CPS1985 Dataset Analysis

Julieta Pappano

2022-Jan-01

Dataset

CPS1985 dataset contains 534 observations from the Current Population Survey data collected in 1985 by the US Census Bureau, with information on wages and other characteristics of workers. Table 1 shows the descriptive statistics of the numeric variables contained in this dataset.

Table 1: Summary of Numeric Variables (CPS1985)

	wage	education	experience	age
	Min. : 1.000	Min. : 2.00	Min. : 0.00	Min. :18.00
	1st Qu.: 5.250	1st Qu.:12.00	1st Qu.: 8.00	1st Qu.:28.00
	Median : 7.780	Median :12.00	Median :15.00	Median :35.00
	Mean : 9.024	Mean :13.02	Mean :17.82	Mean :36.83
	3rd Qu.:11.250	3rd Qu.:15.00	3rd Qu.:26.00	3rd Qu.:44.00
	Max. :44.500	Max. :18.00	Max. :55.00	Max. :64.00

Table 2: Summary of Categorical Variables (CPS1985)

	ethnicity	region	gender	occupation	sector	union	married
	cauc :440	south:156	male :289	worker :156	manufacturing: 99	no :438	no :184
	hispanic: 27	other:378	female:245	technical :105	construction : 24	yes: 96	yes:350
