Research and Analysis of the Causes of Wage Difference

Min Guo

Shanxi University of Finance and Economics, Taiyuan, Shanxi, China

1253866797@qq.com

Keywords: Wage differenses, R, Multiple regression, Factor analysis

Abstract: With the development of social economy, people pay more and more attention to the wage requirements and attach great importance on the salary. Meanwhile, the influence factors are also changing and increasing constantly, so it is extremely urgent to study the influence factors of wage differences. Although many scholars have conducted researches, this paper aims to find the key influence factors and establish appropriate models for analysis. The influence factors of wage differences are quite a lot, and itself has a lot to do with employees, so this paper selects the foreign 935 samples of an area, 9 variables reflect the main information for multiple linear regression, factor analysis as statistical analysis, describe the influence of the specific process, better reflect the reason of the differences of wage, also the corresponding explanation and suggestions is given

1. Introduction

1.1 Research Background

The wage difference is an issue of wide concerns all over the world. Because the wage level directly affects the daily work, life itself and the quality of life of residents, how to solve the employment problems and obtain high-paying jobs has become a big problem. There are many factors affecting the salary of residents, such as IQ, years of education (equivalent to the level of education), work experience, years of work, race and marital status, etc. In particular, racial discrimination still exists in the world. At the same time, enterprises will also refer to marital status when hiring employees. Therefore, this paper is devoted to the study of the factors affecting residents' wages.

1.2 Literature Review

In recent years, many experts and scholars have studied the influence factors of the income gap between urban and rural residents. Ma Bin and Zhang Furao demonstrated in their empirical analysis that the current income gap lies in education level, gender and other factors, and at the same time, through data analysis, they have proved the influence of these factors. In the analysis of the main factors affecting the income gap of urban residents, Zhou Yunbo also analyzed the factors leading to the income gap[1]. Chen Zongsheng and Zhou Yunbo investigated the factors affecting the income growth of urban residents from the perspective of population characteristics such as education level, and found that gender characteristics produced different income levels in the society, and different ages had significantly different impact on residents' income. The higher the level of education, the higher the residents' income was. Chen Lina[2]expressed her views on the influence factors of wage and income gap;Li Huazuo[3] also reached the same conclusion after studying the influence of residents' education level on their incomes, indicating that human capital is an important factor affecting residents' income.

2. Research Methods and Objectives

The data of this study includes a total of 935 samples and 17 variables. Through the analysis of data in the early stage, and in order to make the analysis process more simple and smooth, this paper extractes 9 main variables, which are respectively:

  DOI: 10.25236/edssr.2020.052

wage	monthly salary
exper	years of working
age	time of life
educ	schooling
hours	average working hours per week
tenure	office term
IQ	intelligence quotient
married	marital status
black	racial types

The variable wage is taken as a dependent variable to reflect the wage income of employees in the region, and the other 8 variables are taken as independent variables to explain the reasons for the wage difference.

In the process of this research, the statistical methods of regression analysis, principal component analysis and cluster analysis are mainly used. The regression analysis method is used to determine the interdependence between two or more variables. In the process of analyzing employee absence, there are many variables. Therefore, a multiple linear regression model is established to analyze how the independent variables exper, Tenure, Age, IQ, hours, educ, Black and married affect the dependent variable wage. Factor analysis uses the idea of “dimensionality reduction” to extract the common factor of the original variable and find out the main explanatory variable that can better reflect the variable information under the condition of little information loss.

Cluster analysis is applied to this study. At the end of the study, cluster the groups with similar salaries of employees in different regions according to the main factors, so as to understand the current situation of salaries of employees in this region more clearly and intuitively.

3. Descriptive Statistics

3.1 Total Description of Variables

First, the data was imported into R, and the meaning and attributes of the data were observed. According to the actual situation and literature review, the final 9 variables were left, including 2 categorization variables, 6 continuous independent variables and 1 dependent variable. Based on our knowledge, we know that independent variables have a certain degree of influence on dependent variables, and most of the information of influencing factors is included. In this way, the selection of variables can minimize the error of the model in the final regression. The following codes are:

```
data1<-read.table(“clipboard”,header = TRUE)
data2<-data1[,c(1,2,3,5,6,7,8,9,10)]
data2<-na.omit(data2)
attach(data2)
str(data2)
```

3.2 Variable Analysis

Analyze the distribution of the dependent variable wage and check the kernel density curve of the wage:

```
plot(density(wage))
```

