EMORY UNIVERSITY
DEPARTMENT OF MATHEMATICS
Atlanta (GA) USA
Project 2

MATH345 MATH MODELING

Submitted By:
Group Neptune

Authors:
Ashley Zhou
Hyesun Jun
Noufissa Guennoun
Nina Bilsel
Robyn Goldberg

Spring 2022
Group Report 2
2022.03.02

EMORY
UNIVERSITY

Abstract

Determining employees' salaries requires knowledge and understanding regarding qualifications, cost of living, and the salaries' of comparable jobs. Ideally, salaries would be determined based solely on merit. However, this is often not the case, as many employers may take race, gender, and ethnicity into consideration. Our goal in the mathematical model is to construct a model that solely takes the important qualifications and factors into consideration, and excludes all other factors. We created a mathematical model to ensure formal justice by calculating the annual salary for a given employee based solely on their qualifications, cost of living, and a starting salary. This model takes into account multiple qualifications and the weighting of each different qualification to ensure consistent practices for all employees. We then tested our model using county-level data in Georgia to determine significant qualifications and assign weights to these qualifications. We also tested our model against the small data sets provided, and we found that our model was able to produce consistent results with the data provided.

Keywords: meritocracy, compensation, qualifications

Contents

1 Introduction

1.1 Background

How should an employer decide how much to pay their employees? Ideally, an individual's compensation would be determined based on factors relating to their merit. However, this is unfortunately not always the case in society today. Rather than merit, factors such as race, gender, and ethnicity are often more predictive of an individual's compensation. According to Pew Research Center, women, Black people, and individuals who identify as Hispanic consistently receive lower salaries than males.[5]

This paper will propose an alternative model that employers can utilize when determining salary, in attempt to eliminate current discriminatory practices in salary determination. Our model solely takes into account an employee's qualifications and cost of living rather than demographic factors that contribute to minority communities receiving lower salaries. Our model will allow employers to independently determine job related factors and qualifications to calculate an unbiased salary for their employees. These qualifications may include, but are not limited to, education, seniority, expertise on the subject matter, communication skills, and more. This model allows employers to calculate a consistent salary for all their employees based qualifications the employer considers valuable. The use of this model will help create a fair and just wage scale to combat current discriminatory practices.

1.2 Assumptions

There are several assumptions that must be made in order to form a model:

1. Employers understand how to weight the qualifications.

   Our model takes the employer's preferences into consideration, giving the employer the flexibility to place more weight on certain qualifications. This requires the employer to understand and determine how to weight each qualification to apply the model. While this may be based on a number of years or the level of education, they may want to bring in alternative factors for weighting, and they need to have a basic understanding about that.

2. Employers have all the information on the candidates' skill sets.

   In order for the model to be accurate, the employer must be aware of the employee's skill set. This includes, but is not limited to, their level of education, their level of experience, and their seniority.

3. Employers can measure the skill sets equivalently among candidates.

   In order for our model to address inequalities, we need to assume all skills are measured equivalently to minimize bias.

4. The data based on Georgia is a good indicator of the national trends.

   Therefore, even though we base our cost of living and starting salary from data collected in Georgia, it would be a good indicator for national trends.

5. Weightings are accurate.

   The qualifications in our model are weighted differently depending on other present qualifications, the weight of their qualification, and the bias of the employer. These weightings must be accurate in order for our model as a whole to be accurate.

6. Qualifications are the same for different job occupations.

   For example, if the two qualifications that we choose to examine are education and seniority, we must apply these same qualifications to different job occupations in order to make an accurate comparison.

7. Post-COVID data would be approximately the same as pre-COVID data that we used (1994-1995).

   Our model uses date from 1994-1995, a pre-COVID time. However, in order for the model to be accurate in present day, we assume that our 2019 data is approximately the same as post-COVID data.

8. Starting salary can be based on the minimum wage in Georgia, and cost of living in Atlanta would be same throughout Georgia.

