Please refer to my code here. Thank you!

1 *Solution.*
OLS regression provides a good starting point. The challenge is that the coefficient can hardly be interpreted as the causal effect. When we run an OLS regression of log earnings on years of schooling, the coefficient β can be interpreted as the marginal return (in percentage terms) from a one-year increase in education. However, there are other potential explanatory variables that correlate with lifetime earnings. Although we can manage to include most of these variables in our estimation, some variables are intrinsically non-quantifiable. It is difficult to justify the mean independence between the error term and years of schooling. For instance, ability is one of the most common omitted variables in this regression. Since ability is intractable to measure, it seems inevitable that omitted variable bias will inhibit our identification of the causal relationship.

To overcome this issue, other causal inference regression models are promising alternatives. Given limited resources, experiments and randomised controlled trials are often not available. Difference-in-differences and regression discontinuity models require an exogenous change and stringent conditions on the characteristics of the treated and control groups; therefore, they may not be useful for typical survey data.

The instrumental variable (IV) model can be employed to overcome endogeneity. If there are variables in the survey data that are relevant to years of schooling yet uncorrelated with the error term, we can examine and justify the validity of the instruments. On one hand, good choices for instruments are rare in the literature. The challenge is to argue for the instrument's independence from the error term (the exclusion restriction). From my perspective, compulsory schooling laws and military draft lotteries are clever instruments. On the other hand, weak instruments, which are a key caveat of this method, can have a significant impact on the estimation.

To conclude, the return to schooling has been a very popular topic in labour economics for decades. We should take advantage of each causal inference model based on its strength. Under the specific context of the research and data, we should also be aware of the weaknesses and limitations of the model. □

3 *Solution.*

a The numbers of positive observations for $hrp1$ and $hrp2$ are 5342 and 1194 respectively. This difference may stem from the definition of the variables. Interviewees are often questioned about how many jobs they have and the corresponding earnings. It is reasonable that most interviewees have a main job but not multiple ones. $hrp1$ could be the hourly pay for the main job and $hrp2$ could represent the hourly pay for a second job.

b The unit of analysis is earnings per hour. The observed average is 2634.481. According to public statistics from FRED, the average hourly earnings of all employees in the US was approximately 36 in August 2024. The salient discrepancy suggests that the average of hourly earnings in our dataset is implausibly high.

However, I consider this a measurement error in the variable's units. If we assume the variable was recorded in cents per hour rather than dollars per hour, dividing the average by 100 yields approximately 26.34. This estimate is much more plausible and aligns with national data, 24, published by FRED in 2014.

c Taking the natural log of an income variable is a very common approach in economics. This method provides two advantages. First, the distribution of the log variable tends to be less skewed, which is a significant improvement compared to the distribution of the original variable(check the summary statistics above). Through this monotone transformation, the ordering of the original data is also well-preserved. Second, the log estimates can be interpreted as relative growth rather than nominal change, which can help economists shed light on the analysis of relative effects.

