Explore the Factors Influencing the Total Industrial Output Value

p930018036 Li Xinyi

p930018045 Liu Ainuo

Bejing Normal University-Hong Kong Baptist University United

International College

Content

Introduction

Gross Industrial Output Value (GIOV) is an important indicator that reflects the total scale and level of industrial production in a country or region in a certain period. It is included in the statistical yearbooks of all provinces. The purpose of our group project is to explore the influencing factors of GIOV in each province.

In the process of searching for previous research literatures, we found that many models take the local economic situation into account, but few documents have studied the impact of local population structure and education level on the total industrial output value. Through researching, we will finally select a suitable model to fit the industrial output value, and give relevant policy recommendations about the influencing factors.

Conceptual Framework

From a macro perspective, the demographic structure and the situation of industry employees will affect the overall industrial output value. Internal R&D expenditures are related to technological progress and degree of automation. The average employee salary and per capita disposable income can reflect labour costs and regional economic levels. The education level of the region is closely related to the quality of the labour force. We want to conduct our research by establishing a multiple linear regression model. Additionally, the regression model is tested for heteroscedasticity, and t, f and other statistics.

Econometric Models and Estimation Method

Our research direction is mainly the influencing factors of the province's total industrial output value, so our explained variable is the total industrial output value of each province in 2019. When selecting explanatory variables, we mainly explore the influencing factors of total industrial output value from five aspects, including demographic structure, industrial employment situation, people's living standards, scientific expenditures in industries above designated size, and education level. At the same time, we have added two indicators to explore the interaction between education level and per capita wages, and the multiple effects of per capita wages on the total industrial output value.

(1) Set the variables for the model

Dependent variable Y: the gross industry output value (thousand million CNY)

Independent variables X_1~X_{10}

X_1 : Natural population growth rate (‰)	X_6 : Disposable income per capita (CNY)
X_2 : Male to female ratio	X_7 : Internal expenditure of R&D expenses (thousand million CNY)
X_3 : Percentage of population over 65	X_8 : Proportion of college degree or above
X_4 : Number of employees	X_9 : the interaction term of X_5 and X_8
X_5 : Per capita salary (CNY)	X_{10} : the quadratic term of X_5

The main model type is below

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + u$$

(*u is the error term)

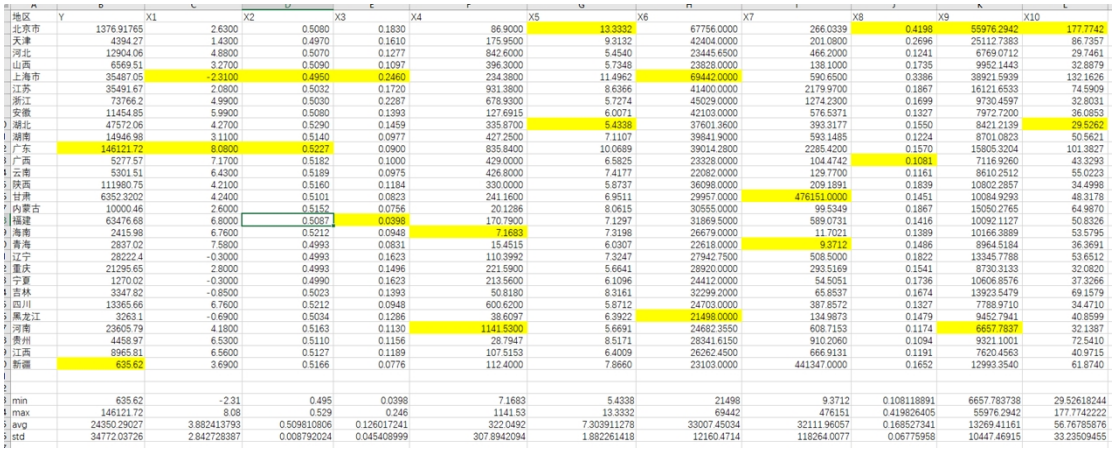
(2) Assumption:

1. The population data is obtained based on random sampling and error adjustment, which can better reflect the overall situation.
2. Zero-conditional mean is fitted.
3. There is no perfect collinearity between independent variables.
4. Linear in parameters.