   Since the starting salary is subjective, and depends on the company's preference, we assumed that the minimum wage in Georgia would be the base for the starting salary. Also, the employers would determine the starting salary by adding qualifications on the minimum wage.

Moreover, the cost of living is also based on data collected in Atlanta, Georgia.[1]

2 Methods

The primary hypothesis, that the education level as well as the seniority would have a strong positive correlation with the compensation will be tested in the mathematical model we constructed. First, we constructed some assumptions to model the real world case. Then, we brainstormed on what qualifications would be contained in both positive and negative criteria that would have a relationship with the compensation. Then, we conducted non-linear regression to identify the significant contributing qualifications to the salary system using the national data from the University of California Irvine website, originally collected by the US Census Bureau website. [3] We developed the mathematical model from a simple linear model to incorporate exponential, quadratic, or even logarithmic models as well as different qualifications.

Data

The national data was collected from the University of California, Irvine, School of Information and Computer Sciences website, originally collected by US Census Bureau [3]. This dataset documents 199523 rows of individual's demographic data like age, class of worker, adjusted gross income, education level, wage per hour, marital status, sex, citizenship status etc., collected in the United States. We filtered this dataset because it included information on unemployed people as well, and we chose to focus on factors like age, education level, wage, race, sex, full time or part time employment, citizenship status, and weeks of work yearly. After the data reduction process, the cleaned dataset contained 11304 rows of data.

Table 1

Demographics of individuals including citizenship status, age, sex, full time or part time employment, race, weeks work yearly, and wage.

Category	Level	n,%
	All	11304, 100%
Citizenship status	Native	10174, 90%
	Foreign	340, 10%
Age	Age 18 to 24	1992, 17.6%
	Age 25 to 34	2653, 23.5%
	Age 35 to 44	2809, 24.8%
	Age 45 to 54	1925, 17%
	Age 55 to 64	1992, 918, 8%
Sex	Male	5446, 48.2%
	Female	5858, 51.8%
Full time or Part time	Full time	4944, 43.7%
	Part time	436, 3.8%
	Children or Armed Forces	5667, 50.1%
Race	White	9558, 84.5%
	Black	1228, 10.8%
	American Indian etc.	134, 1.2%
	Asian or Pacific Islander	268, 2.4%
	Other	116, 1%
Weeks work yearly	0 to 52	11179, 98.9%
	52 or greater	125, 1.1%
Wage($)	0-10,000	397, 3.5%
	10,000-20,000	3305, 29.2%
	30,000-40,000	210, 1.9%
	40,000-50,000	1773, 15.7%
	50,000-60,000	1350, 11.9%
	60,000+	3782, 33.4%

3 Our Model

3.1 Our Mathematical Model

Our model determines the optimal salary for a given employee, taking their qualifications and cost of living into consideration. Our model takes

multiple qualifications into account, as well as the weighting of each qualification. It begins with finding the starting salary, and then incorporating the cost of living in different areas. Then, we have functions for each of the qualifications, and these functions are dependent on both x and y.

The following variables are integral in our model.

- x_i = the weight value for each qualification, where i = 1,2,3,...,n.
- y_i = the qualification value for each of the distinct qualifications, where i = 1,2,3,...,n
- $a, b, c, ...$ = sets of all real numbers
- n = the number of qualifications we are studying
- AS = annual salary
- COL = cost of living
- SS = starting salary

Let us begin with the simple version of our model.

$$AS = COL + SS + F \tag{1}$$

In our equation, we have a function called F. This function is dependent on x and y, and the number of functions, F, is dependent on the number of qualifications, n, we are studying.

Next, we must determine how to define the function F. Since we want to be able to integrate both polynomial and exponential functions in this function, F, we make F be a poly-exponential function. Let us define F:

$$F_i = y_i(a_0 + a_1x_i + a_2x_i^2 + a_3x_i^3 + ... + a_ne^{x_i}) \tag{2}$$

This function for F enables us to create non-linear equations for each of the qualifications, which ensures that the employer can alter compensations on a polynomial or exponential basis.