	other : 67	NA	NA	services : 83	other :411	NA	NA
	NA	NA	NA	office : 97	NA	NA	NA
	NA	NA	NA	sales : 38	NA	NA	NA
	NA	NA	NA	management: 55	NA	NA	NA

Visualization: Histograms and Boxplots

Figure 1 corresponds to the age variable. The graph has bins with a width of five years and shows that the distribution is skewed to the right, which means that the sample consists mainly of younger people. Most observations are found between 28 and 38 years old. On the other hand, the boxplot on the right allows analysing the spread and centres of the data. For the case of age, it can be seen that the median (Q2) of the observations is around 35 years of age. Furthermore, 25% (Q1) of the observations can be found between 18 to 28 years old, 50% of the observations (IQR = Q3-Q1) fall between 28 and 43 years old and 25% (above Q3) between 44 and 68 years. Lastly, the boxplot confirms the skewness to the right as well.

Figure 1. Age Distribution

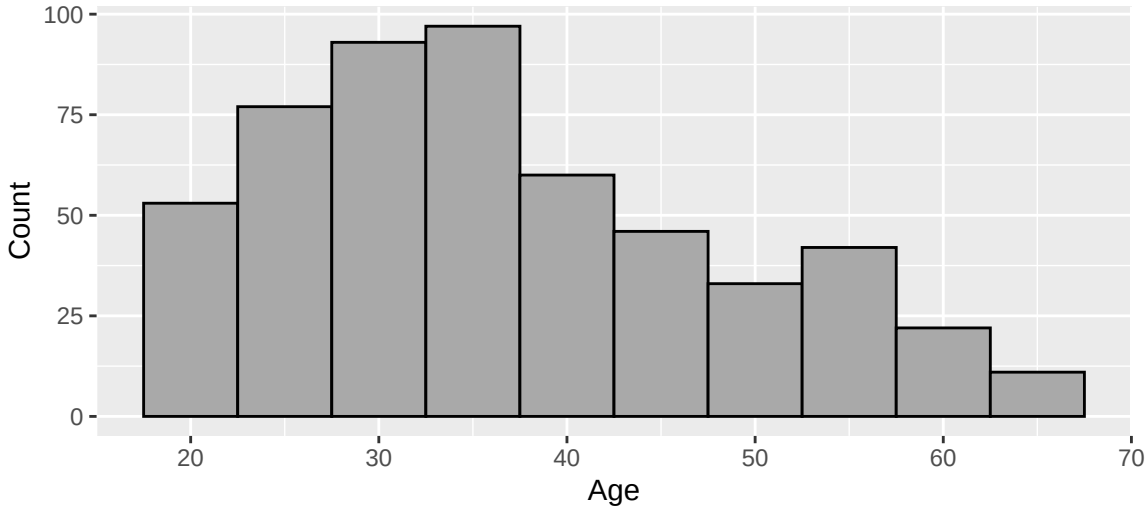
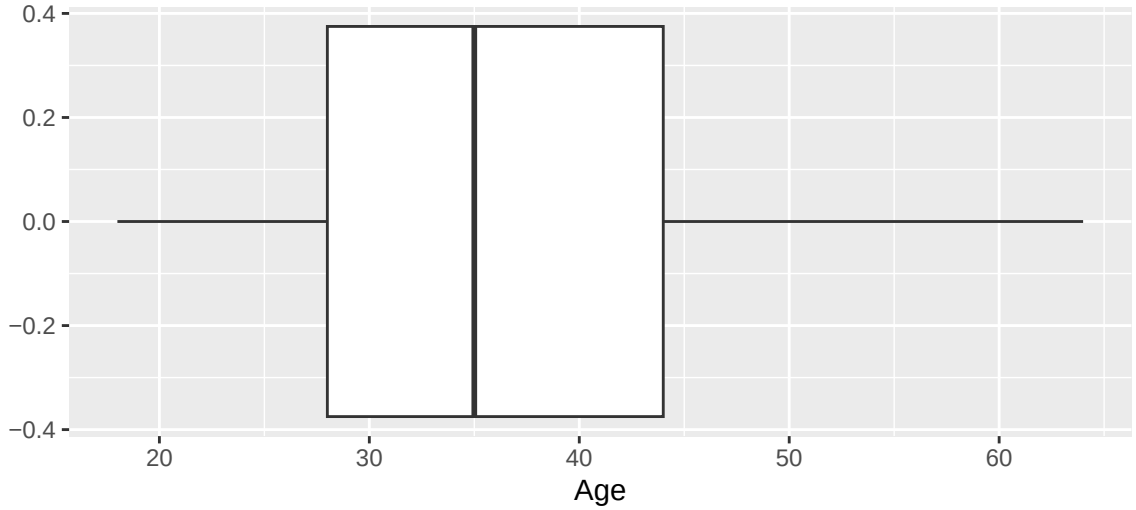


