Is Education the Key to Higher Income?

Trina Beaton, Akeem Coburn*
Department of Economics
University of Toronto, Mississauga

April 12, 2024

ECO475 Term Paper

Abstract

To what effect does the adoption of higher education change one's income distribution, and what does this mean on a economic stage? This study investigates the probabilistic relationship between the number of years of one's education and an indicator variable of whether one's income is over $50,000 or not, and their trend patterns using a cross-sectional data model. By uncovering these inter-variable relationships this study reveals an extensive positive relationship between them and a positive softening internal correspondence in their marginal effects, as communicated in the relevant literature by Card (1999). We follow up by suggesting policies directed towards addressing the impact education has on income, highlighting its social implications. Finally, we address all limitations to the model posed by threats to external and internal validity.

*Contact email: trina.beaton@mail.utoronto.ca, akeem.coburn@mail.utoronto.ca. I would like to thank Professor Eduardo for his guidance and support throughout this project. Thanks also to Bingyao Liu for suggestions and providing this template. All remaining errors are our own.

1 Introduction

The pursuit of higher education and its implications for economic outcomes, particularly income, has long been of interest to economists. Extensive research has established a robust positive association between years of education and earning potential, grounded in seminal theories such as human capital theory and empirical models like the Mincer earnings function. However, while existing literature provides valuable insights into the broad relationship between education and income, we wish to include and dissect important variables that are normally grouped as unobservables to extrapolate additional insights into the income-education relationship.

To answer the fundamental question, *"How does education influence income, and what factors mediate or moderate this relationship?"*, we extend the base model with other potential determinants of the size of ones earning and apply various estimation methods. In this paper we present our findings, compare direct impacts, and make valuable predictions.

2 Literature Review

Economists have researched the relationship between income and education in great detail, and empirical data has always shown that the number of years of school positively correlates with earning capacity. Its intricacies and ramifications have been explored in a number of peer-reviewed econometrics publications that have examined this relationship using different approaches. Here, we summarize the most important discoveries from a few chosen studies and point out the gaps that our study attempts to fill.

The foundation for comprehending the financial benefits of education was established by Becker's groundbreaking work on human capital theory. This idea holds that people invest in education to develop their human capital, which raises their lifelong earning potential and productivity (Becker, 1964). The impact of education on incomes was then investigated using instrumental variable techniques to overcome endogeneity concerns in Card's seminal study on the returns to education. Through the use of exogenous differences in state-level laws requiring education, Card provided strong evidence of the relationship between education and income (Card, 1999). Finally, Oreopoulos et al. investigated the role of field of study in shaping the returns to education, focusing on the earnings differentials between STEM (Science, Technology, Engineering, and Mathematics) and graduates from other fields. Their findings revealed substantial differences in earnings trajectories between STEM and non-STEM graduates, with STEM graduates obtaining greater salaries (Oreopoulos et al., 2006). These studies addressed endogeneity issues and highlighted the importance of considering field-specific factors in understanding the relationship between education and income.

While these and other existing literature provides robust evidence of the positive association between education and income, several gaps remain that our research seeks to address. Our study utilizes advanced econometric techniques, such as logit and probit models, to estimate the effect of education on income while accounting for potential confounding factors. With the use of these methods, we treat endogeneity and omitted variable bias concerns and offer more accurate estimates of the causal relationship between education and income. Furthermore, in our unique methodology, we generate many specifications that account for extrinsic variables such as racial origin, gender, and socioeconomic status in order to investigate the differential impact among various demographic subgroups.

3 Data

We commence our analysis with data collected from a data database, *Kaggle*, which is sourced from the United States Census Bureau and contains census information on income and other factors from 1994. Our data is cross-sectional, presenting probabilistic causation between dependent and independent variables. The data collected is sampled from 16,383 people completing the survey. The census entries provide a variety of intrapersonal factors characterized by age, education level, work class, and other economic performances. Concerning our analysis, our population of interest is the people who had completed the census fully at the time of its release in 1994. For our analysis, we consider only the cases with full completion of the survey; thus, any incomplete answers are removed from the data to control for sample selection errors. The primary variable of interest is the binary variable concerning whether or not a person has a yearly income over $50,000, measured in USD. The main independent variable is the number of years of education. Table 2 presents the variables that are important to our experiment, along with their description and some examples.

Within our data, we can see that the majority of people, 32.31 percent, are high school graduates, closely followed by those with bachelor's degrees at 16.88 percent, as seen in Figure 1. Additionally, 75.33 percent of the census participants earn \$50,000 USD or less, as presented in Figure 2. By glancing at the individual distributions, there is some connection between the number of years of schooling and the income reported. However, we are interested in determining the magnitude to which years of education impact the probability of a person earning a yearly income over \$50,000 USD, which is not quite answered only by looking at each variable's distribution.

