Research on the Factors that Influence Wages - Take the Example of the Data Science Industry

Yuyan Chen[1,a,*]

[1]School of Social Science, The University of Manchester, Oxford Street, Manchester, M1 9PL, United Kingdom
a. yuyan.chen-3@student.manchester.ac.uk
**corresponding author*

Abstract: This paper aims to illustrate which variables affect compensation by examining the factors that influence compensation in the data science industry. A total of six independent variables are considered in this paper: experience level, employment type, country of residence, remote ratio, company location, and company size. The paper recoded the raw data, converted the textual data into numerical data, and then tested for multicollinearity. The results show weak correlations between the independent variables. Finally, these six independent variables were used to analyse the dependent variable wage using a linear regression model, which showed that, except the employment type and the remote ratio, the other four variables had a greater impact on the model outcome, that is, wage. To improve the accuracy of the model further, the sample size needs to be expanded to improve the power of the test, and other variables need to be considered to explain the influences on wages more fully.

Keywords: Wage difference, linear regression model, data science.

1. Introduction

In labor economics, wages and employment are the two most important variables. Wages can be considered the primary source of income and have a significant impact on an individual's well-being. They also play a crucial part in the dynamics of labour supply and demand, which in turn drives market forces that determine employment and unemployment. However, employees' varying qualifications and contributions to the organization lead to different wages among them [1]. Therefore, it is important to understand the factors that contribute to wage differences.

Lazear pointed out that current work experience has a significant impact on wages, and that more accumulated work experience results in higher wage growth [2]. Lazear used Ordinary Least Squares (OLS) regression analyses to estimate the factors affecting wage growth, including the level of education, work experience, age, duration of on-the-job training, and changes in working hours. In addition to this, the geographical location of the labor force is also related to the wage differentials between them. Farrokhi and Jinkins found that people working in more geographically isolated cities have a lower skills wage premium, with 16.5 percent of the difference in the skills wage premium attributable to the location of the city [3]. They applied regression analysis to verify this relationship. The context of the study builds on the intersection of economic geography and labor economics,

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

focusing on how the concentration of the same occupation in each region affects the wages of its practitioners.

Besides firm location contributing to wage differentials, an employee's place of residence also contributes to wage differentials. Vejlin used ordinary least squares (OLS) regression analyses to examine the effects of distance between residence and workplace on wage levels, job mobility, and wage changes. It was found that workers who live farther from their place of work receive higher wages on average, which is consistent with the theory of compensatory wage differentials, whereby workers demand higher wages for jobs away from their place of residence to compensate for traveling costs [4]. Apart from field offices, the surge in remote working during the COVID-19 outbreak has led to the discovery of the convenience of remote office work and its growing popularity, which can also cause differences in salaries. Sabrina and Victoria have pointed this point out in their study. They showed that the wages of remote workers were generally higher than those of on-site workers, and this gap increased significantly during the epidemic [5]. Full-time and part-time jobs also contribute to the wage gap. Baffoe-Bonnie's study provides an in-depth insight into the wage gap between full-time and part-time workers through multiple regression analysis and sample selection modeling. Baffoe-Bonnie found that the wage differential between full-time and part-time workers decreases from 33.1 percent to 23.0 percent when individual characteristics, labor market conditions, and other factors are considered [6]. The study by Gerlach and Schmidt revealed a significant positive correlation between firm size and wages through an in-depth analysis of German socio-economic panel data. The study employed a variety of statistical methods and economic models to control for the diversity of individual worker characteristics and work environments, thus providing strong evidence that firm size affects wage differentials [7].

This paper has selected the data science industry as an example to examine the factors that contribute to the wage gap. Data science is intimately connected to the daily lives in work, learning, and the economy [8]. With time, industries need professionals to process the collected data and help them identify the errors that lead to loss or failure in the industry and find solutions to them. The field that manages these problems is called data science. Thus, the future era will revolve around data, and this paper needs a lot of data science enthusiasts and the compensation they get is huge [9, 10]. This paper focuses on the impact of six variables (experience level, employment type, remote time, employee's residence, company size, and company location) on employee's salary, and further selects an appropriate model to study the correlation between these factors and data scientist’s salary. In review, this paper will use a multiple linear regression model to study the effect of these five factors on salary.