Our model:

$$AS = COL + SS + \sum_{i=1}^{n} F_i \tag{3}$$

3.2 Model explained

Our model takes in an employee's qualifications and the weightings for those qualifications, as well as the cost of living in the area and the starting salary for that job, and outputs the annual salary for that employee.

First, we find the qualifications that we deemed the most important. We began by studying seniority and education. Therefore, out model had two functions of F. Then, we obtain the model:

$$AS = COL + SS + F_1 + F_2 \tag{4}$$

Then, we can find two separate equations for each of the qualifications.

It is also important to note what the values for x_i and y_i are. First, x_i represents the weight given to the qualification. In many cases, this is found to be the number of years working, the number of years of education, or the number of awards received. However, this can also be a completely subjective value given by the employer. Next, y_i represents the qualification, and the actual value for that. That is found from a value from data analysis, linear regression, or non-linear regression. Similarly, the values for a_i can also be found from data analysis or regression models. This real number is helpful in ensuring the value of compensation aligns with the employers goals. Then, we multiply the values of a_i, x_i, and y_i together in order to obtain the monetary value for the qualification.

Our model is helpful as it enables an employer to be flexible with how they compensate their employees. In particular, some employers may decide to increase compensation linearly. They believe that the number of years of education will linearly increase the amount of money they receive for that qualification. However, what if the employer places even more weight on education than this linear model would? What if they believe that the number of years a person is educated for should exponentially increase the quantity of money received? This is the reason our model is helpful.

This mathematical model enables the employer to take many different factors into account, and the salaries of their different employees do not have to have a linear relationship. They have the ability to decide how to increase salaries by different factors.

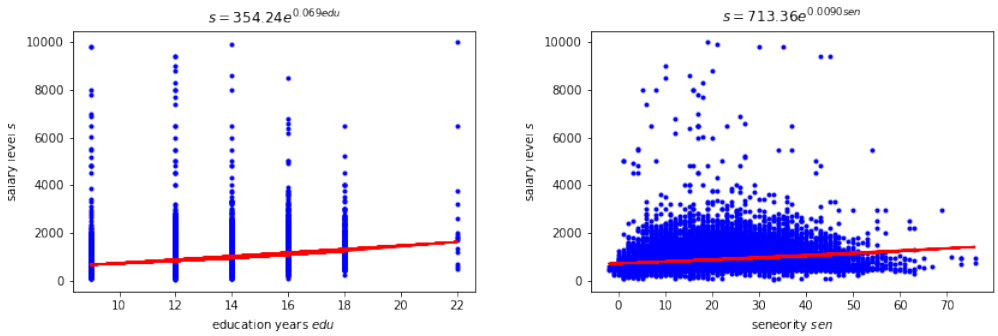
The next part of our model examines the cost of living for different places around the country and helps an employer ensure that they are incorporating this cost into their compensation. Cost of living is described as the amount of money a person needs according to the cost of goods,

location, and their preferences.[4] While it is beyond the scope of our model to examine a single individual's preferences, it can be interesting to examine a cost of living index in order to better understand how annual salaries will change between people in different locations. [2]

4 Solution

4.1 Data Fitting

We conduct exponential regression for wage on seniority and education as shown in **Figure 1**, which shows the growth in salary when individuals are better educated or more experience.



(a) salary $\approx 354.24e^{.069\text{education}}$ (b) salary $\approx 713.36e^{.0090\text{seniority}}$

Figure 1: Regression for salary on education and seniority.

Population	Average Salary
Female	875.22
Male	1089.23
White	989.15
Black	887.49

Table 1: Unfair Salary

We analyzed the data set from national survey by population groups as shown in **Table 1**, which indicates the unfair phenomenon in salary distribution. To build a more fair salary deciding model we conduct

multi-variable nonlinear regression for wage on seniority and education, which gives a nonlinear salary deciding formula:

$$\begin{aligned}\text{salary} = &\ 37.39 * \text{seniority} - 0.60 * \text{seniority}^2 \\ &- 166.16 * \text{education} + 9.38 * \text{education}^2 \\ &+ 1144.39,\end{aligned}$$

which is shown in **Figure 2**. See **Section 6.1** codes.