Data

Our data sources all refer to the 2019 Statistical Yearbooks issued by the National Bureau of Statistics and the Local Bureau of Statistics, and the data from the Seventh Census. Among all the data, the data of Tibet is not complete, so we exclude it from our research. We use the permanent population as total population of each province. The number of employees and per capita wages refer to industrial personnel, and per capita income refers to all residents of the province. Research expenditures use data from industrial enterprises above designated size. An industrial enterprise above designated size refers to an industrial enterprise whose main business income is more than 20 million yuan per year.

Here is a picture of our data.



地区	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
北京市	1376.91765	2.6300	0.5080	0.1830	86.9000	13.3332	67756.0000	266.0339	0.4198	55976.2942	177.7742
天津	4394.27	1.4300	0.4970	0.1610	175.9500	9.3132	42404.0000	201.0800	0.2696	25112.7383	86.7357
河北	12904.06	4.8800	0.5070	0.1277	842.6000	5.4540	23445.6500	466.2000	0.1241	6769.0712	29.7461
山西	6569.51	3.2700	0.5090	0.1097	396.3000	5.7348	23828.0000	138.1000	0.1735	9952.1443	32.8879
上海市	35487.05	-2.3100	0.4950	0.2460	234.3800	11.4962	69442.0000	590.6500	0.3386	38921.5939	132.1626
江苏	35491.67	2.0800	0.5032	0.1720	931.3800	8.6366	41400.0000	2179.9700	0.1867	16121.6533	74.5909
浙江	73766.2	4.9900	0.5030	0.2287	678.9300	5.7274	45029.0000	1274.2300	0.1699	9730.4597	32.8031
安徽	11454.85	5.9900	0.5080	0.1393	127.6915	6.0071	42103.0000	576.5371	0.1327	7972.7200	36.0853
湖北	47572.06	4.2700	0.5290	0.1459	335.8700	5.4338	37601.3600	393.3177	0.1550	8421.2139	29.5262
湖南	14946.98	3.1100	0.5140	0.0977	427.2500	7.1107	39841.9000	593.1485	0.1224	8701.0823	50.5621
广东	146121.72	8.0800	0.5227	0.0900	835.8400	10.0689	39014.2800	2285.4200	0.1570	15805.3204	101.3827
广西	5277.57	7.1700	0.5182	0.1000	429.0000	6.5825	23328.0000	104.4742	0.1081	7116.9260	43.3293
云南	5301.51	6.4300	0.5189	0.0975	426.8000	7.4177	22082.0000	129.7700	0.1161	8610.2512	55.0223
陕西	111980.75	4.2100	0.5160	0.1184	330.0000	5.8737	36098.0000	209.1891	0.1839	10802.2857	34.4998
甘肃	6352.3202	4.2400	0.5101	0.0823	241.1600	6.9511	29957.0000	476151.0000	0.1451	10084.9293	48.3178
内蒙古	10000.46	2.6000	0.5152	0.0756	20.1286	8.0615	30555.0000	99.5349	0.1867	15050.2765	64.9870
福建	63476.68	6.8000	0.5087	0.0398	170.7900	7.1297	31869.5000	589.0731	0.1416	10092.1127	50.8326
海南	2415.98	6.7600	0.5212	0.0948	7.1683	7.3198	26679.0000	11.7021	0.1389	10166.3889	53.5795
青海	2837.02	7.5800	0.4993	0.0831	15.4515	6.0307	22618.0000	9.3712	0.1486	8964.5184	36.3691
辽宁	28222.4	-0.3000	0.4993	0.1623	110.3992	7.3247	27942.7500	508.5000	0.1822	13345.7788	53.6512
重庆	21295.65	2.8000	0.4993	0.1496	221.5900	5.6641	28920.0000	293.5169	0.1541	8730.3133	32.0820
宁夏	1270.02	-0.3000	0.4990	0.1623	213.5600	6.1096	24412.0000	54.5051	0.1736	10606.8576	37.3266
吉林	3347.82	-0.8500	0.5023	0.1393	50.8180	8.3161	32299.2000	65.8537	0.1674	13923.5479	69.1579
四川	13365.66	6.7600	0.5212	0.0948	600.6200	5.8712	24703.0000	387.8572	0.1327	7788.9710	34.4710
黑龙江	3263.1	-0.6900	0.5034	0.1286	38.6097	6.3922	21498.0000	134.9873	0.1479	9452.7941	40.8599
河南	23605.79	4.1800	0.5163	0.1130	1141.5300	5.6691	24682.3550	608.7153	0.1174	6657.7837	32.1387
贵州	4458.97	6.5300	0.5110	0.1156	28.7947	8.5171	28341.6150	910.2060	0.1094	9321.1001	72.5410
江西	8965.81	6.5600	0.5127	0.1189	107.5153	6.4009	26262.4500	666.9131	0.1191	7620.4563	40.9715
新疆	635.62	3.6900	0.5166	0.0776	112.4000	7.8660	23103.0000	441347.0000	0.1652	12993.3540	61.8740
min	635.62	-2.31	0.495	0.0398	7.1683	5.4338	21498	9.3712	0.108118891	6657.783738	29.52618244
max	146121.72	8.08	0.529	0.246	1141.53	13.3332	69442	476151	0.419826405	55976.2942	177.7742222
avg	24350.29027	3.882413793	0.509810806	0.126017241	322.0492	7.303911278	33007.45034	32111.96057	0.168527341	13269.41161	56.76785876
std	34772.03726	2.842728387	0.008792024	0.045408999	307.8942094	1.882261418	12160.4714	118264.0077	0.06775958	10447.46915	33.23509455