4 Model

4.1 Linear Regression Model

Our fundamental model concentrates on the relationship between whether a person earns a yearly income over \$50,000 and their level of education, documented by a total number of years. We start off with a classic OLS model:

$$income = \beta_0 + \beta_1 educnum + u, u \sim N(0, \sigma^2) \tag{1}$$

where β_1 is the main parameter of interest and measures the approximate percentage change in earnings resulting from an additional year of schooling. However, we are aware of negative probabilities that arise as a common problem when using ordinary least squares regression to analyze a probability-centered model. Thus, we document the result and integrate probit and logit estimations to rectify the possible negative probabilities present.

4.2 Single Variable Probit and Logit Models

To better analyze the true relationship between whether a person makes over \$50,000 in income a year and the number of years of their education, specifications (2) and (3) focus on probit and logit estimations, respectively, correcting the mis-specification issue and uncovering the predicted probabilities. Assuming that the properties for probit and logit are met, the model we consider take the form of:

$$
\begin{aligned}
income^{*} &= \beta_0 + \beta_1 educnum + u, u \sim N(0,\sigma^2) \\
income &= 1[income^{*} > 50K] \\
&= \begin{cases} 0 & income^{*} \leq 50K \\ 1 & income^{*} > 50K \end{cases}
\end{aligned}
\tag{2}
$$

where $income^{*}$ is a latent variable and we predict the probability that $income = 1$.

4.3 Logit Model with Multiple Independent Variables

For the logit regression to be accurate and true, the independence of errors condition must be satisfied. This means that there are no omitted variables, and the model itself only includes the necessary and valuable independent variables. Specifications (4), (5), and (6) provide solutions to these concerns and converge to an accurately predicted estimation. These estimations apply variants of the model that is inclusive of all the relevant exogenous variables:

$$
\begin{aligned}
income^{*} &= \beta_0 + \beta_1 educnum + \beta_2 age + \beta_3 hrspw \\
&\quad + \delta_1 d_1 Black + \delta_2 d_2 Asian + \delta_3 d_3 Other \\
&\quad + \delta_4 d_4 female + \delta_5 d_5 married + \delta_6 d_6 separated \\
&\quad + \delta_7 d_7 divorced + u, u \sim N(0,\sigma^2) \\
income &= 1[income^{*} > 50K] \\
&= \begin{cases} 0 & income^{*} \leq 50K \\ 1 & income^{*} > 50K \end{cases}
\end{aligned}
\tag{3}
$$

where $income^{*}$ is again a latent variable and d_i are dummy variables *maritalstatus*, *gender* and *race*.

5 Results

5.1 Linear Regression Model

Specification (1) shows that there is a weak positive relationship between a person having a yearly income over \$50,000 and their years of schooling. On average, a 1-year increase in one's years of education constitutes a 0.06 unit increase in whether a person makes a yearly income over \$50,000. We expect 95 percent of other samples' β_1 measure to fall within 0.05 and 0.06. Therefore, 95 percent of other samples will also reflect this weak positive relationship between the independent and dependent variables. Furthermore, the coefficient is statistically significant at the 5 percent level, with a p-value close to 0. Hence, we can reject the null ($H_0 : \beta_1 = 0$) with a t-value of 42.61. The linear model shows that any increase in a person's education level impacts their yearly income. This reinforces the social idea that higher education is important as it defines the opportunities received in life as well as income, but also that one's education level is economically significant.

Due to the nature of the data, a 0.06 unit increase in whether one makes a yearly income over \$50,000 (Y = 1) causes some concerns. Even though the sign of β_1 is in the expected positive direction, its magnitude does not align with our expectation. Thus, we must change our view to consider probabilistic causation instead of a direct causal effect. Additionally, the Ramsey reset test, which checks for omitted variable bias, and the Hettest, which checks for heteroskedasticity, both fail with prob > F and prob > Chi2 being both zero, respectively. Since the linear model fails these tests, we can conclude that this model is not reliable, resulting in a mis-specification issue. Specifically, the least squares assumptions 1, 5, and 6 fail, generating an inconsistent and potentially biased estimation.

5.2 Single Variable Probit and Logit Models

Considering specifications (3) and (4), the probit estimation yields a marginal effect of 0.0571, while the logit model estimates a marginal effect of 0.0578. Therefore, on average, for each additional year of education, the model predicts an increase of approximately 58 percent in the probability of earning over \$50,000. These values are statistically significant at the 5 percent level, with p-values of 0 and economically significant due to the small standard errors. Additionally, with a McFadden R^2 value of 0.105, the logit model proposes a better fit than the probit model which has a McFadden

R^2 value of 0.103. Therefore, we may conclude that the logit model provides a better fit for the data and our empirical question.