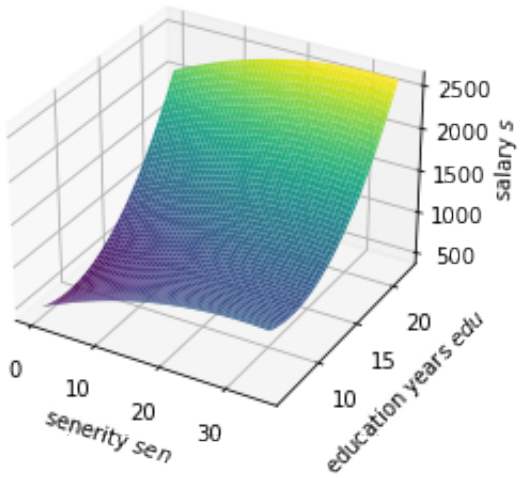


Figure 2: Salary on seniority and education.

4.2 Solution: Company 1 and Company 2

In order to study our model against the data provided, we first acknowledge that we are studying only one qualification, so there will only be one F function. Then, our model will look like this:

$$AS = COL + SS + y_i(a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 + ... + a_n e^{x_i}) \qquad (5)$$

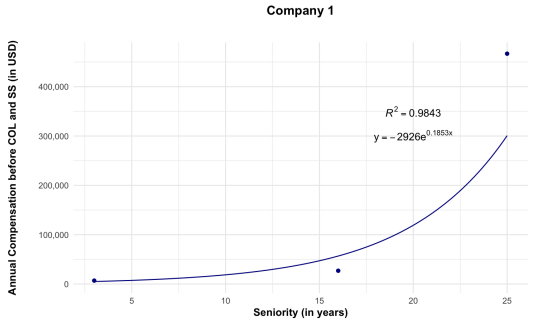


Figure 3: Company 1

We found that the minimum wage in Georgia is \$15,080 and the cost to live in Georgia is \$18,000. Then, we get the following equation from the graph:

$$AS = 15080 + 18000 + 2962.1e^{0.1853x} \tag{6}$$

Then, we can plug in the different values for x, 3, 16, and 25, and we obtain similar annual salaries to those supplied in the data set.

Therefore, we fit our model to the data, and we found that we were able to make our model consistent with the data.

It is important to note that in this part of the project, very few data points were given to us, so it would be probable that a model could not completely portray the data provided. However, because our model is flexible and can be altered to fit an employer's needs, we are able to make the model consistent with the data.

In the second part of the problem, we are given a different set of data, which has a greater number of data points, but contains some duplicate points. This led us to believe that our F function could be made into a piece-wise function.

From this example, we see that the graph is a polynomial, but when it contains the duplicated points, it would be a piece-wise function.

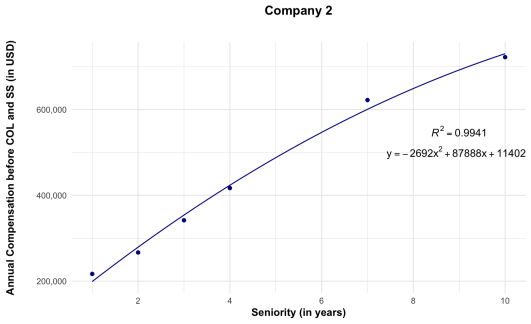


Figure 4: Company 2

Therefore, the equation would look like this:

$$y = \begin{cases} -2629.1x^2 + 87888x + 114026 & x = 1, 2, 3, 4 \\ -2629.1(4)^2 + 87888(4) + 114026 & x = 5, 6 \\ -2629.1x^2 + 87888x + 114026 & 6 < x < 7 \\ -2629.1(7)^2 + 87888(7) + 114026 & x = 7, 8, 9 \\ -2629.1x^2 + 87888x + 114026 & 9 < x \leq 10 \end{cases} \tag{7}$$

In both of these examples, both the employers and the employees are taken into consideration. First, the employer has the ability to account for the starting salary and the type of growth they want to see as the seniority in their employees increases. Second, the players are able to witness how and why they receive the compensation that they receive.