For each variable we summary the city that has the maximum and the minimum

	GIOV	Population gross	Male-Female	⩾65 years old	Practitioner number
Min	635.62 (Xin Jiang)	-2.31 (Shanghai)	0.495 (Shanghai)	0.0398 (Fu Jian)	7.1683 (Hai Nan)
Max	146121.72 (Guang Dong)	8.08 (Guang Dong)	0.529 (Guang Dong)	0.246 (Shanghai)	1141.53 (He Nan)
Avg	24350.29027	3.882413793	0.509810806	0.126017241	322.0492
Std	34772.03726	2.842728387	0.008792024	0.045408999	307.8942094

value and we calculate the average value. The table below is our result.

	Salary	income	R&D expense	education	educ*slaray	salary^2
Min	5.4338 (Hu Bei)	21498 (Hei Longjiang)	9.3712 (Qin Hai)	0.1081 (Guang Xi)	6657.78 (He Nan)	29.526 (Hu Bei)
Max	13.3332 (Bei Jing)	69442 (Shanghai)	476151 (Gan Su)	0.4198 (Bei Jing)	55976.29 (Bei Jing)	177.77 (Bei Jing)
Avg	7.303911278	33007.45034	32111.96057	0.168527341	13269.41	56.76785876
Std	1.882261418	12160.4714	118264.0077	0.06775958	10447.47	33.23509455

In the data we collect, we can find out that the Guangdong Province has the highest GIOV and the Xinjiang Province has the lowest GIOV. For the population structure, we can find that Shanghai has the lowest natural population gross rate, the lowest male-female ratio and the highest the percentage of population over 65 years old. In the contract, the Guangdong Province has the highest natural population gross rate

and the highest the male-female ratio. As for the expenditure on the R&D, the Shanghai has the highest input, and the Hei Longjiang has the lowest input. Beijing has the highest level of education and the highest salaries. Many information can be found out.