5.3 Logit Model with Multiple Independent Variables

In specification (4), we see that a one-year increase in one's education increases the odds of having an income over \$50,000 by 1.4 times as much. This demonstrates an even bigger change in the predicted probability that one has a yearly income over \$50,000 than the previous estimations. Moreover, specification (4) shows that being a male increases the odds of having an income over \$50,000 by 3.5 times as much, and a one-year increase in age barely effect the odds of having an income over \$50,000 with an odds ratio of 1.05.

This shows that along with the number of years of one's education and gender, their age and race; using White as the basis, help to determine the predicted probability of one's binary income distribution. These are all statistically significant at the 5 percent level, with p-values of 0 except race, and economically significant since any change in the predicted probability of one's income distribution has ripple effects across the economy reflecting social and economic changes, such as social importance and the poverty line.

Correspondingly, an increase in the years of education increases those odds of one having an income over \$50,000 by almost 1.407 times as much. This means that now, with more independent variables, education has a minor decrease in effect on the predicted probabilities of one having a yearly income over \$50,000. In similar fashion, being a male has a decreased coefficient by just over 3 times as much. However, as seen in specification (5), a one-unit increase in age and hrspw does not affect the odds of earning a yearly income over \$50,000. Additionally, the odds are decreased with movement from race selections that are not White. These are all statistically significant with p-values of 0, except for race, with a p-values ranging from 0 to 0.021. The marginal effect of increasing years of education increases the predicted probability that one makes over \$50,000 in yearly income by 0.052 percentage points. This is a minor increase, but nonetheless statistically and economically significant.

The consideration of multicollinearity violates the idea that only the necessary variables are included in the model, and all there exists no omitted variable bias. Thus, one of the logit assumptions fails.

Our model correctly predicts those who earn over \$50,000 in yearly income 35.25 percent of the time, and correctly predict the people who make equal to or less than \$50,000 93.37 percent of the time. This is an increase in correctly predicting those with an income over \$50,000 but a decrease in predicting those with an income equal to or less than \$50,000. But overall, we correctly classified 79.03 percent of our model, which is an increase from the previous one. Overall, we have increased our McFadden R^2 by 0.017 units by a value of 0.21, and have significant results on a likelihood ratio test compared to the previous regression. Thus, we have an improvement in fit with this specification.

Finally, from specification (6), the addition of *maritalstatus* increases the odds that an increase in the number of years of one's education has on the predicted probability that one makes over \$50,000 in yearly income to 1.45 times as much. With the exception for race, which has an increased magnitude in comparison to the previous model, the other independent variables have decreased odd ratio magnitudes but still reflect a positive relation to the predicted probability that one makes a yearly income over \$50,000. These are all statistically significant with p-values of 0, allowing us to reject the null (H_0: $\beta_i = 0$), except now gender is only statistically significant at the 0.1 level. *maritalstatus*, with a civilian spouse as the base, decreases the odds of one earning an income of over \$50,000 by a factor of 0.01 to 0.1 at a statistically significant level, except for having a spike in the military, which increases the odds of one having a yearly income over \$50,000 by 1.32 percentage points.

The marginal effect of the number of years of education on the probability of earning over \$50,000 in yearly income, tells us that as the number of years of one's education increases, the educational effect on the predicted probability increases. Therefore, one more year in school from high school to higher education has a larger effect on the predicted probability of the income distribution, than one more year in school from within the years of higher education. Additionally, we can see that the marginal effect of education on the predicted probability that one earns over \$50,000 in yearly income conditional on gender, has a minor impact at a non-significant level since the 95 percent confidence intervals overlap. This can be seen in Figure 4 - 9. The classification tells us that we correctly predict people with incomes over \$50,000 48.8 percent of the time, and we correctly predict people who earn equal to or less than \$50,000 91.9 percent of the time. This is an overall increase in correctly classifying our model by 81.27 percent. The LR test suggests an improvement in fit, which is also reflected in the 0.11 increase in the McFadden R^2, allowing for a better estimation of the years of education on the probability of earning a yearly income over \$50,000.

