Assignment 1

Andreas F. — Marc S. — Andreas L.

November 2, 2025

a)

Generate the log of wages (ln_wage). Produce a table with descriptive statistics for education, motivation, hours and ln_wage. Also calculate correlations between these variables and plot the density of log wages as well as a histogram for the years of education. Briefly comment on your results.

Table 1: Descriptive statistics

variable	n	min	max	median	iqr	mean	sd	se	ci
education	5000.00	10.00	24.00	12.00	4.00	12.8	2.42	0.03	0.07
motivation	5000.00	-4.56	5.14	-0.00	1.88	0.01	1.39	0.02	0.04
hours	5000.00	6.35	9.70	7.95	0.67	7.95	0.49	0.01	0.01
ln_wage	5000.00	2.37	4.12	3.35	0.32	3.36	0.23	0.00	0.01

Table 2: Correlations

education	motivation	hours	ln_wage
1.00	0.14	0.65	0.51
0.14	1.00	0.82	0.35
0.65	0.82	1.00	0.60
0.51	0.35	0.60	1.00

The summary (Table 1) reveals key aspects of the dataframe with 5000 individuals. Education ranges from 10 to 24 years, with a median of 12 years. Hours worked range from 6.35 to 9.7 hours daily, averaging 7.95 hours or about 39.8 hours weekly. Table 2 shows that all variables are positively correlated: motivation and education show a weak correlation of 0.14, while education with *ln_wage* and motivation with hours have stronger correlations (0.51 and 0.82).

The density plot (Figure 1) of *ln_wage* appears to be nearly normally distributed, with a mean of approximately 3.3. The bar chart (Figure 2) indicates that the relative majority of individuals have 10 years of education. Following this, around 600 individuals have 11 years, with increasing numbers for 12, 13, and 14 years. Beyond 14 years, the number of individuals with higher education decreases.

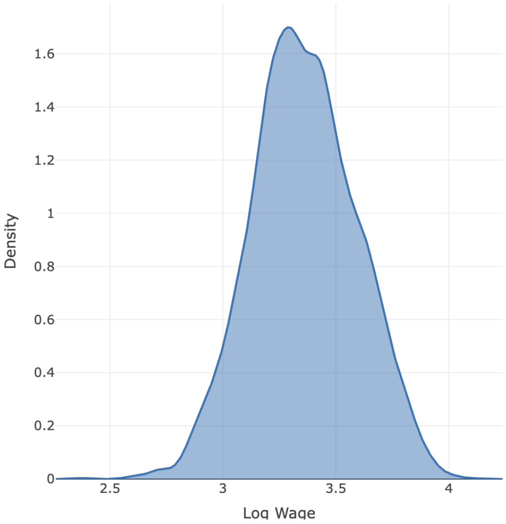

Figure 1: Density plot

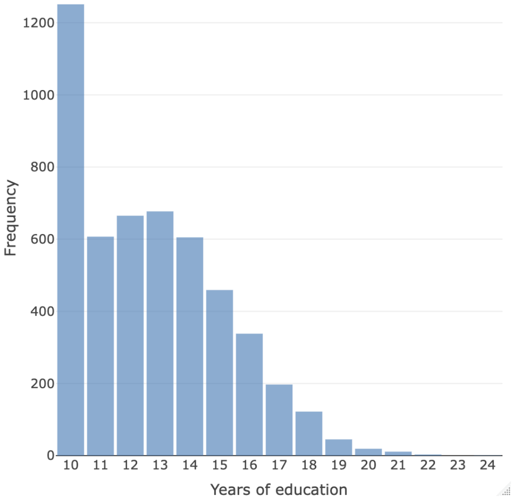

Figure 2: Histogram

b)

Compare descriptive statistics for individuals who have a motivation above or equal to zero versus below. Do the same thing for education, 14 years or more (some college) versus below. Again comment on what you find.

For individuals with 14 or more years of education, we observe higher average motivation (0.216 compared to -0.1), increased hours worked (8.3 versus 7.76), and a greater *ln_wage* (3.49 compared to 3.28). This suggests that motivation and education could be significant confounding variables when assessing the impact of wages on hours worked.

Table 3: Descriptive statistics (motivation)

M	variable	n	min	max	median	iqr	mean	sd	se	ci
0.00	education	2501.00	10.00	22.00	12.00	4.00	12.48	2.30	0.05	0.09
0.00	motivation	2501.00	-4.56	-0.00	-0.92	1.14	-1.10	0.83	0.02	0.03
0.00	hours	2501.00	6.35	9.01	7.62	0.49	7.63	0.36	0.01	0.01
0.00	ln_wage	2501.00	2.37	3.90	3.29	0.31	3.29	0.23	0.01	0.01
1.00	education	2499.00	10.00	24.00	13.00	4.00	13.06	2.51	0.05	0.10
1.00	motivation	2499.00	0.00	5.14	0.96	1.19	1.13	0.84	0.02	0.03
1.00	hours	2499.00	7.43	9.70	8.24	0.52	8.28	0.37	0.01	0.02
1.00	ln_wage	2499.00	2.83	4.12	3.42	0.31	3.42	0.21	0.00	0.01

Table 4: Descriptive statistics (education)