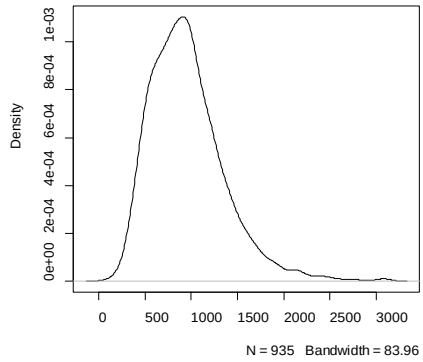

Fig.1 Nuclear Density Curve of the Wage

As shown in figure 1, it can be found that the wage distribution is concentrated in the range of 500-1500, and the distribution range is not very wide. Good data collection can be used to study the causes of wage differences. A descriptive analysis of wage:

```
> summary(wage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  115.0   669.0   905.0   957.9  1160.0  3078.0
>
```

Fig.2 Descriptive Statistics of the Wage

3.3 Analysis of Independent Variables

Descriptive analysis of continuous variables, such as “hours”, “IQ”, “educ”, “exper”, “tenure” and “age”, was conducted using Psych () package in R :

```
install.packages(“psych”)
library(psych)
describe(data2[,c(-1)])
```

```
> describe(data2[,c(-1,-8,-9)])
       vars   n   mean    sd median trimmed   mad min max range  skew
hours     1 935  43.93  7.22     40   42.80  0.00  20  80    60  1.59
IQ        2 935 101.28 15.05    102  101.87 14.83  50 145    95 -0.34
educ      3 935  13.47  2.20     12   13.33  1.48   9  18     9  0.55
exper     4 935  11.56  4.37     11   11.52  4.45   1  23    22  0.08
tenure    5 935   7.23  5.08      7    6.89  5.93   0  22    22  0.43
age       6 935  33.08  3.11     33   33.04  4.45  28  38    10  0.12
       kurtosis   se
hours      4.14 0.24
IQ        -0.03 0.49
educ      -0.74 0.07
exper     -0.57 0.14
tenure    -0.81 0.17
age       -1.26 0.10
>
```

Fig.3 Descriptive Analysis of Independent Variables

Fig.3 Is More Specific in Describing Various Eigenvalues of Independent Variables and Provides More Information.

3.4 Histogram is Used to Show the Frequency Distribution of These 6 Independent Variables:

```
hist(hours)
hist(IQ)
histeduc)
hist(exper)
hist(tenure)
hist(age)
```