The positive probabilistic relationship and significantly positive results between the number of years of education and the probability that one makes over \$50,000 in yearly income allow insight into education's role in biased income gaps. As studies suggest, there is a wide Black-White gap in income and wealth, with a typical White family earning almost 8-times as much wealth as a typical Black family. In addition to race gaps in income, there are also gender-based gaps in income, where women earn less than men. In previous years, the gender wage gap and its impact on income inequality have lessened over the years. Nonetheless, its constant presence is still worth consistently re-studying and analyzing. As reflected in our results, being Black decreases the odds of making \$50,000 in yearly income, by a factor of 0.81. The impact education has on the predicted probability of making a yearly income over \$50,000 is greater for a woman than a man. Thus, being educated as a woman has a considerable marginal effect on income. Many policies to reduce these income gaps have been suggested and enacted, such as programs like Ontario Student Assistance Program (OSAP). However, as Card (1999) suggests and our model reflects, individual returns or marginal effects on education decline with the jump to higher levels of education. Thus more structural policies that address racism and sexism in society and the economy are needed to close the divide in income, targeting children and high school students. Policies like tax reforms and children-based interest-funded bank accounts are among the suggestions.

The role played by education in modern labour markets, as Card (1999) mentions, is overwhelming. The social and economic attention education warrants, suggests that it could be a solution to several income-related social issues. The main issue when discussing income distribution is the poverty line. The poverty line as a measure, defines when and where a person Is dubbed by society as being poor. More specifically, a person makes below the necessary or standard amount to live with some comfort. Economic disturbances such as market crashes or global factors such as war and disease often inflate the dimensions of poverty. Our model, by showing how much education impacts the predicted probability of making over \$50,000 in yearly income reflects where the majority of society lies. This allows insight into the interactions between education and the sorting of a person into poverty, while also providing a solution to reduce the percentage of the population living in poverty.

5.4 Limitations

In the first place, a logit model is quite flexible, making it harder for model limitations to be present. The data, however, shows the presence of bias and reflects variation in external validity. Cross-sectional studies do not conduct causal inference, and have an incompetence to analyze inference. These pitfalls induce information bias. An example of information bias that's prevalent is recall bias. Because the data is constructed from census results, how each person recalls the information and records it may be biased. This bias can affect any of the independent variables due to the inflation or deflation of the correctness of the census, skewing the bias in a positive or negative direction.

Detection bias is another example of information bias that is present. Detection bias is a little more systematic in that it focuses on systematic differences between groups. Since a single person can belong to many groups defined by age, race, gender, etc., the predicted probabilities can be affected by bias. Moreover, due to the nature of the data, we can also see that selection bias is present. We inevitably have a non-responsive bias, resulting in a person's inability to complete the survey. This can skew results in that there is a possibility that the census responses are not reflective of the population, which can only be solved by increasing the responses to the census or by fact checking the responses.

The largest problem within our data and model relies on the idea conceptualized by James Heckman. Our dependent indicator variable depends on the self-selection present since we only see income for people who work or self-select as having an income. This causes a correlation between the unobservables and a negative correlation between the independent variables and unobservables based on the self-selection that one has an income. Mainly, the idea behind having an income is in liaison with being employed legally and in general. This causes the exemption of non-participants presented in the model, causing a negative bias in OLS, and an inconsistent model. The solution to this is to enact the controlled function approach and find an independent variable to incorporate into the inverse Mills ratio to remedy the bias. Our model lacks such a solution due to the difficult nature of finding a viable independent variable in the data that produces an appropriate inverse mills ratio, reflecting omitted variable bias reflected in statistically significant values of _hatsq in the linktest results.

Simultaneously, the model lacks external validity, causing bias that would make the regression futile. Although the predicted probability of one making a yearly income over $50,000 can be generalized to the population, its accuracy after doing so should be questioned. Due to the census nature of the data, it is used to reflect the general population in that time frame. The model's results cannot be expanded to other time frames or countries due to the quick pace of economic activity and the increasing usage of technology. Thus the interpretation of the data on a global demeanor or in a different time frame would be positively or negatively skewed.

6 Conclusion

In order to answer, *"How education influences income, and to what extent related factors mediate or moderate their relationship?"*, we studied the probabilistic relationship between the number of years of one's education and a binary variable of whether one's income is over $50,000 or not. The empirical observations are harmonious of the notion that there is a substantial positive relationship between education and the predicted probability of one's yearly income being over $50,000, as supported in the paper by Card (1999). We then explore the pursuit of higher education and weak internal correspondence in education's marginal effects on income. We refer to studies by the joint economic committee, as well as other global policies, concluding that most policies that aim to reduce income gaps should focus their attention towards tax reforms and youth-centered financial grants to ease access to education. In essence, this study uses a logit regression to analyse the probability centered relationship between education and income, and their average marginal effects to answer the question posed initially about their effects on society.

7 Biblography

Becker, G. S. “Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education”. 1964. Columbia University Press.

Beyer, Don. “Education Can Help Narrow the Racial Wealth Gap, but Structural Solutions Are Needed to Close It.” Joint Economic Committee Democrats, JEC, 2019, www.jec.senate.gov/public/ $_{c}ache/files/1d07cb0d-6ec2-4f49-9fa7-6ee5c771fbe3/education-and-racial-wealth-gap.pdf$.