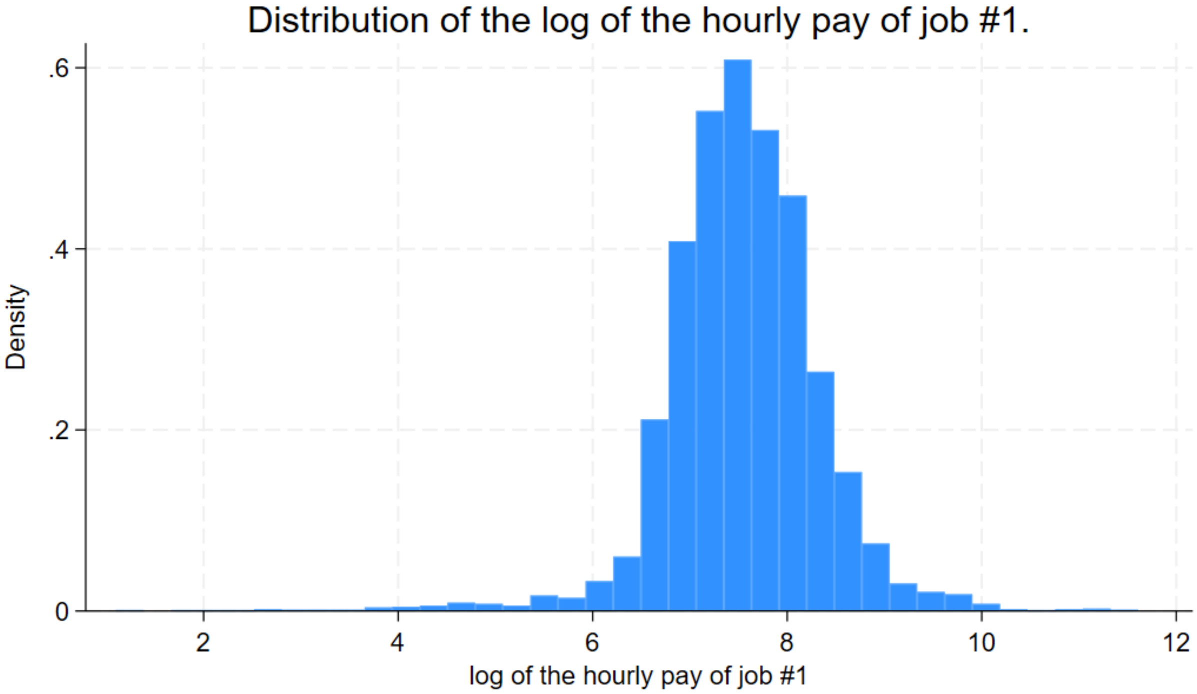

Figure 1: Histogram of the Log Hourly Earnings of Job#1

□

4 *Solution.*

a Sample weights are crucial in survey data, as they allow us to draw national-level inferences from a relatively small number of observations. The mean of the sample weight is $474,607.7$, which implies that an observation represents $474,607.7$ people in the national population on average.

However, this interpretation of the mean could be misleading. The true power of the weights comes from their variation. This variation is intentionally designed to ensure the sample accurately represents key national demographics, such as sex and race, which may have been sampled at different rates. It brings us to the next question which is a deeper look at the weights and will be more insightful.

b This table presents the mean sample weights, conditioned on different demographic groups. The mean of the cross-sectional groups is systematically higher than that of the supplemental groups. The purpose of the weights is to allow us to combine all the samples (cross-sectional and supplemental) into a single, nationally representative dataset. This system decreases the influence of the supplemental groups and emphasises the cross-sectional groups.

Diving into each group, we first find that Hispanic and Black samples have lower weights in both the cross-sectional and supplemental groups. The NLSY79 documentation confirms this is because these groups were intentionally oversampled. Therefore, it is this imbalanced sampling that contributes to the different average sample weights. On the other hand, military samples are deliberately separated. As voluntary armed forces are intrinsically different from the general population, we cannot treat them the same as other groups. Interestingly, the sample weight of Black samples in the military group is not lower; instead, it is Hispanic and white female samples that are assigned a greater sample weight. This likely relates to the actual representation of these groups in the US armed forces during that period.

c Taking sample weights into account can prevent a regression from employing too much variation from the oversampled groups and helps to recover more precise, nationally-representative estimates. However, there are multiple methods to address this concern.

To achieve the desired effect on a national level, using sample weights in a single regression is intrinsically similar to estimating the effect for each group separately and then aggregating those results based on each group's true proportion of the population.

If we treat the sample weight as a person weight in a regression model, the estimation will generate a nationally representative result by weighting the influence of each observation. In contrast, computing the effect for each group first gives us conditional results, which we then aggregate based on their weights to recover the national-level estimates. These two methods are, in essence, conceptually identical.

□

5 *Solution.*

a The modal number of years of schooling in the data is 12th grade, which accounts for approximately 41% of the sample. The second most common number of years of schooling is 16th grade, which accounts for about 14% of the sample.

This seems intuitive and plausible. The largest group of people achieves their highest degree from high school (12th grade). Four-year college graduates (16th grade) come next. There are also some two-year college graduates, which is why 14th grade is the third most common in the data. It is worth noting that the data for those who did not complete high school shows an interesting pattern. The percentage of respondents whose highest grade completed is 11th grade is higher than that for 10th grade, which is in turn higher than for 9th grade. This implies that among the population who did not finish high school, it was more common to leave after completing 10th or 11th grade than after 9th grade.

b Six observations were dropped in the process. This is a very common phenomenon in survey data. Sometimes, we may face nonresponse or encounter cases that do not satisfy our design and requirements. For example, a respondent may intentionally avoid answering certain questions. It is also possible that some people in this sample received education from a particular institution that is not documented in the general education system.