Figure 2. Age Spread



In the case of experience, bins also have a width of five years. The Figure 3 shows that the distribution is skewed to the right and that most observations can be found in the range from approximately 12 to 17 years of experience. On the other hand, the boxplot shows that the median (Q2) is around 15 years of experience. Furthermore, 25% (Q1) of the observations have between 0 to 8 years of experience, 50% of the observations (IQR = Q3-Q1) between 12 and 26 years and the upper 25% (above Q3) between 26 and 48 years. Lastly, the boxplot shows two outliers that are above 50 years and it also confirms the skewness to the right.

Figure 3. Experience Distribution

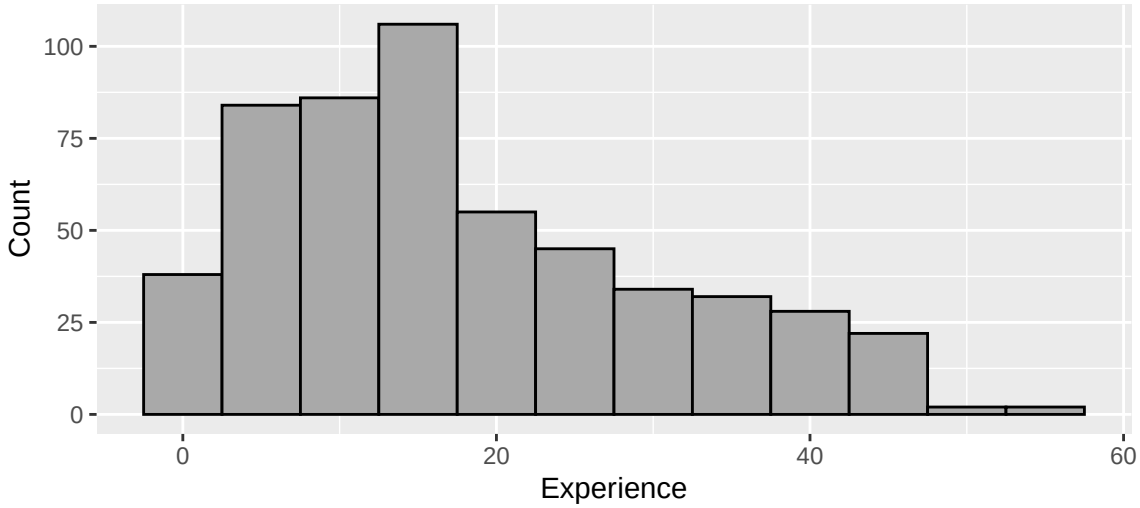


Figure 4. Experience Spread

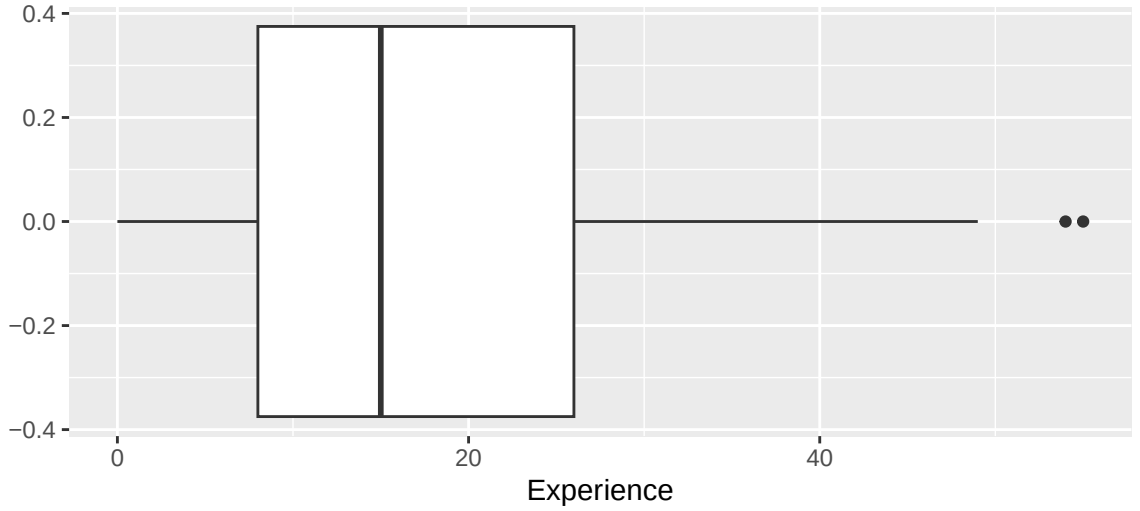


Figure 5 shows the distribution of the wage variable. The histogram bins have a width of three dollars per hour and it shows that the distribution is skewed to the right, with most observations found in the range from approximately 4.5 to 7.7 dollars per hour. On the other hand, the boxplot in Figure 6 shows that the median (Q2) is around 7 dollars per hour. Furthermore, 25% (Q1) of the observations have a wage that ranges approximately from 1 to 6 dollars per hour, 50% of the observations (IQR = Q3-Q1) are between 11 and 6 dollars p/h and the upper 25% (above Q3) are between 26 and 48 dollars p/h. The boxplot shows several outliers that are above 20 dollars p/h and also confirms the skewness to the right.