Card, D. “The Causal Effect of Education on Earnings”. 1999. Handbook of Labor Economics, 3, 1801-1863.

Government of Canada, Statistics Canada. “Dimensions of Poverty Hub.” Government of Canada, Statistics Canada, Statistics Canada, 10 Oct. 2023, www.statcan.gc.ca/en/topics-start/poverty.

Government of Canada, Statistics Canada. “Does education pay? A comparison of earnings by level of education in Canada and its provinces and territories” Census in Brief: Does Education Pay? A Comparison of Earnings by Level of Education in Canada and Its Provinces and Territories, Government of Canada, Statistics Canada, 29 Nov. 2017, www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016024/98-200-x2016024-eng.cfm.

Hagle, Timothy M., and Glenn E. II. “Goodness-of-fit measures for probit and logit.” American Journal of Political Science, vol. 36, no. 3, Aug. 1992, pp. 762–784, https://doi.org/10.2307/2111590.

Heckman, James J. “Characterizing Selection Bias Using Experimental Data. National Bureau of Economic Research, 1998. Hoffman, Julien I.E. “Logistic Regression.” Biostatistics for Medical and Biomedical Practitioners, Elsevier, 2015, pp. 601–611. University of Toronto, 978-0-12-802387-7. Accessed 10 Apr. 2024.

Mincer, J. “Schooling, Experience, and Earnings”. 1974. Columbia University Press.

Mounk, Yascha. “Nothing Defines America’s Social Divide Like a College Education.” 4 Oct. 2023, https://www.theatlantic.com/ideas/archive/2023/10/education-inequality-economic-opportunities-college/675536/. Accessed 10 Apr. 2024.

OECD. “Home.” What Are the Earnings Advantages from Education? — Education at a Glance 2019: OECD Indicators — OECD iLibrary, 2019, www.oecd-ilibrary.org/sites/ab9c46ef-en /index.html?itemId=%2Fcontent%2Fcomponent%2Fab9c46ef-en.

Oreopoulos, P., et al. (2006). "Does Studying Economics Inhibit Cooperation? Evidence from Experiments in 15 Small-Scale Societies". Journal of Economic Behavior & Organization, 61(3), 297-314.

Stoltzfus, Jill. "Logistic Regression: A Brief Primer." Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine, U.S. National Library of Medicine, 18 Oct. 2011, pubmed.ncbi.nlm.nih.gov/21996075/#: :text=Basic%20assumptions%20that %20must%20be,lack%20of%20strongly%20influential%20outliers.

"Taking a Closer Look at Marital Status and the Earnings Gap." Saint Louis Fed Eagle, Federal Reserve Bank of St. Louis, 12 Aug. 2021, www.stlouisfed.org/on-the-economy/2020/september/taking-closer-look-marital-status-earnings-gap.

Xie, Yu. "Values and limitations of statistical models." Research in Social Stratification and Mobility, vol. 29, no. 3, 1 Sept. 2011, pp. 343–349, https://doi.org/10.1016/j.rssm.2011.04.001.

8 Appendix

We completed this assignment independently, but with some support from ChatGPT. We used ChatGPT to search for papers related to our research. Everything else including results, interpretations, discussions and writeup is done by us.

Please suggest economics paper relating to income and years of schooling

Table 1: Variable Descriptions

Variable	Definition	Example
age	Age (years)	38, 42, 71
workclass	8 different job categories	Private, Local-gov, Never-worked
educ	Education level	Bachelors, 9th, Preschool
educnum	Years of education	13, 9, 7
maritalstatus	Marital status	Divorced, Separated, Widowed
occupation	Occupation	Tech-support, Armed Forces, Sales
relationship	Relationship	Wife, Unmarried, Own-child
race	Race	White, Asian-Pac-Islander, Other
sex	Sex	Male, Female
hrspw	Hours worked per week	40, 50, 70
nativecountry	Native country	China, Italy, Vietnam
income	Greater, or lesser than or equal to USD $50,000	>50K, <=50K

Table 2: Specification 1 - OLS

	(1) dinc
educnum	0.0557***
	(42.61)
_cons	-0.318***
	(-23.28)
N	15171
R^2	0.107

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Specification 2 - Probit

	(1) dinc	(2)
main		
educnum	0.202***	0.0572***
	(39.23)	(45.13)
_cons	-2.815***	
	(-49.80)	
N	15171	15171

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Specification 3 - Logit

	(1) dinc	(2)
main		
educnum	0.353***	0.0578***
	(38.74)	(46.53)
_cons	-4.871***	
	(-47.49)	
N	15171	15171