2. Methods

2.1. Data Source

The dataset used in this paper is from the Kaggle website (Data Science Salaries 2023). It contains data such as salaries of employees in data science-related positions in different companies in the year 2023 from aijobs.net The dataset is in CSV format containing 3755 observations.

2.2. Variable Selection

Employees’ experience level, employment type, employees’ country of residence, company country and company size contained in the original dataset were all textual data, so they needed to be recoded into numeric variables to provide more accurate analyses.

For the experience level, the original dataset was divided into four categories: EN, MI, SE, and EX, with increasing levels of experience, so 1-4 was used to represent each category, with larger numbers representing more experience. For the types of employment, the original dataset contains four categories: FT, PT, CT, and FL, which have ascending degrees of job flexibility, so the same 1-4 are used to represent the job flexibility of employees, with larger numbers indicating greater job flexibility. For the country of residence, since previous studies have found that workers who live further away from the company receive higher wages, the employee's country of residence is set to 1 for those who live in the same country as the company and 0 for those who live in a different country. A binary variable was used for the country in which the company is located. Firms located in the US were set to 1 for the rest and 0 for the rest, which was done to avoid the dummy variable trap. Firm size in the original dataset contains three categories: S, M, and L. These are replaced by recoding with numbers 1-3, with larger numbers representing larger firms.

Thus, this paper has obtained six independent variables (experience level, employment type, residence country, remote ratio, company country, company size) and one dependent variable (data science employees' salaries) in Table 1.

Table 1: List of variables

Logogram	Variable	Definition
X_1	Experience level	The experience level in the job during the year: EN (1), MI (2), SE (3), EX (4).
X_2	Employment type	The type of employment for the role: FT (1), PT (2), CT (3), FL (4).
X_3	Residence country	Employee's primary country of residence during the work year. Suppose it is the same as the company country (1), otherwise (0).
X_4	Remote ratio	The overall amount of work done remotely.
X_5	Company country	The country of the employer's main office or contracting branch. If the company is in the US (1), otherwise (0)
X_6	Company size	The median number of people that worked for the company during the year: S (1), M (2), L (3).
Y	Salaryinusd	The salary is in USD.

2.3. Method Introduction

In this paper, multiple linear regression models are used to test whether there is a statistically significant relationship between the selected independent variables (X1-X6) and the dependent variable (Y). The equations are as follows:

$$\mathrm{Y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_6 x_6 + \mu \tag{1}$$

3. Results and Discussion

3.1. Descriptive Analysis

Table 2 shows the descriptive statistics for each variable. The figure demonstrates that the dataset contains a total of 3755 observations, of which only salaryinusd is a continuous variable. The mean value of salary is at $137,570 with a standard deviation of $63,055.63, indicating a high degree of volatility and a more dispersed distribution of salary levels. The standard deviation of the X_1, X_2, X_3, X_5, and X_6 are small, indicating that the data are concentrated. The large variance of X4 indicates that the percentage of telecommuting varies widely among employees and may be fully remote, partially remote, and fully on-site work.

Table 2: Descriptive statistic table.

	Mean	Median	Min	Max	Sd
Y	137570.39	135000	5132	450000	63055.63
X_1	2.65	3	1	4	0.68
X_2	1.02	1	1	4	0.2
X_3	0.97	1	0	1	0.16
X_4	46.27	0	0	100	48.59
X_5	0.81	1	0	1	0.39
X_6	2.08	2	1	3	0.39

3.2. Correlation Analysis

Before conducting multiple linear regression analyses, the degree of correlation between the independent variables needs to be calculated to ensure that there is no multicollinearity between the variables that would bias the regression results.

Table 3: Correlation matrix of independent variables.

