AUAF AMERICAN UNIVERSITY of AFGHANISTAN

# MULTI-REGRESSION ANALYSIS

Muhammad Usman Popal

**ITC 255:** Statistical Data Analysis

Dr. Asadullah Jawid

December 25, 2022

## [PROJECT QUESTION ANALYSIS]

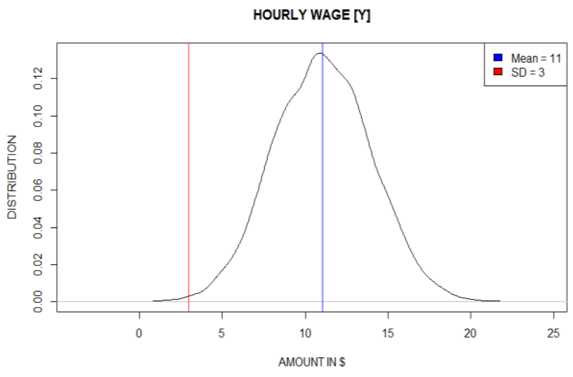
The **Y** variable I have been given to create data with represents the hourly wage of employees, and it should have a normal distribution with a mean that equals **11** and a standard deviation of **3.**

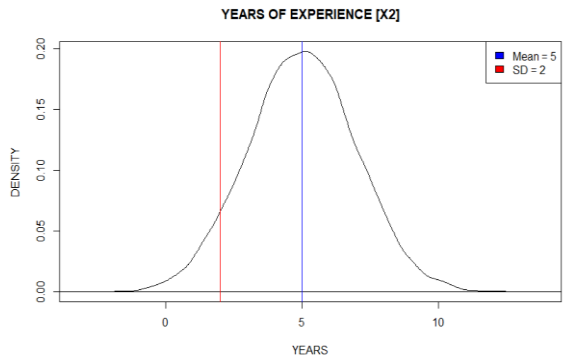
**X1** is a binomial variable that tells whether or not the observation at hand has a postgraduate degree with a probability of **40%** postgraduates.

**X2** is again a normal distribution containing the years of experience of employees under observation, and it should have a mean of **5** and a standard deviation of **2**.

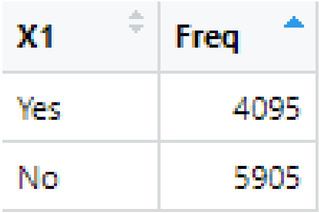
## [DATA CREATION VIA R]

The data I create for both numerical variables is through the "rnorm" function of RStudio. Following are the graphs that conclude **10000** observations over the given numerical variables (**Y** & **X2)** representing a normal distribution and with their respective means and standard deviations as per the requirement.





X1 (the 40% probability of postgraduate employees) is also done for **10000** observations via the random sampling function while specifying a probability of **0.4**.
(4095/10000 = 0.4095)