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

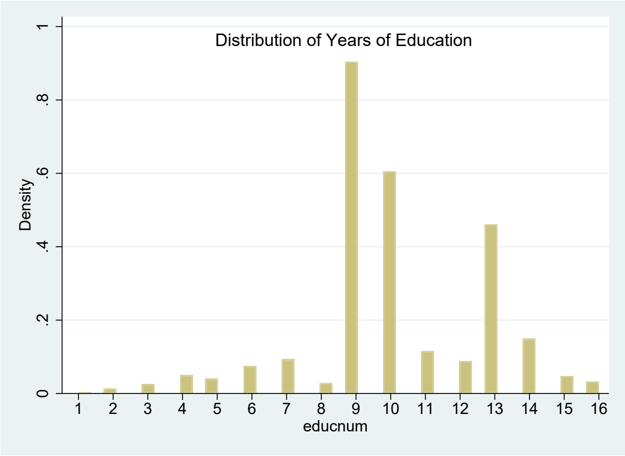


Figure 1: Distribution of Years of Education

Table 5: Specification 4 - Logit

	(1) dinc	(2)
main		
educnum	0.356^{***} (37.21)	0.0517^{***} (44.69)
1.gender	0 (.)	
2.gender	1.260^{***} (23.59)	
age	0.0443^{***} (26.03)	
1.Race	-0.650^{*} (-2.15)	
2.Race	-0.382^{**} (-2.95)	
3.Race	-0.461^{***} (-5.23)	
4.Race	-1.154^{**} (-2.75)	
5.Race	0 (.)	
_cons	-7.557^{***} (-52.29)	
N	15171	15171
R^2		

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Specification 5 - Logit

	(1) dinc	(2)
main		
educnum	0.342^{***}	0.0486^{***}
	(35.36)	(41.73)
gender	1.113^{***}	
	(20.40)	
age	0.0459^{***}	
	(26.04)	
1.Race	-0.702^{*}	
	(-2.30)	
2.Race	-0.328^{*}	
	(-2.51)	
3.Race	-0.408^{***}	
	(-4.60)	
4.Race	-1.262^{**}	
	(-2.92)	
5.Race	0	
	(.)	
hrspw	0.0323^{***}	
	(16.76)	
_cons	-9.863^{***}	
	(-51.71)	
N	15171	15171

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Specification 6 - Logit

	(1)	(2)
	dinc	
main		
educnum	0.377***	0.0461***
	(35.18)	(42.77)
gender	0.120	
	(1.83)	
age	0.0298***	
	(14.22)	
1.Race	-0.587	
	(-1.85)	
2.Race	-0.430**	
	(-3.01)	
3.Race	-0.209*	
	(-2.18)	
4.Race	-1.324**	
	(-2.95)	
5.Race	0	
	(.)	
hrspw	0.0293***	
	(13.78)	
1.maritals	-2.039***	
	(-23.72)	
2.maritals	0.279	
	(0.41)	
3.maritals	0	
	(.)	
4.maritals	-2.188***	
	(-7.46)	
5.maritals	-2.499***	
	(-31.28)	
6.maritals	-2.018***	
	(-10.75)	
7.maritals	-2.187***	
	(-10.93)	
_cons	-6.890***	
	(-33.04)	
N	15171	15171
R^2		