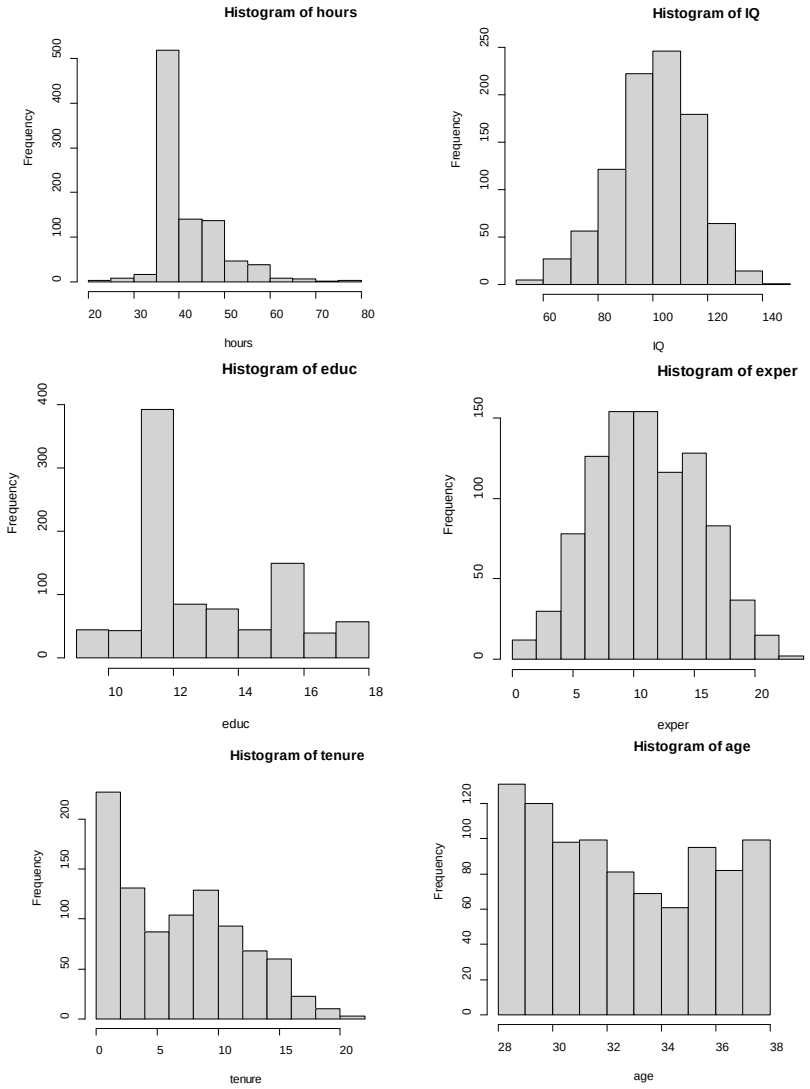

Fig.4 Histogram of Independent Variables

3.5 Do Frequency Statistics on Classified Data

Table(Black)

Table(Married)

The above order can be established and it can be counted as 120 blacks and 815 whites.There were 835 married persons and 100 unmarried persons.

Histogram shows more visually the frequency of blacks and whites and the frequency of marriage and non-marriage:

barplot(table(data2$black),name=c(“white”,”balck”))

barplot(table(data2$married),name=c(“unmarried”,”married”))

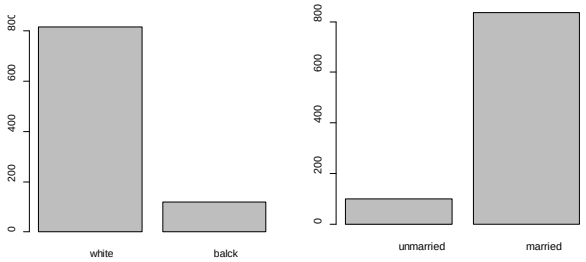

Fig.5 Histogram

4. Linear Regression Analysis Model

4.1 Multicollinearity Check

4.1.1 Coefficient of Association

Round (cor(data.frame(data2[,2:9])),2)

```
> round(cor(data.frame(data2[,2:9])),2)
          hours    IQ  educ exper tenure   age married black
hours      1.00  0.07  0.09 -0.06  -0.06  0.02    0.03 -0.11
IQ         0.07  1.00  0.52 -0.22   0.04 -0.04   -0.01 -0.39
educ       0.09  0.52  1.00 -0.46  -0.04 -0.01   -0.06 -0.18
exper     -0.06 -0.22 -0.46  1.00   0.24  0.50    0.11  0.06
tenure    -0.06  0.04 -0.04  0.24   1.00  0.27    0.07 -0.08
age        0.02 -0.04 -0.01  0.50   0.27  1.00    0.11 -0.04
married    0.03 -0.01 -0.06  0.11   0.07  0.11    1.00 -0.05
black     -0.11 -0.39 -0.18  0.06  -0.08 -0.04   -0.05  1.00
> |
```

Fig.6 Correlation Coefficien

4.1.2 Draw the Scatter Plot Matrix of Each Variable:

Pairs (wage~.,data=data2)

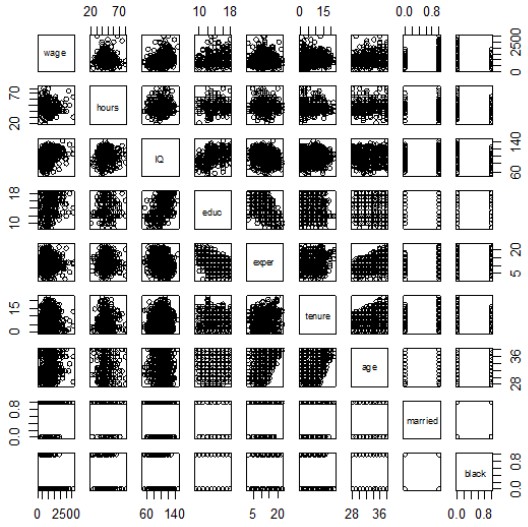

Fig.7 Correlation Coefficient Matrix

4.1.3 Multicollinearity Check

lm1<-lm(wage~.,data=data2)
library(car)
vif(lm1)

```
> vif(lm1)
   hours       IQ     educ    exper   tenure      age  married    black
1.025373 1.566965 1.767874 1.849195 1.116137 1.492517 1.022591 1.198076
> |
```

Fig.8 Vif

From the correlation coefficient and can be seen in the scatterplot matrix between 8 variables related degree is not high, there is no obvious linear relationship, and the variance inflation factor is less than 2 in figure 8, so the data there is no multicollinearity, explanatory variable “hours” and “IQ”, “educ”, “exper”, “tenure”, “age”, “married”, “black” cross between components is not big, the data is valid.

