My OLS regressions

In order to study the causal effect of education on wages and based on my I decided to set up a model first.

$$lnincome = a + b * hgc + \varepsilon \quad \text{(M-1)}$$

Table 1

Variable	Variable Title
lnincome	=Log(income) Income: Total income from wages and salary in past calendar year
hgc	Highest grade completed
afqt	Profiles, armed forces qualification test (AFQT) percentile score
abi	ability

First of all, I want to run the regression of log income only on the highest grade completed of equation M-1. I modified my data set by deleting all invalid or valid skip data. Then, I still have 4769 observations left, which is still a representative and large enough data set. By running that regression, I got Table 2.i

Table 2

Source	SS	df	MS		
Model	49.3286607	1	49.3286607	Number of obs	= 4769
Residual	6805.18456	4767	1.42756127	F(1, 4767)	= 34.55
				Prob > F	= 0.0000
Total	6854.51322	4768	1.43760764	R-squared	= 0.0072
				Adj R-squared	= 0.0070
				Root MSE	= 1.1948

lnincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hgc	.0610417	.0103842	5.88	0.000	.0406838	.0813996
_cons	6.845127	.1236778	55.35	0.000	6.602661	7.087592

From the table, we can observe that both *hgc* and *_cons* are statistically significant at 5% significance level. Which means, increasing the highest grade completed by one year could increase the wage around 6.1 log points.

Secondly, I want to know whether both ability and educational level will influence a person's wage level. I want to get the partial effect of the number of years of schooling on wages. We can get data on the years of education from the NLSY79 data set. But we cannot describe the individual's ability directly through a variable since ability is unobservable. However, we can find some kind of test score data to represent an individual's cognitive ability, then the test score can be treated as a potential proxy variable. There are two assumptions that the proxy variable must satisfy:

$$P(abi|1, hgc, afqt) = p(1, afqt) \quad \text{(M-2)}$$

$$E(income|hgc, abi, afqt) = E(income|hgc, abi) \quad \text{(M-3)}$$

The first hypothesis means that once you know the test scores, there is no correlation between a person's years of schooling and ability. The second hypothesis means that once we know the years of schooling and the data for ability, the individual's test scores have no explanatory power for the income.

The Armed Services Vocational Aptitude Battery (ASVAB) score is used as a proxy variable for at least cognitive ability. This test was initially a basic literacy test for enlisted personnel and was later incorporated into NLSY. Among the questionnaires, the test was generally completed by individuals aged 17-18 (What is the ASVAB). Theoretically, this is a competency test for individuals before entering the labor market.

Then I define the new regression M-4 by modify the equation M-3 as:

$$lnincome = a + b * hgc + c * afqt + u \quad \text{(M-4)}$$

After running this regression, I got the Table 3 below.

Table 3

Source	SS	df	MS		
Model	66.1321613	2	33.0660806	Number of obs =	4769
Residual	6788.38106	4766	1.42433509	F(2, 4766) =	23.22
				Prob > F =	0.0000
Total	6854.51322	4768	1.43760764	R-squared =	0.0096
				Adj R-squared =	0.0092
				Root MSE =	1.1935

lnincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hgc	.038252	.0123131	3.11	0.002	.0141127	.0623914
afqt	.0023767	.000692	3.43	0.001	.0010201	.0037332
_cons	7.000984	.131608	53.20	0.000	6.742971	7.258996

By interpreting the table, we can see *hgc* and *afqt* are both statistically significant at 5% significant level. In this case, the partial effect of education on log wages is around 3.8 log points. Besides, increasing the "Profiles, armed forces qualification test" score by one percent will also increase the log wage by 0.24 log points.

Since I am very interested in understanding the impact of gender on this causal effect, I ran the equation M-4 for the set of all 2384 male observations, and I got the Table 4:

Table 4

Source	SS	df	MS		
Model	10.3903992	2	5.1951996	Number of obs =	2387
Residual	3161.99307	2384	1.32633937	F(2, 2384) =	3.92
				Prob > F =	0.0200
Total	3172.38346	2386	1.32958234	R-squared =	0.0033
				Adj R-squared =	0.0024
				Root MSE =	1.1517

lnincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
afqt	.0007371	.0009251	0.80	0.426	-.0010769	.0025511
hgc	.0292383	.0165285	1.77	0.077	-.0031735	.0616501
_cons	7.407074	.1730873	42.79	0.000	7.067656	7.746491

Similarly, the table I got for the 2387 female observations is in Table 5 below:

Table 5

Source	SS	df	MS	Number of obs =	2382
				F(2, 2379) =	35.71
Model	100.747761	2	50.3738806	Prob > F =	0.0000
Residual	3356.30017	2379	1.41080293	R-squared =	0.0291
				Adj R-squared =	0.0283
Total	3457.04793	2381	1.45193109	Root MSE =	1.1878

lnincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
afqt	.0033171	.0009989	3.32	0.001	.0013583	.0052759
hgc	.0885609	.017873	4.95	0.000	.0535125	.1236092
_cons	6.13049	.1950414	31.43	0.000	5.748022	6.512959

However, from Table 4, we can see both *afqt* and *hgc* are not significant at 5% significance level. So, we cannot conclude that either education level or ability level play an important role for males to get a higher wage. Nevertheless, in Table 5, we can conclude that both education level and ability level have significant influence on female's income increase.