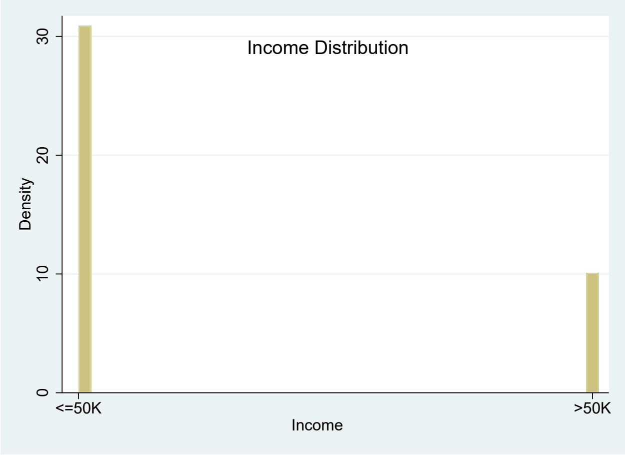


Figure 2: Income Distribution

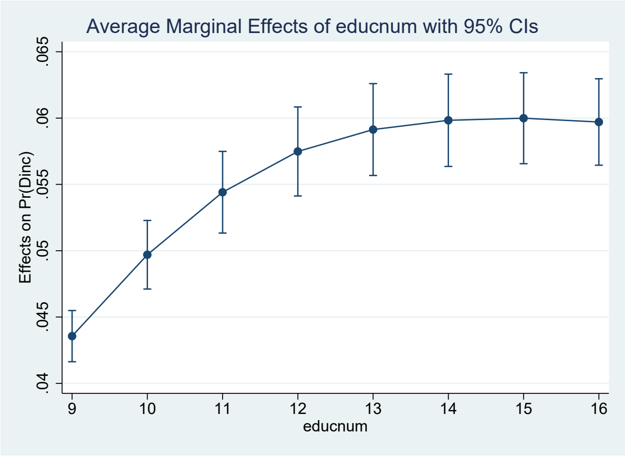
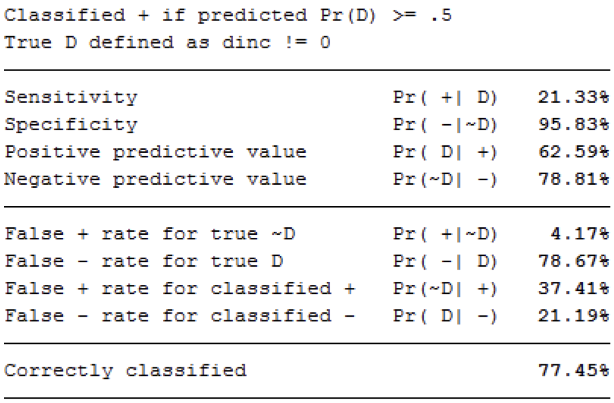
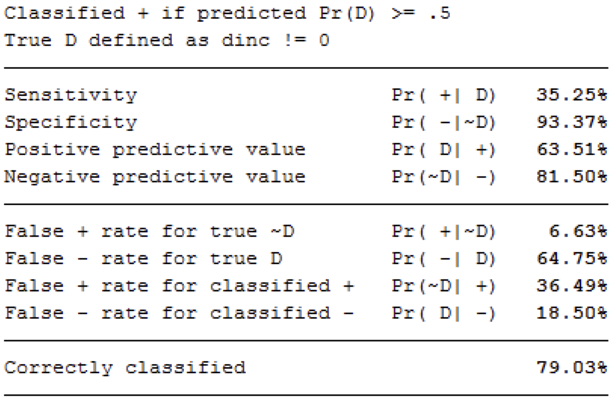


Figure 3: AME of Years of Education



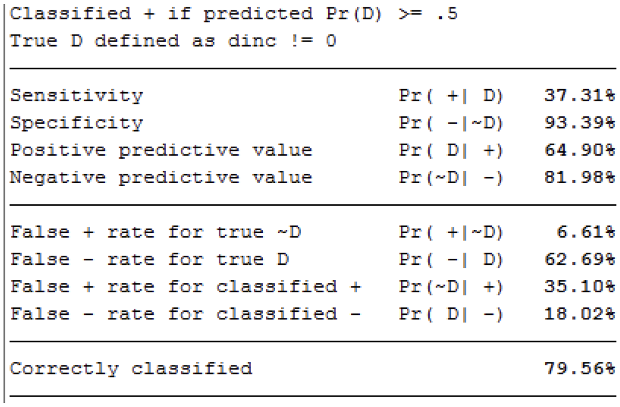
```
Classified + if predicted Pr(D) >= .5
True D defined as dinc != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   21.33%
Specificity                     Pr( -|~D)   95.83%
Positive predictive value       Pr( D| +)   62.59%
Negative predictive value       Pr(~D| -)   78.81%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    4.17%
False - rate for true D         Pr( -| D)   78.67%
False + rate for classified +   Pr(~D| +)   37.41%
False - rate for classified -   Pr( D| -)   21.19%
--------------------------------------------------
Correctly classified                        77.45%
--------------------------------------------------
```

Figure 4: Classification



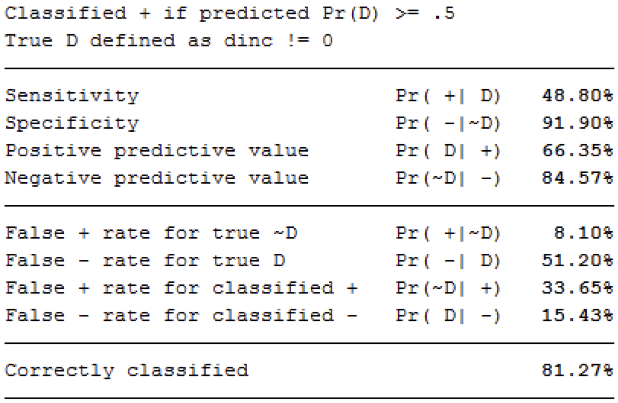
```
Classified + if predicted Pr(D) >= .5
True D defined as dinc != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   35.25%
Specificity                     Pr( -|~D)   93.37%
Positive predictive value       Pr( D| +)   63.51%
Negative predictive value       Pr(~D| -)   81.50%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    6.63%
False - rate for true D         Pr( -| D)   64.75%
False + rate for classified +   Pr(~D| +)   36.49%
False - rate for classified -   Pr( D| -)   18.50%
--------------------------------------------------
Correctly classified                        79.03%
--------------------------------------------------
```