| X1 | Freq |
|---|---|
| Yes | 4095 |
| No | 5905 |

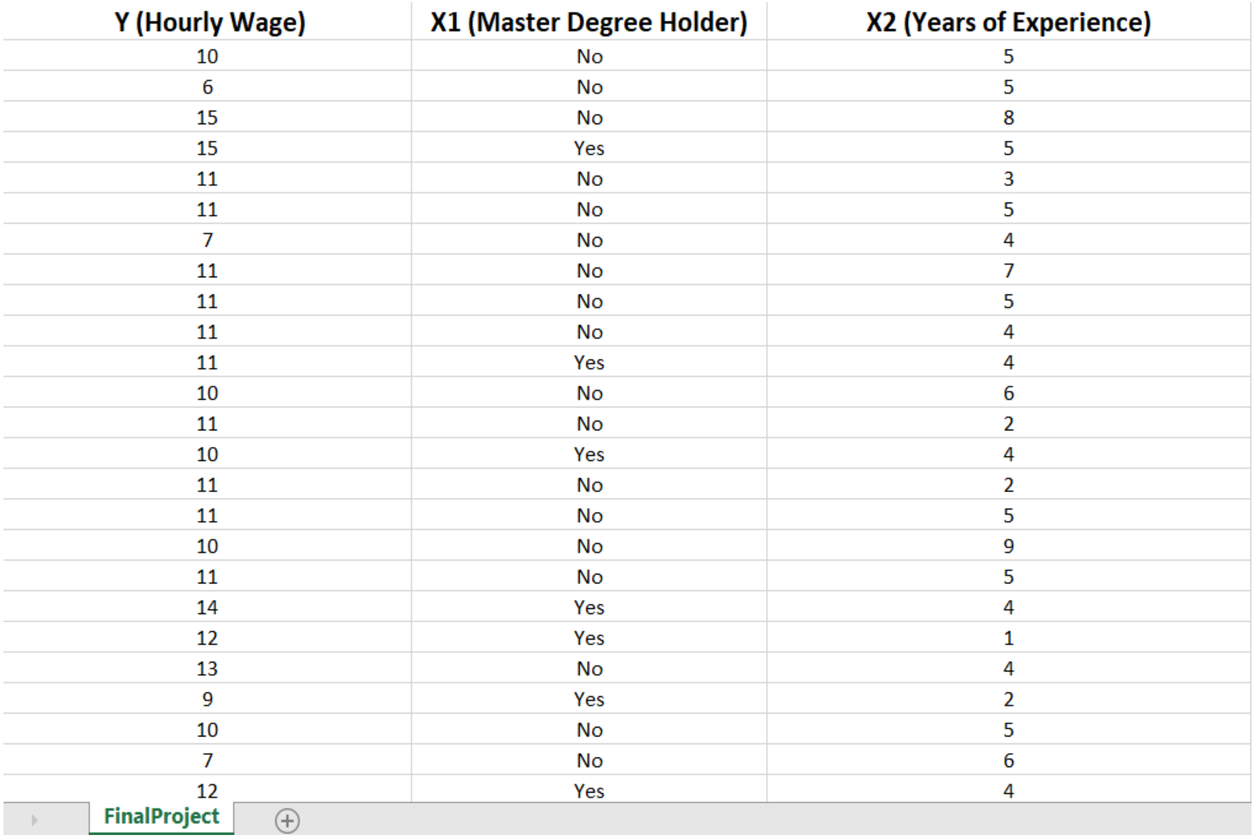
# [DATA CLEANING]

The next step is to organize the data and put it into an excel sheet so it is easier to analyze. The **binomial variable (X1)** is almost 40% but leaning towards approximately 41% which seems about right since the ‘probability’ of employees having a postgraduate was set accordingly.

However, hourly wage (**Y**) as well as years of experience (**X2**) of an employee are usually discrete variables. In order to fix the decimal points that are given by the ‘rnorm’ function in R, we simply use another function called ‘round’ in order to round-off the decimal points to the nearest number. Similarly, when I checked the range of both variables, they started from a negative number which I also fixed for the whole data.

**NOTE:** **[**The standard deviation and mean (double-checked) remain almost the same after rounding off the numerical data.**]**