E	variable	n	min	max	median	iqr	mean	sd	se	ci
0.00	education	3200.00	10.00	13.00	11.00	2.00	11.24	1.18	0.02	0.04
0.00	motivation	3200.00	-4.29	5.14	-0.12	1.85	-0.10	1.39	0.03	0.05
0.00	hours	3200.00	6.35	9.19	7.75	0.56	7.76	0.41	0.01	0.01
0.00	ln_wage	3200.00	2.37	3.97	3.27	0.29	3.28	0.21	0.00	0.01
1.00	education	1800.00	14.00	24.00	15.00	2.00	15.48	1.53	0.04	0.07
1.00	motivation	1800.00	-4.56	4.17	0.20	1.85	0.22	1.37	0.03	0.06
1.00	hours	1800.00	7.08	9.70	8.30	0.56	8.30	0.42	0.01	0.02
1.00	ln_wage	1800.00	3.00	4.12	3.47	0.29	3.49	0.20	0.01	0.01

c)

Plot a scatter of the hours worked against ln_wage. What do you notice? There is a clear positive correlation between hours worked and *ln_wage*.

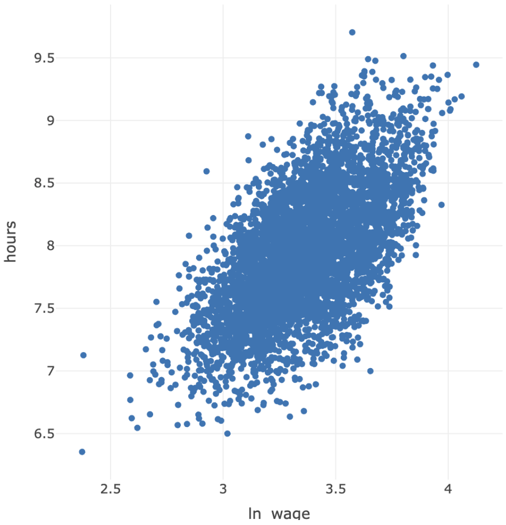

Figure 3: Scatter plot

d)

Do a simple regression of hours on ln_wage. Add the regression line to the plot from c). Include the 95% confidence interval and interpret your results.

The linear regression offers some evidence for the association between *ln_wage* and hours, aligning with our previous plot analysis. The narrow confidence band reflects a high level of precision in our estimation. Nevertheless, we exercise caution in interpreting this as a causal effect based on earlier considerations.

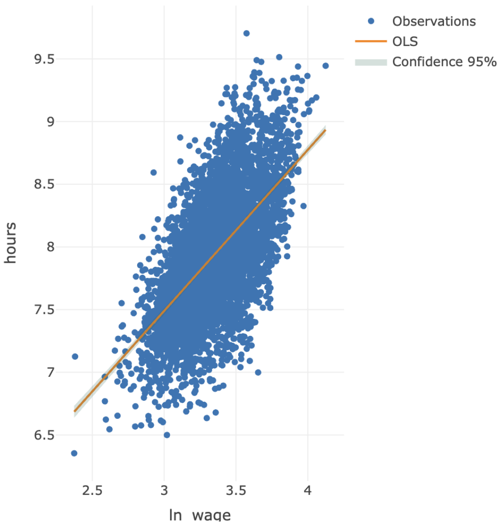

Figure 4: Regression plot

e)

Now add education to your regression and explain to what extent and why results change.

The coefficient for education in relation to hours worked is positive and statistically significant. Additionally, the coefficient on *ln_wage* decreases by one third, indicating the presence of omitted variable bias related to education. Nonetheless, wages remain significant in their effect on hours worked.

Table 5: simple OLS education

	Dependent variable:
	hours
lnwage	0.787^{***} (0.024)
education	0.092^{***} (0.002)
Constant	4.134^{***} (0.071)
Observations	5,000
R^2	0.518
Adjusted R^2	0.518
Residual Std. Error	0.340 (df = 4997)
F Statistic	$2{,}690.303^{***}$ (df = 2; 4997)
Note:	*p<0.1; **p<0.05; ***p<0.01

f)

Finally, do the full multivariate regression of hours on ln_wage, education, and motivation. Compare your results to before, i.e., out of d)–e), what is your preferred regression specifi- cation and results?

As expected, all coefficients for hours in relation to motivation, education, and *ln_wage* are positive and statistically significant, as indicated by the low p-values (less than 0.05 and less than 0.01). Overall, these three variables (*ln_wage*, education, motivation) show a significant relationship with the dependent variable (hours). The preferred approach is the full multivariate regression. We should include factors that likely influence both individuals' hours and wages simultaneously. Controlling for these variables in the regression is a classic method to mitigate OVB.

Table 6: simple OLS motivation

	Dependent variable:
	hours
lnwage	0.210^{***}
	(0.008)
education	0.100^{***}
	(0.001)
motivation	0.250^{***}
	(0.001)
Constant	5.968^{***}
	(0.022)
Observations	5,000
R^2	0.959
Adjusted R^2	0.959
Residual Std. Error	0.099 (df = 4996)
F Statistic	$38{,}858.290^{***}$ (df = 3; 4996)
Note:	*p<0.1; **p<0.05; ***p<0.01