Result

For the data we collect, we use the R studio, we get the multivariate regression of our model

```
Residuals:
   Min     1Q Median     3Q    Max
-36637 -12339  -1445  11520  42007

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.031e+04  4.006e+05  -0.076 0.940532
X1           1.306e+03  2.406e+03   0.543 0.593891
X2           2.210e+05  6.799e+05   0.325 0.748951
X3          -2.580e+05  1.799e+05  -1.434 0.168602
X4           2.472e+01  1.597e+01   1.548 0.139016
X5          -9.186e+04  3.607e+04  -2.547 0.020230 *
X6           2.023e+00  7.718e-01   2.621 0.017308 *
X7          -2.449e-02  3.798e-02  -0.645 0.527255
X8           2.583e+06  5.321e+05   4.855 0.000127 ***
X9          -3.106e+01  6.368e+00  -4.878 0.000121 ***
X10          9.570e+03  2.988e+03   3.203 0.004933 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21980 on 18 degrees of freedom
Multiple R-squared:  0.7431,    Adjusted R-squared:  0.6004
F-statistic: 5.207 on 10 and 18 DF,  p-value: 0.001223
```

figure 1 the regression result of the main model

From the results above, we can find out:

i. The increase of population gross rate and the increase of male-female ration both contribute to the increase in the GIOV. The increase of population percentage over 65 has a negative effect on the GIOV. (X_1, X_2, X_3)

ii. More practitioners are there in the industry, more GIOV there is. The higher the salaries of the practitioners have, the higher the cost of labours is, the lower the GIOV. (X_4, X_5)

iii. The higher the disposable income per capita is, the more active the economy in the region is, thereby increasing GIOV. (X_6)

iv. More spending on R&D makes GIOV lower. Because the R&D expenditures are difficult to have return in the short term, which generally reduce the GIOV. (X_7)

v. The higher education level of the region, the higher the quality of the labor force, which has positive impact on GIOV. (X_8)

vi. The interaction term reflects, with the level of education improves, if the salaries increase the same unit, the GIOV decreases less than a lower level of education. The education increasing will cover the expense of the salaries. (X_9)

vii. At first, the increase of salaries decreases the GIOV, but after some interval, the increase of salaries will increase the GIOV, and maybe the high salaries attract more excellent people to come and encourage people to work hard. (X_{10})

1) Testing for heteroscedasticity
Using White Test to check

$$\hat{u}^2 = \delta_0 + \delta_1 Y + \delta_2 Y^2$$

For the regression above, we have the null hypothesis H_0: $\delta_1 = \delta_2 = 0$

By running the model, the F statistics is 0.9347 (2 and 26 DF), which has the p-value of 0.4055. We can conclude that there is no heteroscedasticity in the model at 5% confident interval.

For further analysis, we also calculate the heteroscedasticity-robust standard errors.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.031e+04  4.006e+05  -0.076 0.940532
X1           1.306e+03  2.406e+03   0.543 0.593891
X2           2.210e+05  6.799e+05   0.325 0.748951
X3          -2.580e+05  1.799e+05  -1.434 0.168602
X4           2.472e+01  1.597e+01   1.548 0.139016
X5          -9.186e+04  3.607e+04  -2.547 0.020230 *
X6           2.023e+00  7.718e-01   2.621 0.017308 *
X7          -2.449e-02  3.798e-02  -0.645 0.527255
X8           2.583e+06  5.321e+05   4.855 0.000127 ***
X9          -3.106e+01  6.368e+00  -4.878 0.000121 ***
X10          9.570e+03  2.988e+03   3.203 0.004933 **
---
```

figure 2 the regression result of the main model

```
t test of coefficients:

              Estimate  Std. Error t value Pr(>|t|)
(Intercept)  4.6566e+03  1.3795e+05  0.0338  0.97344
X1           1.2466e+03  1.7529e+03  0.7112  0.48608
X2           1.3707e+01  8.2432e+00  1.6628  0.11367
X3          -2.1260e+03  2.2998e+03 -0.9244  0.36749
X4           1.8992e+01  1.0796e+01  1.7592  0.09554 .
X5          -4.6530e+04  5.3402e+04 -0.8713  0.39505
X6           2.0532e+00  8.8847e-01  2.3109  0.03289 *
X7          -2.1955e-02  2.1326e-02 -1.0295  0.31688
X8           1.6684e+04  1.1000e+04  1.5168  0.14670
X9          -1.8588e+01  1.2672e+01 -1.4669  0.15966
X10          4.7678e+03  4.9215e+03  0.9688  0.34550
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

figure 3 the robust standard error

The heteroscedasticity-robust standard errors are somewhat larger, in all cases, than the usual OLS standard errors. We can find out that the standard errors of X_2, X_3, X_8, X_9 have large difference with their robust-standard errors.

We then use B-P test to test the residuals on these variables,

$$\hat{u}^2 = \theta_0 + \theta_1 X_2 + \theta_2 X_3 + \theta_3 X_8 + \theta_4 X_9$$

The null hypothesis of this test is H_0: $\theta_1 = \theta_2 = \theta_3 = \theta_4 = 0$

The F statistics is 2.087 (4 and 24 DF) and the p-value of it is 0.114. We can finally conclude that there is no heteroscedasticity in the model at 10% confident level. Because there is no heteroscedasticity in the model, the t test and f test are valid.

2) The T/F test, R square, Adjusted R square

From the output of our main model, we can find out the variables X_1, X_2, X_3, X_4, X_7 (p-value of each is: 0.59, 0.75, 0.17, 0.14, 0.53) are statistically insignificant at 10% confident level.

Next, we use F test to examine the joint significance of these variables. We have the null hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_7 = 0$

The F statistic of the test is 1.471, which has the p-value of 0.2479. The null hypothesis should not be rejected at 10% confident level.

The main model: R square = 0.7431, adjusted R square = 0.6004. If we exclude these variables: R square = 0.6382, adjusted R square = 0.5595. The adjusted R square is less than the original model if we exclude these variables.

The gross population rate, male-female ratio, percentage over 65, the partitioners' number in the industry have little impact on the GIOV. But from our intuition, these variables have some impact on the GIOV. For the future study, we should use more data to run the model, and get a more precise conclusion.

3) Multicollinearity Problem (VIF)

```
      X1          X2          X3          X4          X5
2.711141    2.071136    3.855327    1.401049 267.126405

      X6          X7          X8          X9         X10
 5.105125    1.169248   75.326410 256.510769 571.595635
```

The VIF of X5, X8, X9, X10 are larger than 10, which means they may have multicollinearity with other variables.

4) Residuals Analysis

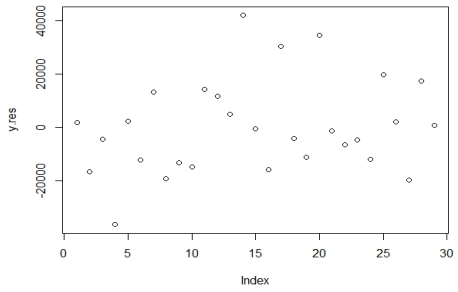


figure 4 the scatter plot of the residuals

From the figure 4, we can find the residuals obey the assumptions basically.