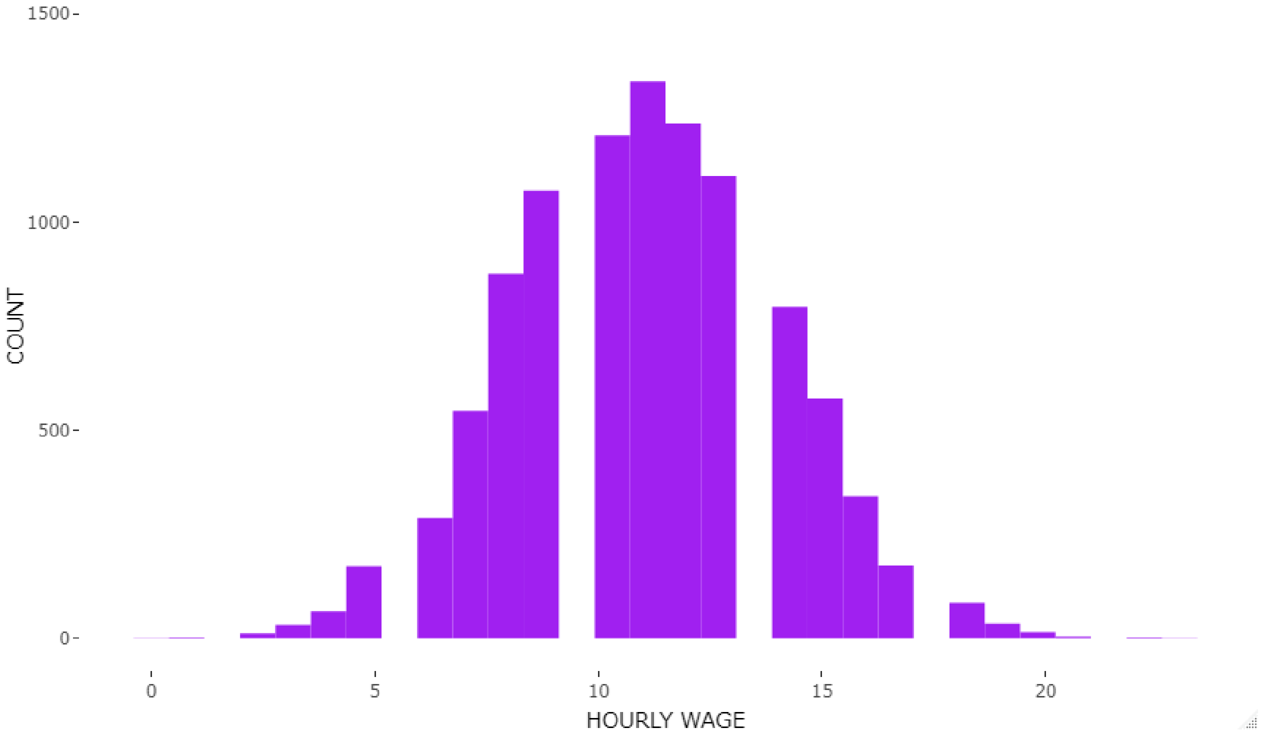
Finally, the data created, collected, and cleaned takes a following shape in an excel file.



| Y (Hourly Wage) | X1 (Master Degree Holder) | X2 (Years of Experience) |
|---|---|---|
| 10 | No | 5 |
| 6 | No | 5 |
| 15 | No | 8 |
| 15 | Yes | 5 |
| 11 | No | 3 |
| 11 | No | 5 |
| 7 | No | 4 |
| 11 | No | 7 |
| 11 | No | 5 |
| 11 | No | 4 |
| 11 | Yes | 4 |
| 10 | No | 6 |
| 11 | No | 2 |
| 10 | Yes | 4 |
| 11 | No | 2 |
| 11 | No | 5 |
| 10 | No | 9 |
| 11 | No | 5 |
| 14 | Yes | 4 |
| 12 | Yes | 1 |
| 13 | No | 4 |
| 9 | Yes | 2 |
| 10 | No | 5 |
| 7 | No | 6 |
| 12 | Yes | 4 |