4.2 Establishment of Regression Model

Take wage as dependent variable, hours, IQ, educ, exper, Tenure, age, marriage and Black as independent variables, and make multiple linear regression:

lm1<-lm(wage~.,data=data2)

Display the detailed estimation results of the model:

summary(lm1)

The results are shown in the figure below:

```
> summary(lm1)

Call:
lm(formula = wage ~ ., data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-757.73 -238.79  -42.09  181.80 2113.79

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -684.0910   178.0966  -3.841 0.000131 ***
hours         -3.0266     1.6632  -1.820 0.069125 .
IQ             4.0039     0.9868   4.058 5.38e-05 ***
educ          54.2801     7.1824   7.557 9.87e-14 ***
exper          9.3655     3.6886   2.539 0.011278 *
tenure         5.2269     2.4701   2.116 0.034602 *
age           10.8089     4.6645   2.317 0.020707 *
married      167.6526    38.8052   4.320 1.73e-05 ***
black       -116.8366    38.8111  -3.010 0.002680 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 362.6 on 926 degrees of freedom
Multiple R-squared:  0.2026,    Adjusted R-squared:  0.1957
F-statistic: 29.41 on 8 and 926 DF,  p-value: < 2.2e-16
```

Fig.9 Linear Regression

According to the results, in this model, the variables IQ, educ and married pass the significance test. At any significance level, the T value is greater than the critical value, while the P value is less than the significance level, and falls in the rejection domain. Therefore, these three variables have a great effect on the interpretation of the explained variable wage.The variable Black passed the significance test at the significance level of 0.001. The variables exper, Tenure and AGE passed the significance test at the significance level of 0.01. While hours is not very significant, indicating that the independent variables exper, Tenure, age have an explanatory effect on the dependent variable wage, but the effect is not significant.

At the same time, taking into account the fact that the salary is closely related to the contribution made, the efficiency achieved and the result achieved, and these factors are closely related to a person's ability level and the time and energy that can be contributed, so the regression model has a good illustrative effect.

We removed the insignificant variable hours and made another regression model:

```
> lm2<-lm(wage~IQ+educ+exper+tenure+age+black+married,data=data2)
> summary(lm2)

Call:
lm(formula = wage ~ IQ + educ + exper + tenure + age + black +
    married, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-778.24 -241.12  -39.78  175.20 2156.14

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -798.833    166.768  -4.790 1.94e-06 ***
IQ             3.993      0.988   4.042 5.74e-05 ***
educ          53.814      7.187   7.488 1.63e-13 ***
exper          9.625      3.690   2.608  0.00925 **
tenure         5.522      2.468   2.237  0.02549 *
age           10.366      4.664   2.223  0.02648 *
black       -110.629     38.709  -2.858  0.00436 **
married      165.242     38.831   4.255 2.30e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 363.1 on 927 degrees of freedom
Multiple R-squared:  0.1997,    Adjusted R-squared:  0.1937
F-statistic: 33.05 on 7 and 927 DF,  p-value: < 2.2e-16
```

Fig.10 Removes the Regression of Hours

4.3 Heteroscedasticity Was Tested by Bp Test

install.packages(“zoo”)
install.packages(“lmtest”)
library(zoo)
library(lmtest)
bptest(lm1)
The results are shown in the figure below:

```
> bptest(lm2)

        studentized Breusch-Pagan test

data:  lm2
BP = 28.893, df = 7, p-value = 0.0001513
```