□

6 *Solution.*

a By subtracting their birth years from 2014, we can compute the respondents' ages. The minimum age is forty-nine, and the oldest sample in the data is fifty-nine. cI would expect age to be correlated with wages. It is very common that experienced labour tends to acquire higher earnings than younger labour, ceteris paribus. It is definitely a control variable that has to be taken into consideration.

b Presumably, the effect of age is likely not just linear. A simple linear term assumes that the wage increase from age 49 to 50 is the same as from age 58 to 59, which is unlikely. A more realistic hypothesis is that age has a diminishing marginal impact. Although age is expected to be positively correlated with wages, the marginal effect of an additional year of age may be less salient as people get older. In this sense, a more flexible model would include a squared term for age. We would expect the coefficient on the linear term to be positive, while the coefficient on the squared term would be negative, capturing this classic concave age-earnings profile. Hence, the squared term is definitely a control variable that has to be taken into consideration.

c Mather's average years of schooling is slightly higher than father's given sample weights. The 95% confidence intervals of these two estimates are not overlapping. However, a proper t-test may be still required to draw further inferences.

d As parents' average years of schooling pattern, female respondents tend to have a higher average education level than male respondents. This time, the 95% confidence intervals for the male and female averages are overlapping. As mentioned previously, this makes a formal t-test necessary to conduct a meaningful comparison and determine if this difference is statistically significant. More importantly, it appears that respondents of both sexes have a high average level of education. Although we would need a statistical test to confirm, this large increase in educational attainment across generations is a very common and important finding here.

e I incorporated two methods to estimate the correlation suggested by Bill Sribney, StataCorp. The first method (Method 1) employs a direct point estimation using sample weights. The second method (Method 2) is based on a regression. The estimated correlation between the child's and mother's years of schooling was 0.2486 (Method 1) and 0.283 (Method 2), respectively. In contrast, the estimated correlation between the child's and father's years of schooling was 0.2677 (Method 1) and 0.404 (Method 2), respectively.

If we have more confidence in the first method (the direct weighted correlation), then the correlations of the child's schooling with the mother's or father's schooling are relatively similar. In this case, the child's schooling appears slightly more correlated with the father's years of education than the mother's which aligns with the findings from Taiwan(Liu et al. (1999)).

f According to NLSY documentation, AFQT scores are calculated from the ASVAB tests, which are multi-aptitude tests originally designed for military entrance. This provides us with measures of the samples' ability in arithmetic reasoning, mathematics knowledge, word knowledge, and paragraph comprehension. Because ability is intrinsic and hard to quantify, this test serves as a proxy for comprehension and cognitive ability at most. However, there are other dimensions of ability that this test falls short of capturing (e.g., motivation, social skills, creativity).

g Around 25% of observations report having health problems that hinder their ability to work.

□

7 *Solution.* For the following analysis, please refer to table 1

a Because we are considering a log dependent variable, the coefficient is often interpreted as percentage change. On average, one additional year of schooling is associated with a 11.7% increase in hourly pay. Note that hourly earnings is possibly measured by cents. It is reasonable to observe that the percentage change in hourly earnings is positively correlated with an increase in years of schooling.

b After incorporating control variables, the magnitude of the effect of years of schooling shrunk to 10.8%, ceteris paribus. This is comprehensible, as the original estimation was likely overestimated due to omitted variable bias.

c The magnitude of the estimated returns to schooling fell to 8.29% when the AFQT test score was added as a control variable. This decline is understandable as the original model suffered from omitted variable bias. Ability is expected to be positively correlated with both schooling and earnings. Because the original estimation did not control for ability, the schooling variable was capturing both the true effect of education and the effect of ability, leading to an overestimation of the returns to schooling.

d The estimated coefficient on `heath_problem` is -35.9%, implying that people who have a health issue limiting their working ability tend to earn 35.9% less than those who do not, ceteris paribus. The estimated returns to schooling, at 8.08%, did not vary much. However, adding this control is crucial, since health problems not only affect earnings but are also correlated with years of schooling. Disability is a major concern for people not attending school.

e Using years of schooling as dummies enables the analysis of the returns for different groups. Compared to people whose highest degree is ungraded, there is no evidence that the estimated returns are different. However, the number of observations for those who received ungraded education is scarce. This suggests the key comparisons may not be against this small group.