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1					
X_2	-0.1001	1				
X_3	0.1442	-0.2336	1			
X_4	-0.0437	0.0612	-0.1236	1		
X_5	0.3126	-0.0697	0.1578	-0.0777	1	
X_6	-0.0707	-0.0636	0.0035	0.0369	-0.0670	1

According to Table 3, the correlation between the independent variables is not strong. None of the correlation coefficients exceed 0.5, indicating a weak correlation between all the independent variables. Hence the next step of multiple linear regression analysis can be carried out.

3.3. Model Analysis

Table 4 shows the results of the regression model. Based on the p-value, all the independent variables are significant except for X_2 and X_4, which are not significant. This indicates that the employment type and remote ratio have less impact on data scientists' salaries.

Table 4: Regression results.

	β	Std.Error	T value	p-value
Intercept	-20268.78	9894.55	-2.048	4.06×10^{-3}
X_1	29835.67	1327.04	22.483	2×10^{-16}
X_2	-7908.85	4447.69	-1.778	7.54×10^{-2}
X_3	27062.25	5631.60	4.805	1.6×10^{-6}
X_4	-18.58	17.63	-1.054	2.92×10^{-1}
X_5	57054.22	2295.02	24.860	2×10^{-16}
X_6	7329.77	2176.84	3.367	7.67×10^{-4}

On top of that, X_5, the location of the company has the biggest impact on wages. It has a coefficient of 57,054.22, indicating that data scientists at companies located in the US are paid $57,054.22 more than data scientists at companies that are not located in the US when all other things are held constant. X1, experience level, also has a significant effect on wages. Executive-level (EX) data scientists receive $89,507.01 more in wages than entry-level (EN) data scientists. Where an employee lives also has a large impact on wages. Data scientists who live in the same place as the company is located will make $27,062 more than those who live in a different place, given the other variables are held constant in the model. The effect of firm size on wages is also significant. Holding all other variables constant, a large firm will pay $14,659.54 more than a small firm.

4. Conclusion

The above analyses reveal that work experience, company location, place of residence, and company size all have an impact on wages. Although the regression results show that remote ratio and employment type do not have a significant effect on salaries, this may be due to insufficiently large sample sizes. With a small sample size, even if some of the independent variables have some effect on the dependent variable, the regression analysis may fail to detect this significance due to insufficient statistical power. Insufficient sample size can lead to an increase in the standard error and hence a larger p-value. Therefore, the sample size needs to be increased to improve the statistical power.

In addition, more variables such as gender, age, and so on can be considered to analyse the impact on wages in addition to those mentioned in this paper. The inclusion of more variables can make the analysis more comprehensive and accurate.

References

[1] *Tachibanaki, T. (1998) Introduction to Wage Differentials: An International Comparison. In: Tachibanaki, T. (eds) Wage Differentials. Palgrave Macmillan, London.*

[2] *Lazear, E. (1974) Age, experience and wage growth. National Bureau of Economic Research.*

[3] *Farrokhi, F. and Jinkins, D. (2019) Wage inequality and the location of cities. Journal of Urban Economics, 111, 76–92.*

[4] *Vejlin, R. (2013) Residential location, job location, and wages: Theory and Empirics. LABOUR, 27(2), 115–139.*

[5] *Pabilonia, S.W. and Vernon, V. (2023) Remote work, wages, and hours worked in the United States Sabrina Wulff Pabilonia, Victoria Vernon. Essen: Global Labor Organization (GLO).*

[6] *Baffoe-Bonnie, J. (2004) Interindustry part-time and full-time wage differentials: Regional and National Analysis. Applied Economics, 36(2), 107-118.*

[7] *Gerlach, K. and Schmidt, E.M. (1990) Firm Size and Wages. LABOUR, 4, 27-50.*

[8] *Cao, L. (2017) Data science: A comprehensive overview: ACM computing surveys. University of Technology Sydney, 3.*

[9] *Kaur, A., Verma, D. and Kaur, N. (2022) Utilizing quantitative data science salary analysis to predict job salaries. 2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT).*

[10] *Akobeng, A.K. (2016) Understanding type I and type II errors, statistical power and sample size. Acta Paediatr, 105, 605-609.*