Fig.11 Heteroscedasticity Test

Due to the P-value is less than 0.05 and falls into the rejection domain at the significance level of 0.05, heteroskedasticity exists, so heteroskedasticity needs to be corrected. 1/ ABS (e) is selected as the weight. The correction process is as follows:

```
> lm3<-lm(wage~IQ+educ+exper+tenure+age+black+married,data=data2,weights=1/abs(e))
> summary(lm3)

Call:
lm(formula = wage ~ IQ + educ + exper + tenure + age + black +
    married, data = data2, weights = 1/abs(e))

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-28.895 -14.805  -4.907  14.156  47.124

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -769.1223    66.4887 -11.568  < 2e-16 ***
IQ             4.0008     0.3948  10.134  < 2e-16 ***
educ          51.8709     3.0009  17.285  < 2e-16 ***
exper          8.2725     1.2896   6.415 2.24e-10 ***
tenure         6.0772     0.8604   7.063 3.19e-12 ***
age           10.4005     1.7508   5.940 4.02e-09 ***
black       -111.1950    14.0436  -7.918 6.89e-15 ***
married      161.2165    15.9211  10.126  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.68 on 927 degrees of freedom
Multiple R-squared:  0.6187,    Adjusted R-squared:  0.6158
F-statistic: 214.9 on 7 and 927 DF,  p-value: < 2.2e-16
```

Fig.12 Corrected Heteroscedasticity

By observing and comparing the above two figures' R^2 and the corrected R^2, the corrected heteroscedasticity is significantly larger and all the variables are very significant, so the revised model is better.

4.4 The Regression Model is Analyzed Practically

$$wage = -796.12 + 4IQ + 51.87educ + 8.27\exp er + 6.08tenure + 10.4age - 111.2black + 161.22married$$

In the regression model, whether “black” is a black or not is a binary variable whose regression coefficient reflects the difference between black and white wages of about $111.2.The regression coefficient of “married”, reflecting the difference in wages between married and unmarried, was $161.2.The coefficient of IQ reflects the $4 increase in wages for every 1 increase in IQ. For each additional year of schooling, the salary increases by 51.87 yuan. The salary increases by 8.27 yuan for each additional year of working experience. The salary will be increased by 6.08 yuan for each additional year of service.

Salary is positively correlated with IQ, years of education, years of work, years of employment, age and whether they are married, and negatively correlated with whether they are black, indicating that employees with more experience, higher ability, higher education level and longer working hours will have higher salary. It may be due to regional discrimination that the salary of black people is lower. Among these factors, years of education, IQ and age are three objective factors, and the wage difference caused by them lies in the subjective consciousness of employees themselves. The results of the model are in line with the trend of psychological dynamic changes. The higher an employee's IQ and education, the more adaptable he or she is to the job, the better he or she is at

solving problems, and the more mentally agile he or she is. Experienced employees know the rules of the workplace, know what to do and what not to do, and will be likable. Older employees who have worked for a long time understand the workings of the company and the boss's temperament and can adapt to the job. Married workers are more eager to work and therefore take their work more seriously and devote more energy to it. All of these are in line with human nature, so the establishment of the model is a more scientific reflection of life.

5. Factor Analysis

install.packages("psych")
brary(psych)
KMO(data2)

```
> data3<-data2[,c(-2)]
> KMO(data3)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = data3)
Overall MSA =  0.59
MSA for each item =
   wage      IQ    educ   exper  tenure     age married   black
   0.72    0.65    0.55    0.53    0.77    0.50    0.65    0.66
>
```

Fig.13 Kmo Value

factor.pa(data3)
summary(factor.pa(data3))
fa.parallel(data3)

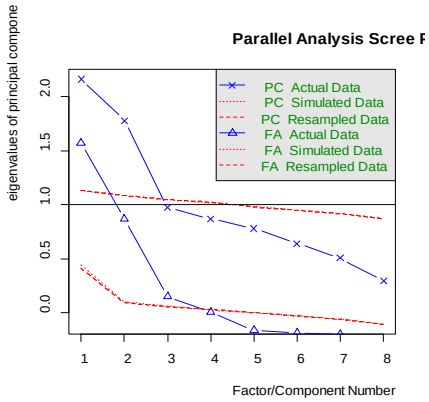

Fig.14 Lithotripsy

In factor analysis, KMO > 50% ensures the implementation of factor analysis. However, it can be seen from the gravel diagram that the broken line has a downward trend, so it is difficult to find the best inflection point position. Combined with the correlation matrix between independent variables, it can be considered that the correlation between variables is not strong, and each variable has an explanatory effect on the wage difference. It is consistent with the conclusion of regression model.