Figure 5. Wage Distribution

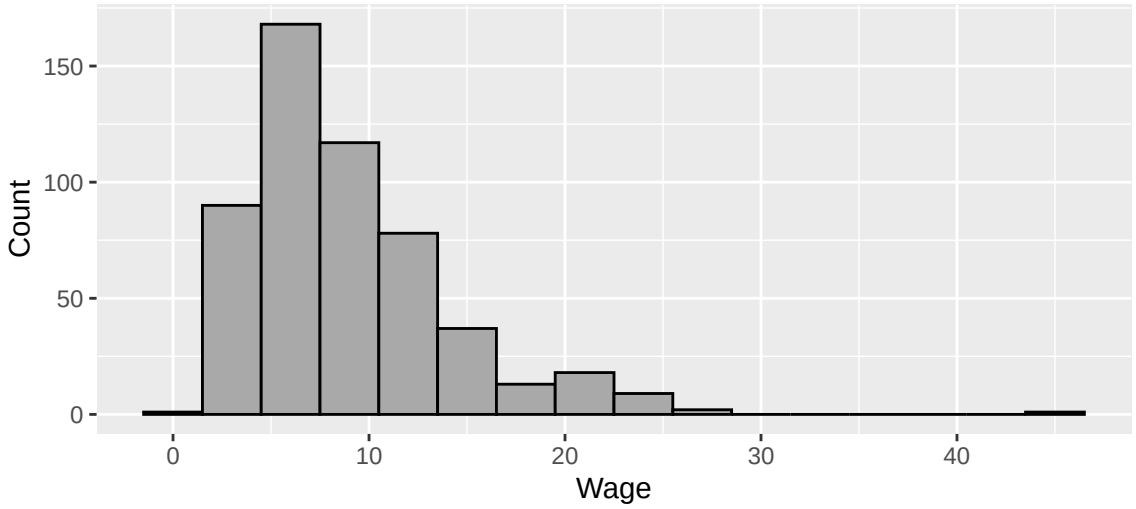
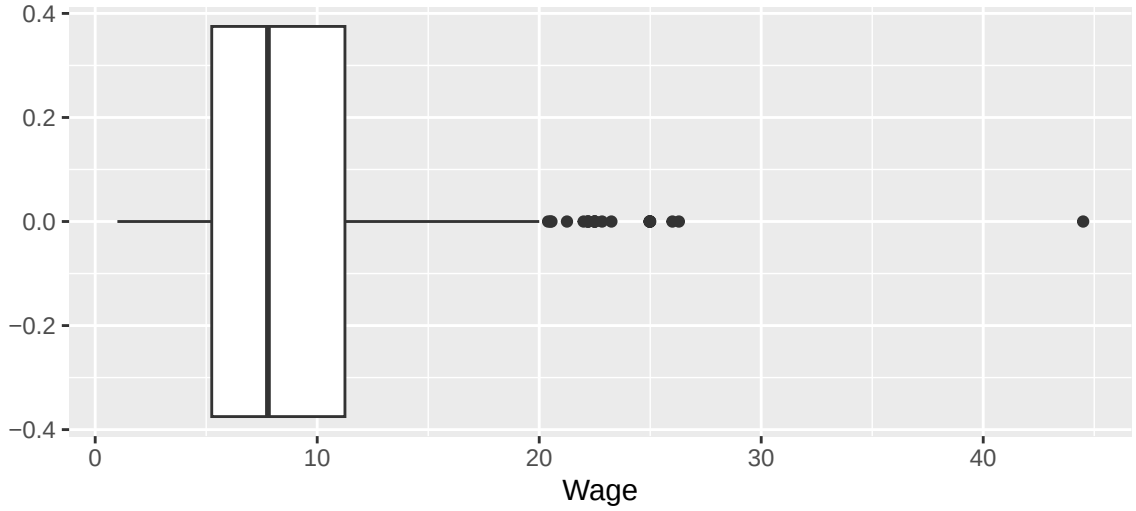


Figure 6. Wage Spread



In the case of education, bins have a width of five years. The graph (Figure 7) shows that the distribution is moderately skewed to the left and that most observations can be found in the range from approximately 11 to 13 years of education. On the other hand, the boxplot (Figure 8) shows that the median (Q2) is around 11 years of education and coincides with Q1. This could be explained by the several outliers located on the left side of the graph, below 8 years of education, which distort the distribution of observations. Furthermore, 50% of the observations (IQR = Q3-Q1) between 11 and 15 years and the upper 25% (above Q3) between 15 and 18 years.