a) Linearity assumption

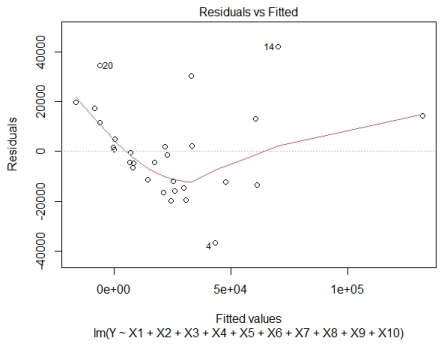
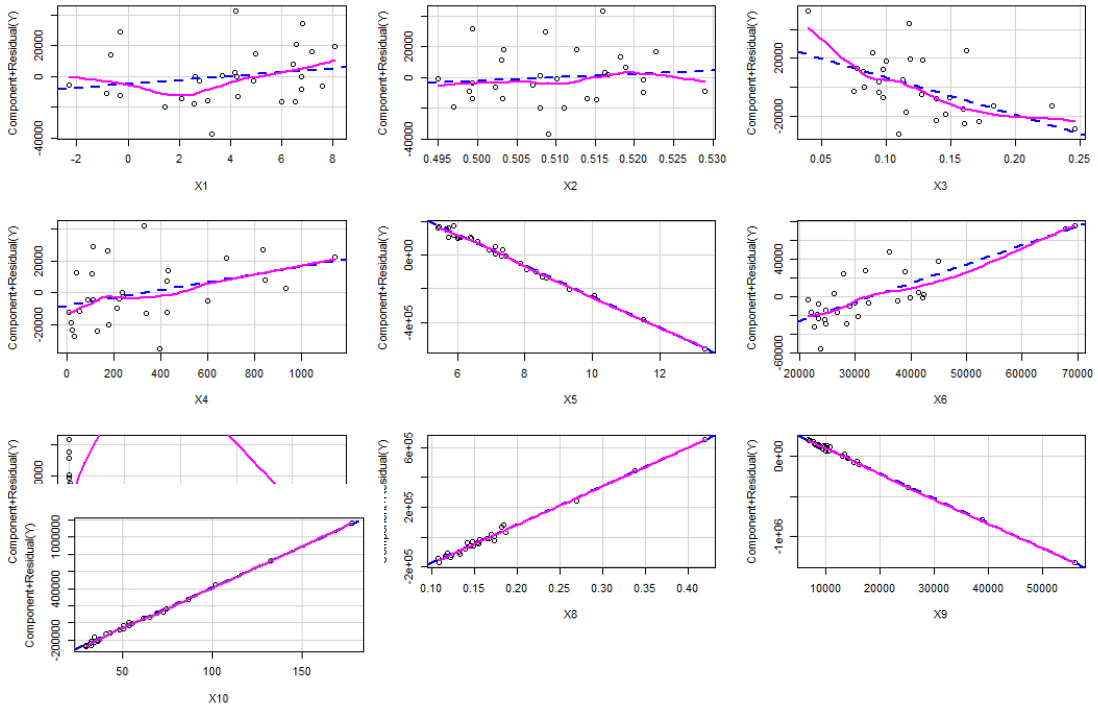


figure 5 the scatter plot of the residual vs fitted y value

If the linear hypothesis is satisfied, there should be no trend relationship between the two, that is, the red line should basically coincide with y = 0. Therefore, the linear hypothesis of the original model may not be valid.



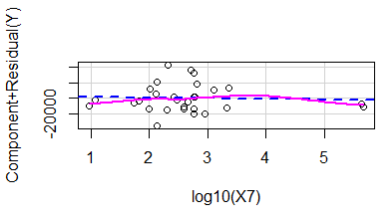


figure 6

We plot the graph of residuals with each independent variable. After the analysis of component residual diagram, we find the X_7 has the problem so we try the quadratic form and logarithm form. Finally, logarithm form is better from the figure 6.

b) Normal distribution assumption

From the figure 7, we know that this graph doesn't deviate from normal generally (normal is a straight line).

After calculation with Galton's skewness, the skewness is 0.08679917, quantifying the direction of the distribution, which is nearly full symmetry. From the shapiro.test, W and p- value are large relatively, so it is normal.

```
        Shapiro-Wilk normality test

data:  residuals(ols)
W = 0.96317, p-value = 0.3923
```