Once we restrict our sample to at least 8 years of schooling, the evidence suggests that higher education degree groups tend to experience higher earnings compared to 8th-graduates. Particularly, after the 14th grade, equivalent to a 2-year college degree, the effect becomes significantly positive. This aligns with the modern literature on wage polarisation.

Finally, if we change the baseline group to 12th-grade graduates, the pattern becomes even more intriguing. People with less education experience significantly negative log hourly earnings (relative to high school graduates), while more educated people enjoy significantly positive log hourly earnings. This is consistent with the previous findings and with wage polarisation in relation to skills (education) that the high-skilled labour(college graduates) receive higher earnings than the low-skilled labour(high school graduates) do.

f We have attempted to control many potential omitted variables to improve the estimates. If we can only exploit this survey without other resources, I think we can add region characteristics, marital status, and family structure and composition. These variables are correlated with both years of schooling and earnings.

Please refer to my code here. Thank you!

Table 1: Returns on Years of Schooling: Regression Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	ln_hr_wage	ln_hr_wage	ln_hr_wage	ln_hr_wage	ln_hr_wage	ln_hr_wage	ln_hr_wage
yschl	0.117***	0.108***	0.0829***	0.0808***			
	(19.73)	(17.49)	(12.25)	(12.10)			
female		-0.344***	-0.331***	-0.318***	-0.331***	-0.332***	-0.332***
		(-14.13)	(-13.19)	(-12.78)	(-13.21)	(-13.25)	(-13.25)
age		-0.114	-0.0485	-0.0558	-0.0488	-0.0509	-0.0509
		(-0.39)	(-0.16)	(-0.19)	(-0.17)	(-0.17)	(-0.17)
agesq		0.00106	0.000458	0.000554	0.000463	0.000478	0.000478
		(0.39)	(0.17)	(0.20)	(0.17)	(0.17)	(0.17)
mom_schl		0.00711*					
		(1.73)					
pop_schl		0.0129***					
		(4.11)					
afqt			0.00608***	0.00585***	0.00549***	0.00550***	0.00550***
			(11.58)	(11.28)	(10.59)	(10.59)	(10.59)
health_problem				-0.359***			
				(-7.80)			
2.yschl					0.188		
					(0.36)		
3.yschl					-0.749		
					(-1.38)		
4.yschl					-0.323		
					(-0.61)		

5.yschl					0.0898		
					(0.16)		
6.yschl					-0.162		
					(-0.30)		
7.yschl					-2.507*		
					(-1.76)		
8.yschl					-0.224		-0.106
					(-0.42)		(-0.89)
9.yschl					-0.553	-0.328*	-0.434***
					(-1.02)	(-1.85)	(-3.25)
10.yschl					-0.321	-0.0965	-0.202***
					(-0.61)	(-0.72)	(-3.06)
11.yschl					-0.369	-0.145	-0.251***
					(-0.70)	(-1.09)	(-3.89)
12.yschl					-0.118	0.106	
					(-0.23)	(0.89)	
13.yschl					-0.0328	0.191	0.0854*
					(-0.06)	(1.50)	(1.75)
14.yschl					0.0361	0.260**	0.154***
					(0.07)	(2.14)	(4.56)
15.yschl					0.0476	0.272*	0.166**
					(0.09)	(1.90)	(2.04)
16.yschl					0.293	0.517***	0.411***
					(0.56)	(4.12)	(9.95)
17.yschl					0.240	0.464***	0.358***
					(0.45)	(3.24)	(4.45)
18.yschl					0.341	0.565***	0.459***
					(0.65)	(3.98)	(5.82)
19.yschl					0.554	0.777***	0.672***
					(1.04)	(4.79)	(6.06)
20.yschl					0.529	0.753***	0.647***
					(1.00)	(5.29)	(8.15)
_cons	6.025***	9.122	7.616	7.803	8.763	8.609	8.714
	(74.74)	(1.18)	(0.97)	(1.00)	(1.12)	(1.10)	(1.12)
N	5337	5337	5131	5131	5131	5094	5094

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

□

References

Liu, J.-T., Hammitt, J. K., and Jeng Lin, C. (1999). Family background and returns to schooling in taiwan," economics of education review, elsevier. *Taiwan Economic Review*, 19(1):113–125.