Figure 7. Education Distribution

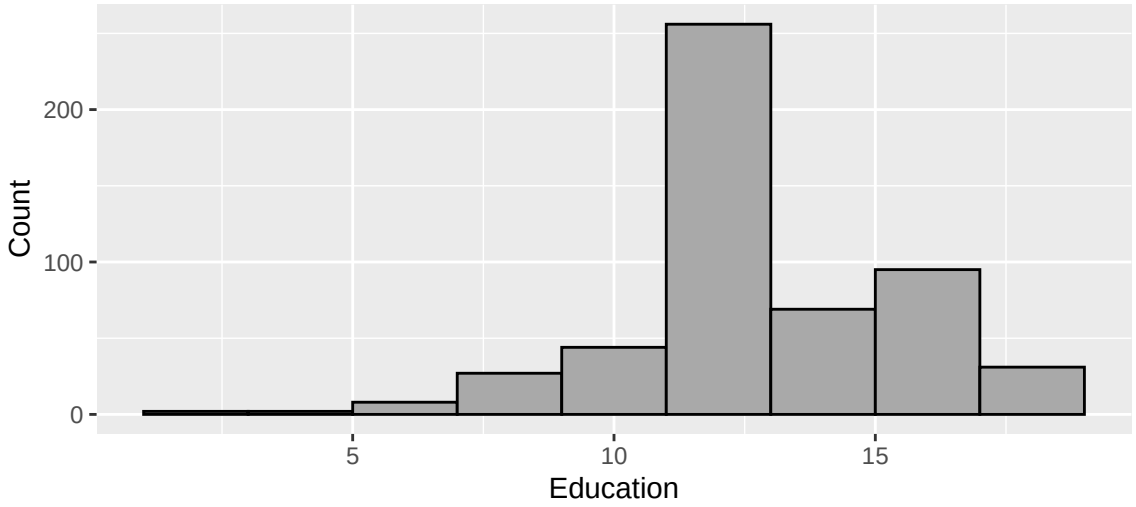
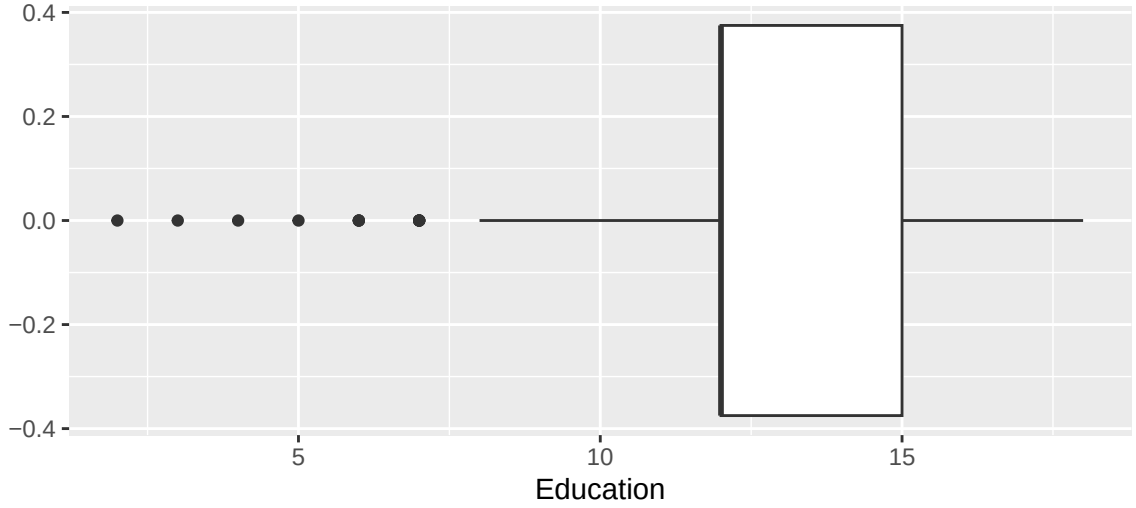
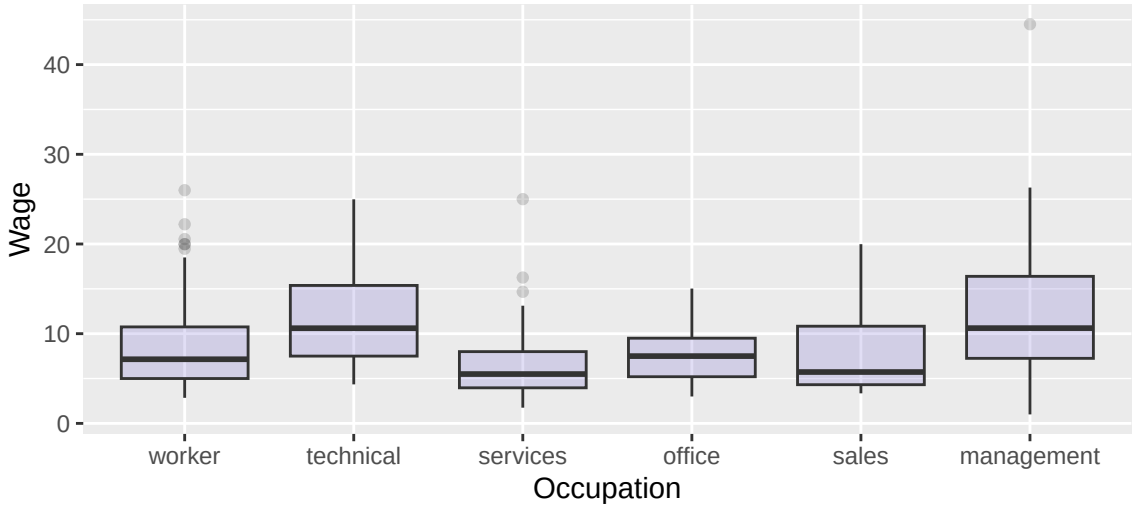