# [GRAPHICAL ANALYSIS OF THE VARIABLES]

## [Y: HOURLY WAGE]



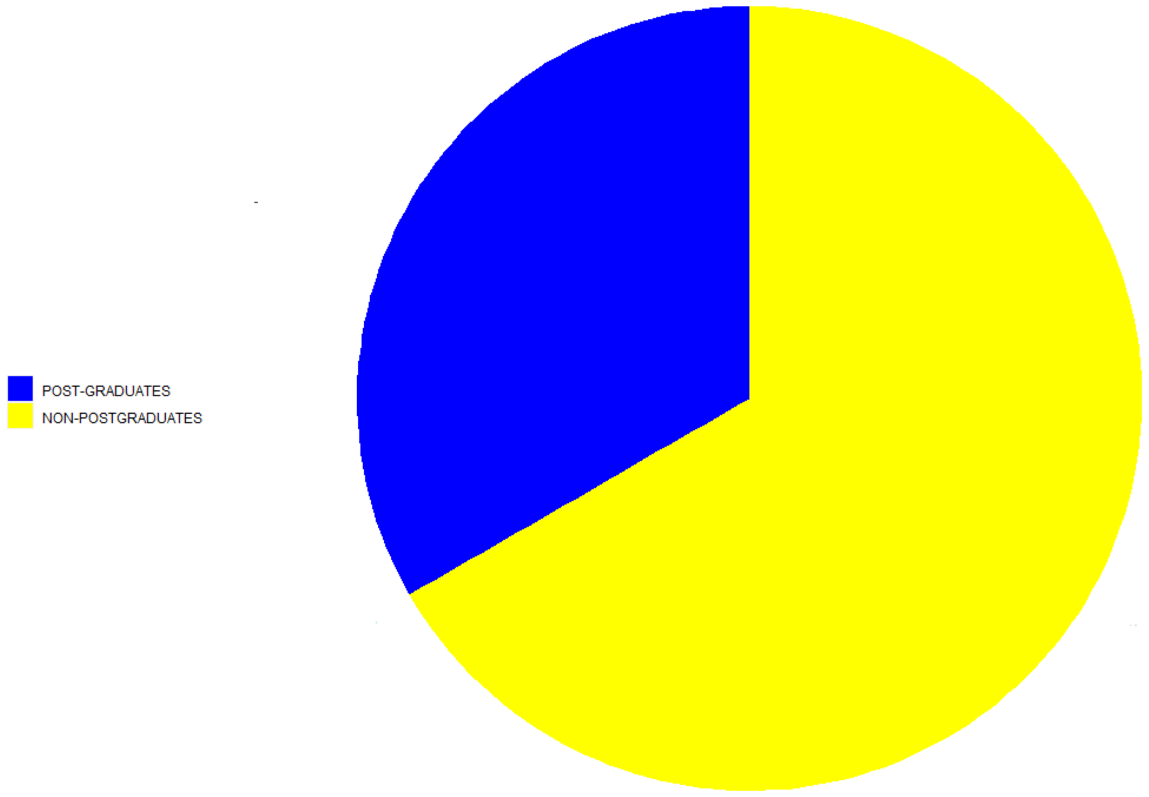
Numerous analyses can be made from the histogram above.

More than 5000 employees obtain an hourly wage between 10$ and 13$ an hour, the highest of which are 1338 employees that receive 11$ an hour (which makes sense because the mean of our data set is also 11, and that is where most of the observations in the graph lie).

There are 12 (lowest earning) employees that receive 2$ and 4 (highest earning) employees that earn 21$ an hour if the low and high ends of the histogram are observed respectively.

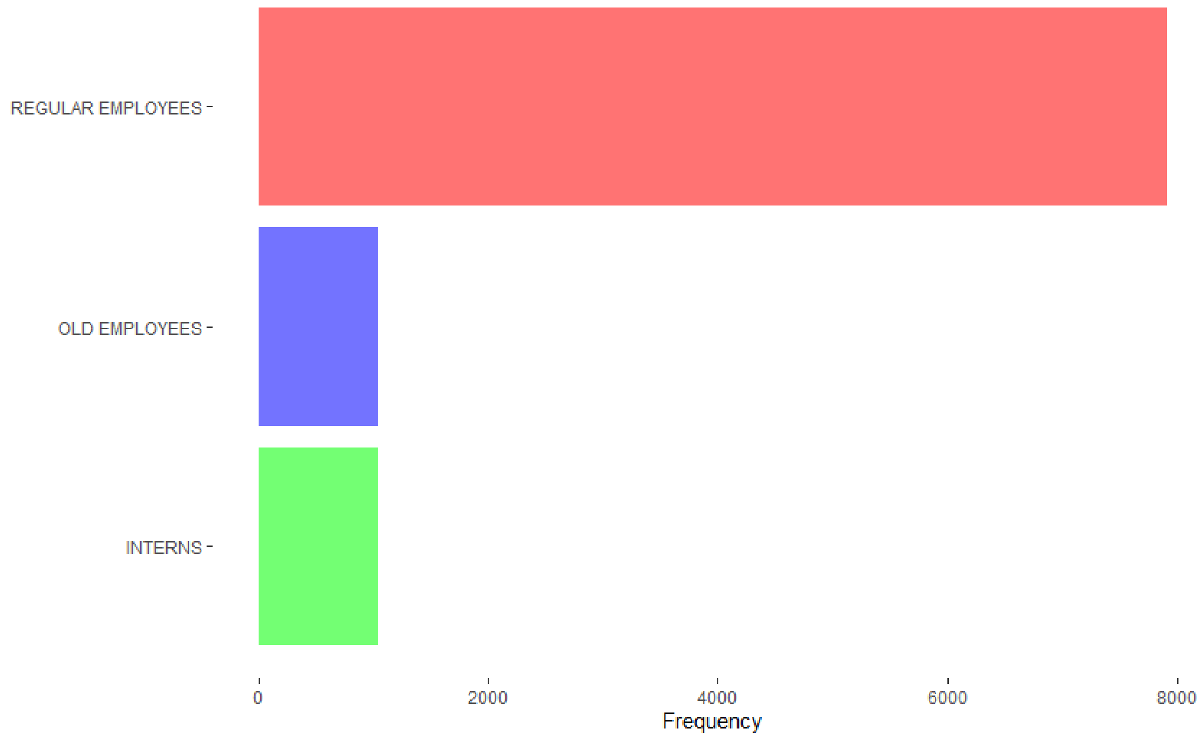
An hourly wage of 6$, 9$, 14$ and 16$ is not received by any employee in the company, visible by the gaps in the histogram.

