# The Determinations of Total Compensation: A U.S. Census Study

Andy Xiang
USC ECON 318
Professor Maggie Switek
November 15, 2020

## Abstract

Compensation is not just a paycheck that people receive biweekly but is a major factor in determining people's life choices. This paper attempts to determine the relationship between total compensation and demographic factors such as age, race, gender, education, and region. Salary compensation can serve as a powerful motivating factor to help ensure a strong labor force and high-quality products for the consumer market. The results reveal a positive correlation between age, highest level of education attained, and being Asian or White.

### 1. Motivation

A few months ago, I faced the difficult decision of enrolling into graduate school at USC or starting my full-time job at Ernst & Young (E&Y). I was very grateful to have these choices but felt that either decision would determine my career for at least the next five years. This prompted me to question, which option would result to having the higher yearly compensation and to what extent would other factors affect my salary? Hence, I developed an econometric model that focuses on compensation salary in the United States versus age and other influential variables such as race, gender, education, and region.

## Data Selection & Limitations

### 2. Data Selection

I used 2019 public U.S. Census Data on Current Population Survey (CPS) Annual Social and Economic (ASEC) Supplement. The data focuses on the *person* record type and I renamed many variables from the provided data dictionary to help make the model more understandable (See STATA code in Appendix for variables renamed). Because the data was from the U.S. Census Survey, many important explanatory variables such as tenure-level, years in the work force, or cost of living,

were not collected and thus unobserved. The U.S. Census Data also provided mostly categorical variables. This led to the creation of various dummy variables.

**3. Variable Selection**

I recently only turned 21 years old, so I thought that because of my young age, my salary must be lower given my inexperience. However, I realized that as my parents got older, their salary should start having a diminishing effect as their physical condition may inhibit their ability to work. This led me to the result of using the variable age and age^2.

I also found race to be a critical factor for salary as racial biases can discourage paying employees higher wages. As an Asian-American, I am often stereotyped in being proficient in "math" or quantitative work and wanted to see if my race would have any correlation with salary.

Gender is a crucial variable in the workforce as men may choose different occupations from women or even because of employer's discrimination on gender. I also knew that I could not have attained the offer from E&Y, if I did not have a bachelor's degree. I was offered the opportunity to attend graduate school immediately after my bachelor's degree and wanted to see the relationship between salary and the highest level of degree attained. Thus, I created an interaction term to capture the correlation of a graduate degree on salary given that I am a male.

Another factor that I found important in determining my salary was the cost of living. The cost of living in Los Angeles is relatively high compared to other areas of the United States and I thought my salary would reflect that. Although the data for urbanization or cost of living was not given, I used the variable "region" as a proxy.

**4. Data Limitations**

With any data, there are limitations. Since the data was collected by a U.S. Census, there may be biases in the survey results. Executives who have exponentially high salaries are unlikely to provide their data as their free time may be limited. The highest salary reported was only $160,000. The survey results may also exclude the salaries of visa holders or undocumented immigrants since they may be more fearful of providing their data to the government. When tabulating the results based on citizenship, only 9.28% of full-time workers receiving wages are not U.S. Citizens (See Table 1 in Appendix). Lastly, since this is a data from a survey, the data is subject to common examples of survey biases such that participants are providing their data on a voluntary basis and that there is a limited ability to factually check the data for accuracy.

## Model Analyses

**5. Summary Statistics of Subset Population (Full-time Employed Workers)**

The survey data results consist of a sample size of approximately 180,000. However, I only used a subset of the survey population data focusing on individuals who are full-time employees receiving wages (N=63,278) in order to have a more fair and comparable analysis. This helps exclude subjects who are children, unemployed, part-time employees, or not in the labor force. Log of salary was used provide a more normalized distribution (See Table 2 in Appendix). See summary of statistics of the variables listed below for a more descriptive summary.

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| logsalary | 63,278 | 10.74444 | 0.8267065 | 0.693147 | 14.28551 |
| age | 63,278 | 42.70669 | 12.92624 | 15 | 85 |
| age_square | 63,278 | 92.24283 | 54.761 | 1 | 289 |

| | | | | | |
|---|---|---|---|---|---|
| male | 63,278 | 0.550223 | 0.4974752 | 0 | 1 |
| white | 63,278 | 0.779339 | 0.4146957 | 0 | 1 |
| black | 63,278 | 0.114005 | 0.3178197 | 0 | 1 |
| american_indian | 63,278 | 0.012564 | 0.111382 | 0 | 1 |
| asian | 63,278 | 0.069787 | 0.2547902 | 0 | 1 |
| bachelor | 63,278 | 0.253453 | 0.4349915 | 0 | 1 |
| graduate | 63,278 | 0.112235 | 0.3156577 | 0 | 1 |
| northeast | 63,278 | 0.160245 | 0.3668363 | 0 | 1 |
| midwest | 63,278 | 0.191346 | 0.393364 | 0 | 1 |
| south | 63,278 | 0.371156 | 0.4831179 | 0 | 1 |
| west | 63,278 | 0.275009 | 0.4465222 | 0 | 1 |

## 6. Multilinear Regression

To determine a more thorough understanding on the salary, I developed a multilinear regression model that includes other explanatory and binary variables such as gender, race, education, and region. First, I ran a Breusch-Pagan test to ensure model does not have heteroskedasticity. Since the p-value is less than 0.01, I reject the Ho. Hence, I used a regression with robust standard errors estimated to help mitigate for heteroskedasticity.