4.3 Fitted Formula on Benchmark Problems

We also attempted to test our model from data fitting on benchmark problems to see whether it is consistent. Since the data set is not up to date we regarded the *salary* in the model as *salarylevel* instead which reflecting the relative amount of salary individuals get.

By picking one entry of compensation in the benchmark data and assuming individuals get average 10 years education, we got the expected salary for other entries with the model as shown in **Table 2**. The results indicates our model from data fitting could predict the salary from seniority to some degree but could not predict the exponential increase in salary for the employee with 25 years experience.

To make the model more consistent, we expect fitting the model with more data from the same industry instead of national data would make a difference.

Company1		
Qualif.: Seniority	Actual Compensation	Predicted Compensation
25 yrs	500,000	67980.5
16 yrs	60,000	base case
3 yrs	40,000	36575.5
Company2		
Qualif.: Seniority	Actual Compensation	Predicted Compensation
Rookie	250,000	299105.79
2nd y	30,000	350554.9
3rd y	375,000	base case
4-6th y	450,000	421331.2
7-9th y	655,000	484430.6
10th y	755,000	601808.3

Table 2: Test Model from Data fitting

5 Discussion

5.1 Strengths and Weaknesses

Even though many assumptions were made in order to form a more comprehensive mathematical model, it still gives a strong understanding of the relationship between the qualifications and the compensation received for their work. Some strengths for this model were that the model can implement all types of factors. The model is very flexible, and more variables can be added depending on the employer's preference. Moreover, the model incorporates the weightings for different qualifications depending on the other qualifications as well as the employer's preference. This enables the employer to implement different types of weights like quadratic, exponential, or logarithmic models. We were aware that the qualifications of professionals contain the positive criteria and the negative criteria, and we purposely didn't include factors that would be in the negative criteria because we wanted to make a more just model excluding the unjust

variables. Some examples are, race, gender, marital status, and citizenship status. Even though the majority of the data available with salary distribution are represented with race, gender, marital status, citizenship status, we decided not to include those factors to design a more fair and just model that doesn't include demographic data that shouldn't relate to the qualifications of professionals. Our model also takes into account the cost of living and the starting salary, which leads to a more accurate representation of the relationship between the qualifications of professionals and compensation received for their work.

Some potential problems in this model are that even though we take it into account of cost of living, there is a possibility for an inaccurate implementation of cost of living since it depends on the individual's preferences. Another weakness is that the employers using this model would need some mathematical background knowledge in order to assign weights for different qualifications. Moreover, we assumed that the employers would be able to quantify the candidates' skill sets and then assign weights. However, it might be ambiguous and vague for the employers to quantify the candidates' skill sets. The uncertain approximations for the qualifications may lead to a biased model. Also, the list of assumptions we have for the model may lead to a biased model. We assumed that the starting salary would be based on the minimum wage of Georgia, so that would potentially lead to a biased model. Moreover, we estimated the cost of living to be data collected in Atlanta from the Numbeo website.[1] Therefore, these estimations and assumptions may lead to a biased model.

5.2 Potential Future Research

A possibility for future research is to take into account for different job occupations or different industries. It would be interesting to compare how employers prioritize certain type of qualifications for different industries or job occupations. Another possible future research idea could be to compare pre-covid compensation system to that of post-covid. Even though we assumed that there wouldn't be much of a difference between pre-covid and post-covid, it would be interesting to investigate more deeply on the change in the compensation system. Moreover, a comparison in the compensation system in different states, and implementing the cost of living to the model more accurately would be another potential research idea. Since the data we used was national level data, as well as state-level data in Georgia,

future research could be to conduct research on census-tract level data. A model based on census-tract level data would lead to a more accurate model. Also, a research that further examines on how to make the model to be more user-friendly could be potentially conducted.

