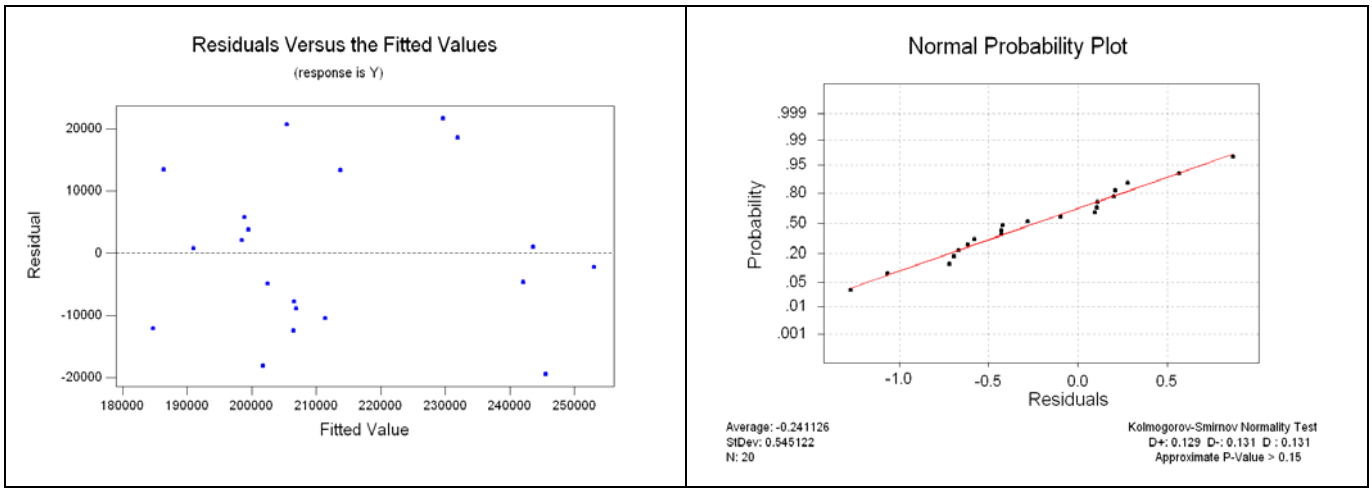
Observation 16 is Penghu County and observation 22 is Taipei City.

Penghu is an isle out of Taiwan, and Taipei is the capital holding the most abundant resource. It is of no doubt that these two cases will bring some distortion to our model. Excluding them and repeating the process, we find there is still an outlier, Hsinchu City.

```
Unusual Observations
Obs        C2        C1        Fit     SE Fit    Residual    St Resid
 17      0.98    278438     234093       6846       44345       2.73R
R denotes an observation with a large standardized residual
```

Industrial Park attracts lots of companies, factories and people to move into Hsinchu, and therefore it also gathers much resource, which again results in a distortion to our model.

With the three places left unconsidered, all the assumptions are satisfied.



By stepwise regression, we have

Stepwise Regression: Y versus Z1, Z2, Z3

```
 Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15
Response is Y on 3 predictors, with N = 20
   Step         1       2
Constant     215903   211295
Z1           19581    16371
T-Value       5.12     4.81
P-Value      0.000    0.000

Z3                   -17113
T-Value                -2.89
P-Value                0.010
R-Sq(adj)     57.08    69.50
C-p            8.3      2.3
```

And the output is:

```
The regression equation is
Y = 211295 + 16371 Z1 - 17113 Z3
Predictor       Coef     SE Coef         T       P
Constant      211295        3412     61.92   0.000
Z1             16371        3407      4.81   0.000
Z3            -17113        5928     -2.89   0.010
R-Sq = 72.7%     R-Sq(adj) = 69.5%
```

The R-square is 0.695, which is an acceptable explanatory power; unfortunately, factor2 is expelled. The problem is probably resulted from the selection of the regressand. If we could have a more adequate response variable, we may have a result that matches our expectation.

Conclusion

As Table 1 shows, three common factors can be extracted from the original eight indices. The factors that influence regional development are Government Expenditure, Degree of Industrialization, and Public Safety. This matches the real situation. And the result also tells us that the government plays the most important role in regional development. If the government budget is able to be allocated more fare and unbiasedly, the difference of development among regions will be reduced.