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: age age_square 0b.male 1.male 0b.white 1.white 0b.black 1.black 0b.american_indian
                    1.american_indian 0b.asian 1.asian 0b.bachelor 1.bachelor 0b.graduate 1.graduate
                    0b.male#0b.graduate 0b.male#1o.graduate 1o.male#0b.graduate 1.male#1.graduate 0b.northeast
                    1.northeast 0b.midwest 1.midwest 0b.south 1.south 0b.west 1.west

         F(14 , 63263)=     8.28
         Prob > F     =   0.0000
```

Some key results of the multilinear regression show that as a person grows older by 1 year, their salary is expected to increase by ~7%. Males are expected to make ~30% more than females and being white has a positive correlation with salary. A person who has a graduate degree is also expected to have a higher salary. See a more detailed analysis in Section 6.2.

Linear regression

| | |
|---|---|
| Number of obs | = 63,278 |
| F(14, 63263) | = 1050.13 |
| Prob > F | = 0.0000 |
| R-squared | = 0.1994 |
| Root MSE | = .73977 |

| logsalary | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | .0782148 | .001566 | 49.95 | 0.000 | .0751455 | .0812842 |
| age_square | -.0007434 | .0000176 | -42.18 | 0.000 | -.000778 | -.0007089 |
| 1.male | .303425 | .0063788 | 47.57 | 0.000 | .2909225 | .3159275 |
| 1.white | .0672748 | .0201049 | 3.35 | 0.001 | .0278692 | .1066803 |
| 1.black | -.0902538 | .0218317 | -4.13 | 0.000 | -.133044 | -.0474636 |
| 1.american_indian | -.1380936 | .0334276 | -4.13 | 0.000 | -.2036118 | -.0725754 |
| 1.asian | .1391318 | .0231107 | 6.02 | 0.000 | .0938348 | .1844289 |
| 1.bachelor | .4142053 | .0070147 | 59.05 | 0.000 | .4004565 | .4279541 |
| 1.graduate | .6139374 | .0120819 | 50.81 | 0.000 | .5902569 | .6376178 |
| male#graduate | | | | | | |
| 1 1 | .0083107 | .0177814 | 0.47 | 0.640 | -.0265408 | .0431622 |
| 1.northeast | .3262237 | .0771702 | 4.23 | 0.000 | .17497 | .4774775 |
| 1.midwest | .2752634 | .0770556 | 3.57 | 0.000 | .1242344 | .4262924 |
| 1.south | .2347626 | .076954 | 3.05 | 0.002 | .0839326 | .3855927 |
| 1.west | .2683993 | .0769898 | 3.49 | 0.000 | .1174992 | .4192993 |
| _cons | 8.226995 | .0856452 | 96.06 | 0.000 | 8.05913 | 8.394859 |

(Predicted) logsalary = 8.227+ 0.0782age - 0.0074age^2 + 0.3034male + 0.06727white - 0.0903black – 0.1381american_indian + 0.1391asian + 0.4142bachelor + 0.6139graduate + 0.0083male*graduate + 0.3262northeast + 0.2752midwest + 0.2347south + 0.2684west + u

Although the R^2 = 0.1994, meaning that only 19.94% of the independent variables in the multilinear regression model represent the variation in logsalary, the

independent variables (except for male*graduate) are still statistically significant. This means that a large proportion of the variation are from other explanatory variables that are not captured in the model. Given the constraints with the U.S. Census Data, I believe if I had the other data such as years in the workforce, years of education, or job performance, then the variance explained by the independent variables can be higher.

*6.1 F-Test for Overall Significance*

```
 ( 1)  age = 0
 ( 2)  age_square = 0
 ( 3)  1.male = 0
 ( 4)  1.white = 0
 ( 5)  1.black = 0
 ( 6)  1.american_indian = 0
 ( 7)  1.asian = 0
 ( 8)  1.bachelor = 0
 ( 9)  1.graduate = 0
 (10)  1.male#1.graduate = 0
 (11)  1.northeast = 0
 (12)  1.midwest = 0
 (13)  1.south = 0
 (14)  1.west = 0