6. Cluster Analysis

To better reflect the results of the regression analysis, the following analysis is based on the independent (significant) variable IQ. In this process, k-means clustering method is used for cluster analysis. In the process of analysis, because three-dimensional space is difficult to grasp intuitively on the plane, the principal component method is used to project samples from three-dimensional space to two-dimensional space, and according to the calculated principal component score, absenteeism levels are divided into three categories, which reflect the absenteeism status of employees in the company.

principal component analysis :
data4<-data2[,c(3,4,5,6,7,8,9)]
data5<-scale(data4)

```
prcomp(data5)
summary(prcomp(data5))
score<-as.matrix(data5)%*%prcomp(data5)$rotation[,1:2]
score
cluster analysis :
clus<-kmeans(data4,3)
clus1<-clus$cluster
plot(score[,1],score[,2],pch=clus1)
```

```
> clus
K-means clustering with 3 clusters of sizes 424, 179, 332

Cluster means:
        IQ     educ    exper   tenure      age   married      black
1 98.81368 13.01651 11.96934 7.363208 33.11085 0.9009434 0.08254717
2 78.58659 11.93855 12.90503 6.793296 33.22346 0.8938547 0.39664804
3 116.67169 14.87048 10.32229 7.307229 32.96386 0.8825301 0.04216867
```

Fig.15 Cluster Analysis

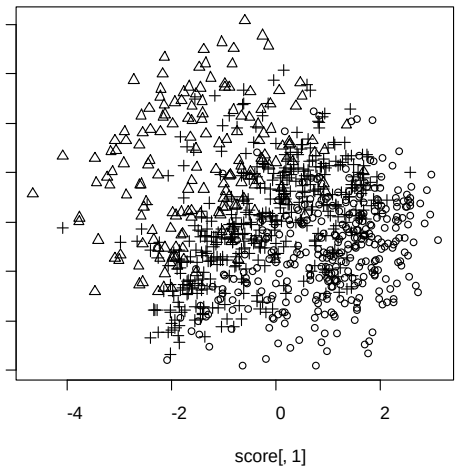

Fig.16 Cluster Diagram

Fig.16 Shows the Results after the K-Means Clustering Method is Used. the Salaries of These 935 Employees Can Be Divided into Three Categories: High, Medium and Low.

7. Research Results and Analysis

Through the above analysis of the wage situation, it is found that the wage difference of employees in this region is affected by many factors.

This study mainly explores the causes of wage difference from the perspective of individual employees. In the personal aspect, it is mainly due to the employees' personal ability and work experience. (1)Ability to work. A standardized set of requirements for a person to hold a position. Ability usually refers to the power that a person can exert. Human ability includes instinct, potential, ability and skill, which directly affects the quality and efficiency of a person's work. There are two psychological explanations for improving one's ability to work: one is that one's capable, and the other is that one is likely to be capable in the future. The actual ability shown by individual behavior is called "achievement" in psychology, while the ability shown through learning and training or in behavior is called "potential" in psychology. Work ability directly affects the quality and efficiency of the work, and further affects the income of wages. (2)Work experience. It refers to the applicant's work history, whether paid or unpaid, full time or part time. Work experience is one of the main reference factors in selecting and recruiting personnel. Work experience affects the level of employees to adapt to the work, but also affects the response to the work of sensitivity and adaptability, interpersonal skills, people who have rich work experience will generally have better adaptability, so the higher the salary. (3)Marital status. Marital status has a great impact on employees' work attitude and working time. Married people's responsibility to family makes them

need to make money seriously, while unmarried people have no family pressure and are less serious about work than married people.

8. Suggestions and Countermeasures

From the perspective of employees themselves, their salary is positively correlated with IQ and years of education. They can improve their ability and quality through continuous learning and further study, enhance their knowledge, obtain higher education and rich theoretical knowledge, and thus obtain higher salary. At the same time, the salary of employees is also positively correlated with their working experience and working years, so they can gain more work experience, accumulate more practical ability and further improve their business ability, thus increasing their salary.

From the perspective of the society and employers, racial discrimination should be eliminated instead of being differentiated with colored glasses. Instead, individual's true level and working ability should be considered, and promote racial equality. At the same time, the worry and influence of marital status should be weakened, and talents suitable for job positions should be sought from the essence of work.

References

[1] Zhou Yunbo. Analysis of The Main Factors Affecting the Income Gap of Urban Residents.

[2] Chen lina. Influencing factors of wage income gap [J].Theoretical Discussion, 2011, (11): 12 -- 13.

[3] Li Huazuo. The Influence of Residents' Education Level on Income.