Figure 8. Education Spread



Lastly, Figure 9 shows the distribution of wages across occupation categories. As it can be seen, the highest wages can be found in the management category, which is the one with the highest values since 50% of observations (above the median) have salaries between approximately 11 and 27 dollars per hour. However, it should also be noted that this category is the one that presents the biggest dispersion of data. The next category with the highest wages is the technical category, in which 50% of observations (above the median) have wages that range from 11 to 25 dollars per hour. On the other hand, the category with the lowest wages is services. In this case, the median salary is approximately 6 dollars per hour and the overall dispersion of the category goes from approximately 2.5 to 14 dollars per hour. Lastly, worker is the category with the highest number of outliers (5 outliers), followed by services (3 outliers) and management (1 outlier).

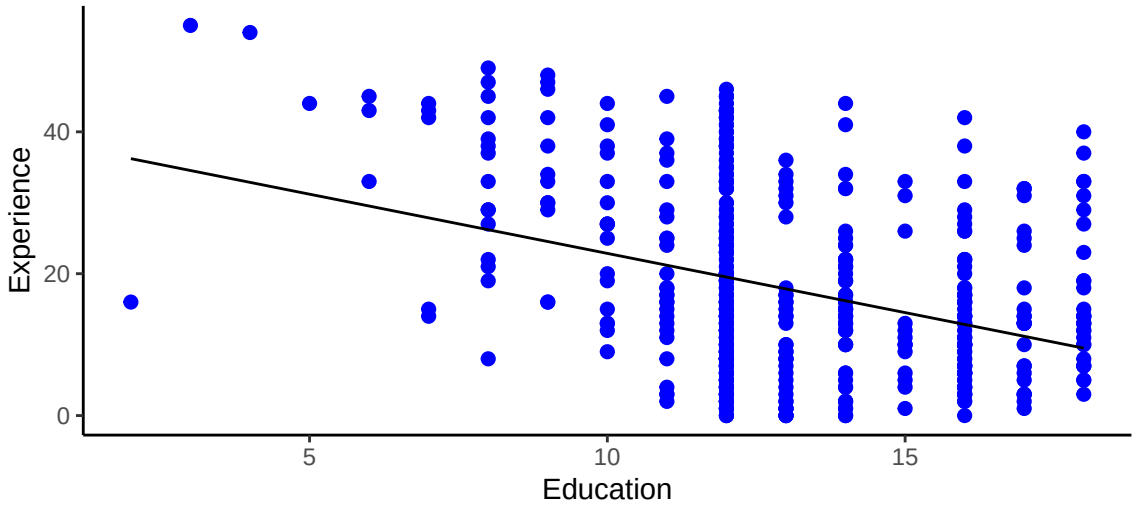
Figure 9. Data distribution per Occupation category