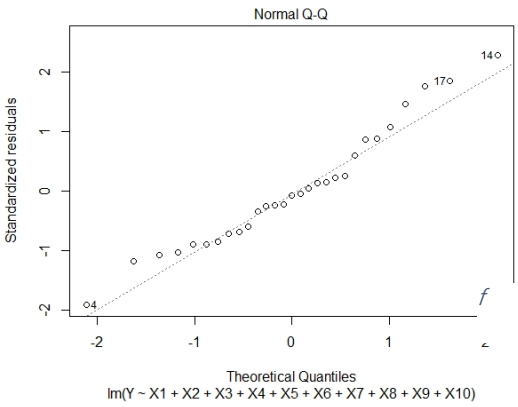


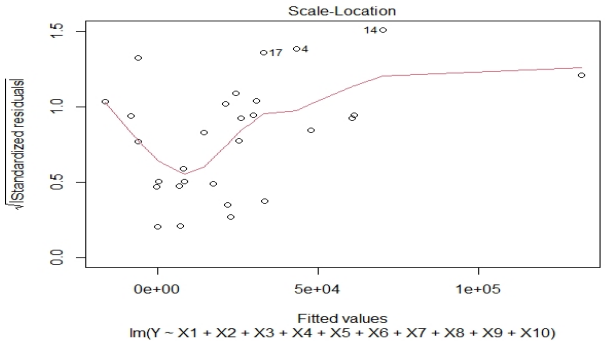
figure 7

c) Independence assumption

```
 lag Autocorrelation D-W Statistic p-value
   1      -0.1310243      2.261654   0.564
 Alternative hypothesis: rho != 0
```

1< D-W statistic <3, they are independent.

d) Constant variance for all levels of X



The trend line is almost horizontal, which satisfies the assumption.

Conclusion

We use OLS to study the established model, and then use the robust standard errors, White test, B-P test, and find that the model does not have heteroscedasticity at 10% confidence. Next, we carry out research on the explanatory variables T, F test, and find that some of these variables are not statistically significant. However, we recognize the economic significance of these variables.

During the multicollinearity test, we find that several variables have serious multicollinearity. This may be due to the improper model design and the limitation of economic data, resulting in a general correlation among the explanatory variables. Finally, we conduct residual analysis from the four assumptions and find that the linear hypothesis requires us to change the X_7 from level form to logarithm form. Because of the small sample size and the limited knowledge of economy, it is hard to avoid the problem of multicollinearity, and the model needs to be optimized.

Through our research, we have found that the increase in the natural population growth rate and the increase in the ratio of males to females have a positive impact on the value of industrial output. However, the increase in the proportion of the population over 65 has a negative impact on GIOV. The increase in the number of employees and the increase in per capita disposable income can increase the value of industrial output to a certain extent. At the same time, we find that improving the education level of the region can effectively offset the negative impact of wage growth on industrial output value. An effective increase in salary levels can help

increase the industrial output value of the region. We hope that our research will have a reference value on future policy.

Reference

1. 鲁晓东、连玉君，“中国工业企业全要素生产率估计：１９９９—２００７”，*China Economic Quarterly*,2012,11(2),541-558.
2. 上海市统计局，“如何解读工业总产值？”，http://tjj.sh.gov.cn/tjwd/20140722/0014-224877.html

Data resource:

http://www.stats.gov.cn/ (National Bureau of Statistics)

R studio Code:

```
ols<-lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +X10,data=data3)
summary(ols)

fitted.ols<-ols$ols.values
fitted_square2=fitted.ols^2
w<-lm(ols$residuals^2~fitted.ols+fitted_square2)
summary(w)

f=coeftest(ols,vcov=vcovHC(ols,"HC1"))
f%>%print()

lm(residuals(ols)^2~X2+X3+X8+X9,data=data3)
lm(Y~X5+X6+X8+X9+X10,data=data3)
anova(ols,l1)
```

```
vif(ols)

y.res<-residuals(ols);plot(y.res)
plot(ols)

summary<-summary(ols$residuals)
Q1<-summary[[2]]
Q2<-summary[[3]]
Q3<-summary[[5]]

skewness<-((Q1-Q2)+(Q3-Q2)/(Q3-Q1))
skewness

crPlot(ols)
```