6 Appendix

6.1 Model Program

```
import pandas as pd
import re
import numpy as np
import matplotlib.pyplot as plt
from sklearn import linear_model
df = pd.read_excel('dataset.xlsx')
def process_education(x):
    res = 0
    m = re.match(r"[!\d]*([\d]{1,2})th.*grade.*", x)
    if m:
        res = int(m.group(1))
    elif x == " High school graduate":
        res = 12
    elif re.match(r".*Bachelors degree.*", x):
        res = 16
    elif re.match(r".*Some college.*", x):
        res = 14
    elif re.match(r".*Masters.*", x):
        res = 18
    elif re.match(r".*Associates.*", x):
        res = 14
    elif re.match(r".*((Doctorate)|(Prof School)).*", x):
        res = 22
    return res if res else 9

def process_age(x):
    m = re.match(r".*(\d{2}).*", str(x))
    return int(m.group(1))
```

```
df = df[df.wage != 0]
df['education'] = df['education'].apply(process_education)
df['age'] = df['age'].apply(process_age)
df['seniority'] = df.apply(lambda row: row.age - row.education - 5, axis = 1)


x = df["education"].astype('float32')
y = df["wage"].astype('float32')
A = np.vstack([x, np.ones(len(x))]).T
beta, log_alpha = np.linalg.lstsq(A, np.log(y), rcond = None)[0]
alpha = np.exp(log_alpha)
print(f'alpha={alpha}, beta={beta}')

# Let's have a look of the data
plt.plot(x, y, 'b.')
plt.plot(x, alpha*np.exp(beta*x), 'r')
plt.xlabel('education years $edu$')
plt.ylabel('salary level $s$')
plt.title('$s = 354.24e^{0.069edu}$')
plt.savefig('pic1.png')


x = df["seniority"].astype('float32')
y = df["wage"].astype('float32')
A = np.vstack([x, np.ones(len(x))]).T
beta, log_alpha = np.linalg.lstsq(A, np.log(y), rcond = None)[0]
alpha = np.exp(log_alpha)
print(f'alpha={alpha}, beta={beta}')

# Let's have a look of the data
plt.plot(x, y, 'b.')
plt.plot(x, alpha*np.exp(beta*x), 'r')
plt.xlabel('seneority $sen$')
plt.ylabel('salary level $s$')
plt.title('$s = 713.36e^{0.0090sen}$')
plt.savefig('pic2.png')
```

```
regr = linear_model.LinearRegression()
df["seniority2"] = df["seniority"] ** 2
df["education2"] = df["education"] ** 2
regr.fit(df[["seniority", "seniority2", "education", "education2"]], df["wage"])
# s = regr.predict([[40, 16]])

print(regr.intercept_)
print(regr.coef_)

def level(sen, edu):
    return 37.39 * sen - 0.60 * sen ** 2 \
    - 166.16 * edu + 9.38 * edu ** 2 + 1144.39

sen = np.linspace(0, 35, 100)
edu = np.linspace(6, 22, 100)
sen, edu = np.meshgrid(sen, edu)
salary = level(sen, edu)
ax = plt.axes(projection='3d')
ax.plot_surface(sen, edu, salary, cmap='viridis', edgecolor='none')
plt.xlabel('senerity $sen$')
plt.ylabel('education years $edu$')
ax.set_zlabel('salary $s$')
plt.savefig('pic3.png')
```

References

[1] Cost of living in atlanta, 2022.

[2] Cost of living index by state 2022, 2022.

[3] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[4] Sangkwon Lee and John Harris. Managing excellence in usa major league soccer: An analysis of the relationship between player performance and salary, May 2012.

[5] Eileen Patten. Racial, gender wage gaps persist in u.s. despite some progress, Aug 2020.