       F( 14, 63263) = 1050.13
            Prob > F =    0.0000
```

The F-Test calculates a statistic of 1050.13, making this model statistically significant.

*6.2 Detailed Analysis of Explanatory Variables*

The multilinear regression model shows that as a person grows older by 1 year, their expected salary increases by ~7%. This is expected as many firms may annually adjust for a ~2-3% increase for inflation as well as reward job performance and retain retention of employees.

A dummy variable for gender was created where male = 1 and female = 0. The results indicate that males tend to earn ~30.343% higher salaries than females with a t-statistic of 47.57, meaning this finding is statistically significant at the 1% level.

Dummy variables for race were categorized into White, Black, American Indian, and Asian. All other races such as mixed races were omitted as they served as the benchmark group (See data dictionary for full list of races). Asians, surprisingly, have the highest coefficient, meaning Asians are expected to make ~13.913% more than the benchmark group. Asians, however, only accounted for ~7% of full-time employed workers in this subset population data. The second highest coefficient for race was White as White people are expected to make ~6.727% more than the benchmark group. African Americans and Indian Americans have a negative correlation with salary (See coefficients in Multilinear Regression Table).

Dummy variables for the highest level of education were categorized into high school, some college, bachelor's degree, and graduate degree. High school and "some college" variables were omitted because of collinearity and served as the benchmark group. The results indicate that having a graduate degree (41.421%) is most positively correlated than having just a bachelor's degree (61.394%) compared to those who do not hold the minimum of a bachelor's degree. Given this result, it is worth considering attending graduate school in hopes of achieving a higher salary as I may be able to provide more value in the labor force.

Since the data such as urbanization or cost of living was not given, I created dummy variables based on region, which were categorized into Northeast, Midwest, South, and West in the United States. The omitted group were salaries abroad. Because I will be working in the West Coast in the United States, I am expected to earn approximately 26.84% higher than people working outside of the United States. In general, people who work in the United States are expected to make between ~23-32% higher salaries than workers abroad. Workers who also live in the Northeast tend to have higher salaries since the cost of living in areas such as New York, New Jersey, or Washington D.C. are higher. The results for all explanatory variables (except for male*graduate) above are statistically significant at the 1% level.

*6.3 Interaction Term*

Male*graduate

Since I was offered the opportunity to attain a graduate degree, I wanted to see the interaction between my gender and receiving a graduate degree on salary. The results show that having a graduate degree given that a person is male increases expected salary by approximately 0.831%. However, since the t-value is 0.47, which is very small, this conclusion is statistically insignificant.

*6.4 Partial Effect on Salary of Age and Age^2*

0 = 0.07821-2*0.00743age

To determine at which age results to the highest earnings, I took the first derivative while holding all other variables other than age constant. The results show that on average, a person reaches their highest salary at around 53 (52.6) years old. In some cases, as a person becomes older, their physical ability starts deteriorating which can help explain the diminishing effect of age.

## Conclusion

Age is positively correlated with salary; however, age starts to have a diminishing effect on salary after the age of ~53 years old. I have determined that gender, race, educational degree, and the region of work are all significantly correlated with salary. Workers with a graduate degree are also expected to make more than those who hold only a bachelor's degree. These results largely appear to be consistent with findings from other research studies. Since this is also an observational study, I can only conclude correlation rather than causation. Moving forward, an inclusion of additional explanatory variables (i.e. years in the work force, tenure level, etc.) would further strengthen the model.

## Appendix

**Table 1: Summary Statistics of Summary**

| Citizenship | Freq. | Percent | Cum. |
|---|---|---|---|
| Native, born in US | 50,871 | 80.39 | 80.39 |
| Native, born in PR or US outlying area | 409 | 0.65 | 81.04 |
| Native, born abroad of US parent(s) | 589 | 0.93 | 81.97 |
| Foreign born, US cit by naturalization | 5,539 | 8.75 | 90.72 |
| Foreign born, not a US citizen | 5,870 | 9.28 | 100 |
| **Total** | **63,278** | **100** | |

**Table 2: Histogram of Logsalary**

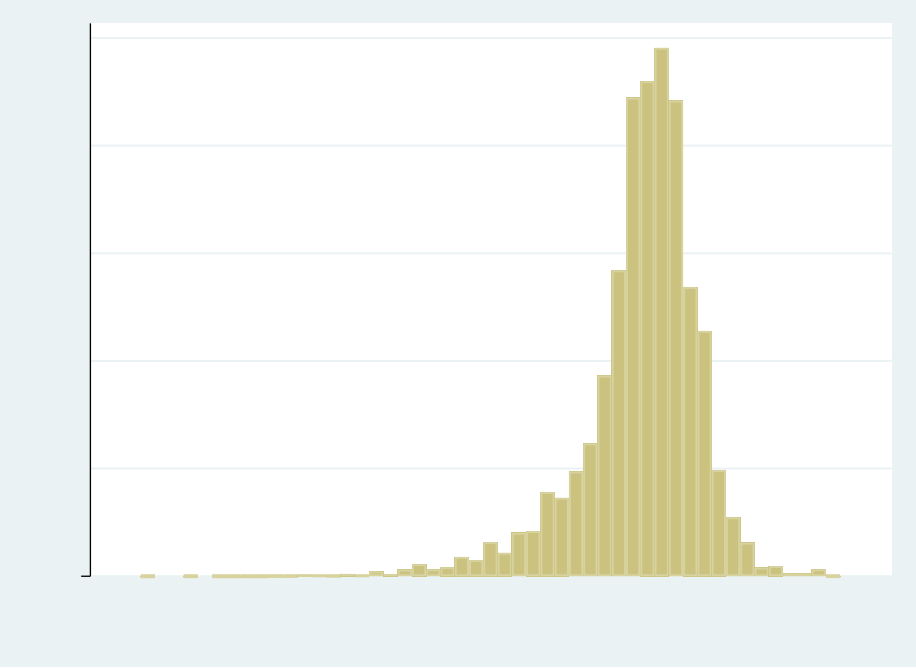

**Table 3: Simple Linear Regression**

(Predicted) Logsalary = 10.1958+0.01285age + u

Linear regression

| | |
|---|---|
| Number of obs | = 63,278 |
| F(1, 63276) | = 2339.56 |
| Prob > F | = 0.0000 |
| R-squared | = 0.0404 |
| Root MSE | = .80986 |

| logsalary | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | .0128473 | .0002656 | 48.37 | 0.000 | .0123267 | .0133679 |
| _cons | 10.19578 | .0117848 | 865.16 | 0.000 | 10.17268 | 10.21887 |

## STATA Code