Figure 5: Classification



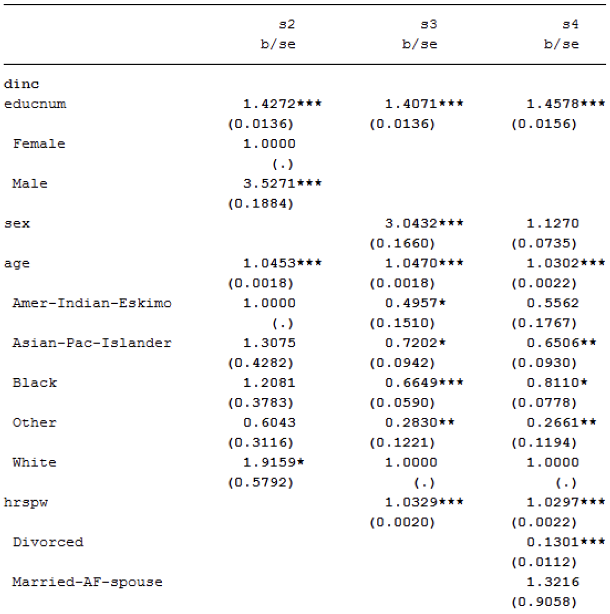
```
Classified + if predicted Pr(D) >= .5
True D defined as dinc != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   37.31%
Specificity                     Pr( -|~D)   93.39%
Positive predictive value       Pr( D| +)   64.90%
Negative predictive value       Pr(~D| -)   81.98%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    6.61%
False - rate for true D         Pr( -| D)   62.69%
False + rate for classified +   Pr(~D| +)   35.10%
False - rate for classified -   Pr( D| -)   18.02%
--------------------------------------------------
Correctly classified                        79.56%
--------------------------------------------------
```

Figure 6: Classification



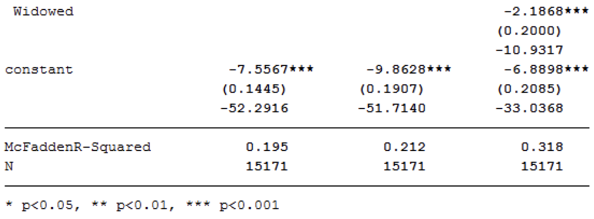
```
Classified + if predicted Pr(D) >= .5
True D defined as dinc != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   48.80%
Specificity                     Pr( -|~D)   91.90%
Positive predictive value       Pr( D| +)   66.35%
Negative predictive value       Pr(~D| -)   84.57%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    8.10%
False - rate for true D         Pr( -| D)   51.20%
False + rate for classified +   Pr(~D| +)   33.65%
False - rate for classified -   Pr( D| -)   15.43%
--------------------------------------------------
Correctly classified                        81.27%
--------------------------------------------------
```

Figure 7: Classification



	s2 b/se	s3 b/se	s4 b/se
dinc			
educnum	1.4272***	1.4071***	1.4578***
	(0.0136)	(0.0136)	(0.0156)
Female	1.0000		
	(.)		
Male	3.5271***		
	(0.1884)		
sex		3.0432***	1.1270
		(0.1660)	(0.0735)
age	1.0453***	1.0470***	1.0302***
	(0.0018)	(0.0018)	(0.0022)
Amer-Indian-Eskimo	1.0000	0.4957*	0.5562
	(.)	(0.1510)	(0.1767)
Asian-Pac-Islander	1.3075	0.7202*	0.6506**
	(0.4282)	(0.0942)	(0.0930)
Black	1.2081	0.6649***	0.8110*
	(0.3783)	(0.0590)	(0.0778)
Other	0.6043	0.2830**	0.2661**
	(0.3116)	(0.1221)	(0.1194)
White	1.9159*	1.0000	1.0000
	(0.5792)	(.)	(.)
hrspw		1.0329***	1.0297***
		(0.0020)	(0.0022)
Divorced			0.1301***
			(0.0112)
Married-AF-spouse			1.3216
			(0.9058)

Figure 8: Classification



Widowed			-2.1868***
			(0.2000)
			-10.9317
constant	-7.5567***	-9.8628***	-6.8898***
	(0.1445)	(0.1907)	(0.2085)
	-52.2916	-51.7140	-33.0368
McFaddenR-Squared	0.195	0.212	0.318
N	15171	15171	15171

* p<0.05, ** p<0.01, *** p<0.001

Figure 9: Classification