# [X1: POSTGRADUATE DEGREE HOLDING EMPLOYEES]



The Pie Chart above shows a simple graphic representation of the probable 40% postgraduate degree holders, which means that among the 10000 employees, **4095** possess a Master's degree and the rest of the **5905** employees do not.

# [X2: EMPLOYEE EXPERIENCE STATUS]



An analysis of the bar graph above tells that employees of the company are divided on the base of their time inside the company.

1) The employees that have spent less than 3 years in the company are considered **interns**.
2) The employees that have spent more than 3 but less than 7 years are considered **regular employees**.
3) The employees that have spent more than or equal to 7 years are considered as **old employees**.

Based on the above distribution, the highest number of employees that are inside the company are regular ones, with an extremely high frequency of **7903**, which is about 79.03% of the total employees. The interns and old employees are almost equal, as the interns are **1052** (10.52%) and old employees are **1045** (10.45%) precisely, which might mean that the company might have a good package for their employees, since there are over a thousand employees that have worked for more than 7 years, and there are over 7000 employees working for more than 3 years.

# [MULTI-REGRESSION]

**Y = *a + b1*X1 + b2*X2***

Following are the values extracted by the linear model formula in RStudio.

*a* = 11.079198

*b1* = -0.013548

*b2* = -0.005001

The H0 or null hypothesis for this case is that there is no effect of having a post-graduate degree or experience in the company on your hourly wage (as in the amount of time the employee has to work.

The H1 is two-sided, such that it either has an indirect effect on Y or direct effect (either having a postgraduate degree and higher amount of experience results in a higher hourly wage or vice versa).

(The p-value for both X1 (**0.822**) and X2 (**0.737**) is much greater than 0.1 [which is my specified limit for error (**alpha = 0.1 or 10%**)] which means we have to reject H0, since H1 (having a negative effect of X on Y) is true.

Similarly, both the b-hats are negative, which means there is an indirect effect of X on Y.

Interpretation can be done as the more employees have worked in the company; the less likely they are to work as enthusiastically as an intern would, decreasing the hourly age. Similarly, employees that have a postgraduate degree are less likely to be able to work more hours in order to get a high wage, as the time to get a higher salary decreases by less working hours, resulting in a lower hourly wage.

Putting the values in the formula.

**Y = 11.079198 -0.013548 * X1 -0.005001*X2**

We can simulate an example as follows:

Y = 11.079198 -0.013548 * 1 -0.005001*5 (1 being a postgraduate) = 11.04064

Meaning that a regular employee (5 years of experience) who has a masters degree earns about 11$ per hour.

Y = 11.079198 -0.013548 * 0 -0.005001*2 = 11.0692

Meaning that an intern who does not have a master's degree earns 2 cents more.