```
*gen female = 1 if a_sex == 2
*replace female = 0 if a_sex == 1

*gen age_square = age^2

*rename pemlr employed
*rename a_wkstat fulltime
*rename a_age age
*rename a_sex male
*rename wsal_val salary

*gen race = 1 if prdtrace ==1
*replace race = 2 if prdtrace ==2
*replace race = 3 if prdtrace ==3
*replace race = 4 if prdtrace ==4
*replace race = 5 if prdtrace >=5

*tab race, gen(m)
*rename m1 white
*rename m2 black
*rename m3 american_indian
*rename m4 asian
*rename m5 other_race (base)

*gen educ = 1 if a_hga <=39
*replace educ =2 if a_hga >=40 <=42
*replace educ =3 if a_hga == 43
*replace educ = 4 if a_hga == 44

*tab educ, gen(m)
*rename m1 high_school
*rename m2 some_college
*rename m3 bachelor
*rename m4 graduate

*gen logsalary = log(salary)

*gen female = 1 if a_sex == 2
*replace female = 0 if a_sex == 1

*gen male = 1 if a_sex == 1
*replace male = 0 if a_sex == 2

*rename wsal_yn receive_wage

*rename mig_reg region
*tab region, gen(m)
*rename m1 no_data
```

```
*rename m2 black
*rename m3 american_indian
*rename m4 asian
*rename m5 other_race (base)

*gen educ = 1 if a_hga <=39
*replace educ =2 if a_hga >=40 <=42
*replace educ =3 if a_hga == 43
*replace educ = 4 if a_hga == 44

*tab educ, gen(m)
*rename m1 high_school
*rename m2 some_college
*rename m3 bachelor
*rename m4 graduate

*gen logsalary = log(salary)

*gen female = 1 if a_sex == 2
*replace female = 0 if a_sex == 1

*gen male = 1 if a_sex == 1
*replace male = 0 if a_sex == 2

*rename wsal_yn receive_wage

*rename mig_reg region
*tab region, gen(m)
*rename m1 no_data
*rename m2 northeast
*rename m3 midwest
*rename m4 south
*rename m5 west
*rename m6 abroad


sum logsalary age age_square male white black american_indian asian bachelor graduate northeast midwest south west if employed == 1 & fulltime==2 & receive_wage==1

reg logsalary age age_square i.male i.white i.black i.american_indian i.asian i.bachelor i.graduate i.male#i.graduate  i.northeast i.midwest i.south i.west if employed == 1 & fulltime==2 &
receive_wage==1, robust

test age=age_square=1.male=1.white=1.black=1.american_indian=1.asian=1.bachelor=1.graduate=1.male#1.graduate=1.northeast=1.midwest=1.south=1.west

reg logsalary age if employed == 1 & fulltime==2 & receive_wage==1, robust

tab prcitshp if employed == 1 & fulltime==2 & receive_wage==1
```