Visualizations: Scatterpolots

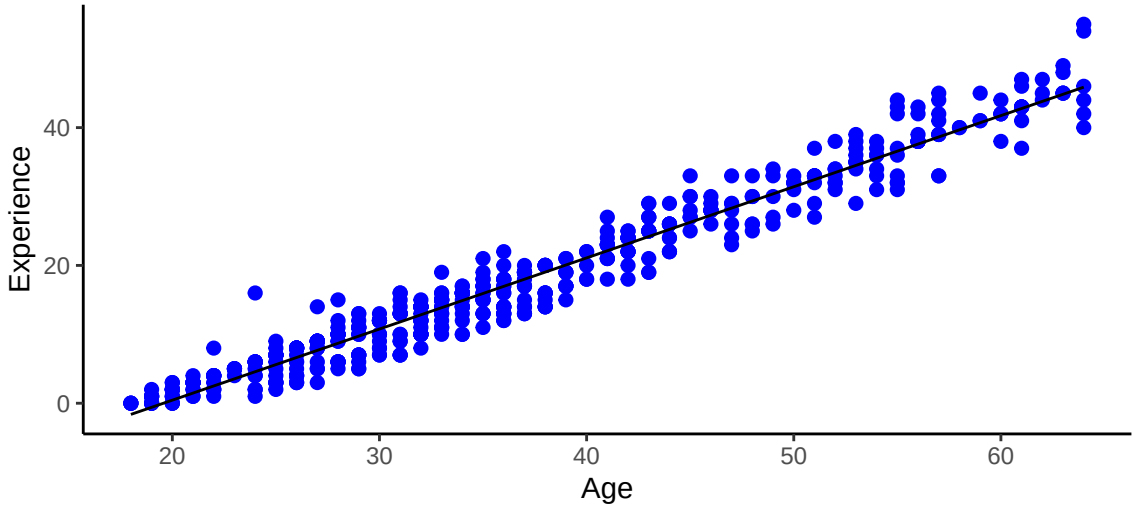
In the first case, considering the trend line, experience and education seem to be negatively associated. This is counterintuitive since this negative relation would imply that the higher the education, the lower the experience. Furthermore, the observations in the scatterplot do not seem to support the said trend. For instance, the observations are very far from the line and are vertically aligned, which shows that the trend line is not correctly representing the relationship between the two variables.

Figure 10. Relation between Experience and Education



On the other hand, when education is replaced by age, the trend line shows a positive relationship, which agrees with the general notion that the older a person is, the more experience that person has. Furthermore, the observations are closer to the line which means that there could be a positive association between both variables.

Figure 11. Relation between Experience and Age



Regression Models

Three simple models are analyzed to understand how age is affected by education, age, and a combination of both. 1) Short model, in which the the effect of education on wage is analyzed 2) Short model, in which the effect of age on wage are analyzed 3) Long mode, in which the effect of education and age on wage are analyzed.

Table 3: Regression model comparison

	Dependent variable:		
	Education	Wage	Education and Age
	Age		
	(1)	(2)	(3)
Education	0.750***		0.821***
	(0.079)		(0.077)
Age (Years)		0.078***	0.105***
		(0.019)	(0.017)
Intercept	−0.746	6.167***	−5.534***
	(1.045)	(0.723)	(1.279)
Observations	534	534	534
R^2	0.146	0.031	0.202
Adjusted R^2	0.144	0.029	0.199
Residual Std. Error	4.754 (df = 532)	5.063 (df = 532)	4.599 (df = 531)
F Statistic	90.852*** (df = 1; 532)	17.199*** (df = 1; 532)	67.210*** (df = 2; 531)
Note:	*p<0.1; **p<0.05; ***p<0.01		

However, this model may be affected by omitted variable bias (OVB). For instance, as a person ages they are more likely to have more years of schooling. Furthermore, age could also be related to higher earnings since

generally people who are more experienced in the job market are better prepared, not only in terms of ability but also in relation to their knowledge of the labour market and salary negotiation strategies. As an example, a forty-year-old probably has more years of schooling, is better prepared for an interview, more aware of the wage offered in the market and already knows which companies pay better than a twenty-year-old. This would have a positive effect on both variables: education and earnings.

To calculate the total OVB, the effect of age on education and wage should be taken into consideration. The effect of age on wage was previously calculated in the second regression model outlined above. In Table 1, it can be seen that the effect of age on wage in the more complete model (3) is 0.105. To obtain the total OVB, a regression is conducted to analyze the effect of age on education, as shown in Table 2.

Table 4: Regression OV

	Dependent variable:
	Education
Age (Years)	-0.673^{***}
	(0.192)
Intercept	45.590^{***}
	(2.552)
Observations	534
R^2	0.023
Adjusted R^2	0.021
Residual Std. Error	11.605 (df = 532)
F Statistic	12.249^{***} (df = 1; 532)
Note:	*p<0.1; **p<0.05; ***p<0.01

As it can be seen in Table 1 and 2, the effect of age on wage (Y) is 0.105 and the effect of age (OV) on education (X) is -0.672. As such, the total OVB is 0.150*(-0.672) = -0.071.