Impact of Internet Accessibility on Earnings among Uneducated Population:

An Analysis of Current Population Survey Microdata, 2015

Shuto Araki

DePauw University

November 26, 2017

Abstract

In today's information society, having Internet access enables people to utilize high quality information and often lead to more opportunities in life. This paper focuses on the uneducated with 13 or less years of education in order to find whether the Internet is serving as a new platform on career development and how it can overcome the lack of education. The empirical analysis concludes that having Internet access at home leads to 0.9-8.9% increase in income, holding other variables constant. Comparing to 10-15% income increase with Internet access at home in 1984, having Internet access does not seem to have much economic impact any more.

I. Introduction

In 2017, we *google* everything. Wherever on earth you are, you are always connected to the entire world as long as you have a smartphone or laptop that is connected to the Internet. Any kind of question can be answered through the Internet. We have never had more information on our fingertips than ever before in human history. For example, Massive Open Online Courses (MOOCs) are becoming more and more popular worldwide and the underprivileged who cannot afford good education can now take courses from high level institutions such as Harvard and MIT. However, this platform requires the Internet access. What about those who don't have Internet access? What kind of impact does it have? There are significant number of people who does not have access to the Internet at home in the United States today: those who have less years of education. According to the Current Population Survey (CPS) data from 2015, only two thirds of people who have 13 or less years of education had Internet access at home.

I was curious whether Internet accessibility makes a significant difference to income of people with less years of education. Utilizing the resources online, some people might be able to compensate their lower level of education. An extreme example would be Mark Zuckerberg, a co-founder and current CEO of Facebook, Inc. Even though he dropped out of college, he is a billionaire because the Internet provided him an opportunity to start his own company. Although Zuckerberg may be an outlier, Internet accessibility could positively affect some uneducated population's income by taking online courses or utilizing web market opportunities, for instance.

II. Literature Review

Evidence from prior research shows that there are "robustly significant positive associations between Web use and earnings growth" (DiMaggio and Bonikowski, 2008, p. 227).

This result partially answers this paper’s question: How much does Internet accessibility affect the uneducated population in terms of income? DiMaggio and Bonikowski (2008) used OLS regression with microdata from CPS to measure the effect of Internet use on income. Although this research did not filter the population with levels of education to focus on the uneducated, the outline of the empirical research is close to this paper. By creating a dummy variable whether a person used the Internet in 2000 and 2001 and by controlling for “race and Hispanic ethnicity, gender, age (and age squared), marital status, educational attainment, region of residence, and metropolitan residence, … industry and occupation,” the study found that “the median earner who used the Internet in both years was paid $.96 per hour more than a comparable nonuser” (DiMaggio and Bonikowski, 2008, pp. 238-239). This result suggests that Internet accessibility affects income positively and matches with their theories that will be introduced in the next section. However, the data used in this analysis are from 2000 and 2001, which is outdated considering the technological advancement in the past decade such as smartphones. Thus, the result of my empirical study should be unique, which makes this paper worth studying.

Young (2006) approaches this paper’s question from educational perspective. He studied the effect of Internet use on the academic performance among high school students. Since income and Internet use (DiMaggio and Bonikowski, 2008), and, income and educational level (Mincer, 1958) are respectively positively correlated, and higher academic performance predicts higher educational level, it seems that academic performance and Internet use also have positive association. However, Internet use in general did not have a significant impact on student’s academic performance measured by GPA (Young, 2006). The regression analysis shows that the influence of Internet use on academic performance depends on its purpose. The result reports that, with P-value < 0.05, Internet use for entertainment negatively predicts academic performance. It is interesting to see that, even with learning oriented usage of the Internet, there is no statistically significant relationship between Internet use and academic performance for high school students. However, one pitfall in this study as an application to this paper might be that the sample data are drawn from high school students. Their behaviors are likely to be different from those of adults. Therefore, it would be meaningful to analyze the latest data from the Current Population Survey to measure the effect of Internet use among uneducated population and see if the Internet could be a new platform for education.

In terms of how much computer usage affects earnings among U.S. workers, Krueger (1993) conducted an empirical study using the Current Population Survey data from 1984 and 1989. The study shows that workers who use computers on their job earn approximately 10 to 15 percent higher wages on average. Furthermore, since more highly educated people tend to use computers especially in 1980's when computer and the Internet were not as common as today, the estimates suggest that rapid increase of computer use at that time "can account for between one-third and one-half of the increase in the rate of return to education observed between 1984 and 1989" (Krueger, 1993, p. 55).

III. Theoretical Analysis

There are 3 different theories in how Internet use increases one's earnings: Human Capital theory, Social Capital theory, and Cultural Resource and Signal theory (DiMaggio and Bonikowski, 2008). The result of empirical analysis will be explained under the light of these theories.

There are several obvious advantages in using the Internet such as access to high quality information, faster communication, and exposure to learning opportunities. These features increase productivity significantly (Krueger, 1993). In economics, society's productivity is measured by the following function:

$$\frac{Y}{L} = A \cdot f(\frac{K}{L}, \frac{H}{L}, \frac{N}{L})$$

Y: quantity of output, A: technological knowledge, H: quantity of human capital,
K: quantity of physical capital, N: quantity of natural resources, L: quantity of labor

The function *f* shows how the inputs are combined to produce output. Division by quantity of labor creates output per worker i.e. productivity. Technological knowledge is "the understanding of the best ways to produce goods and services" (Mankiw, 2013, pp. 530). The Internet has changed the way some firms produce their goods and services. For example, taxi drivers can now pick up their customers more efficiently by specifying their locations using the Internet. Since the extent to which one's job uses the Internet depends upon the job and its industry, occupation and industry have to be controlled in empirical analysis. According to Human Capital Theory, Internet use affects both technological knowledge and human capital per worker because of the features of the Internet mentioned earlier. Firms may also invest in human capital when they implement new technologies by training their employees (Fernandez, 2001).

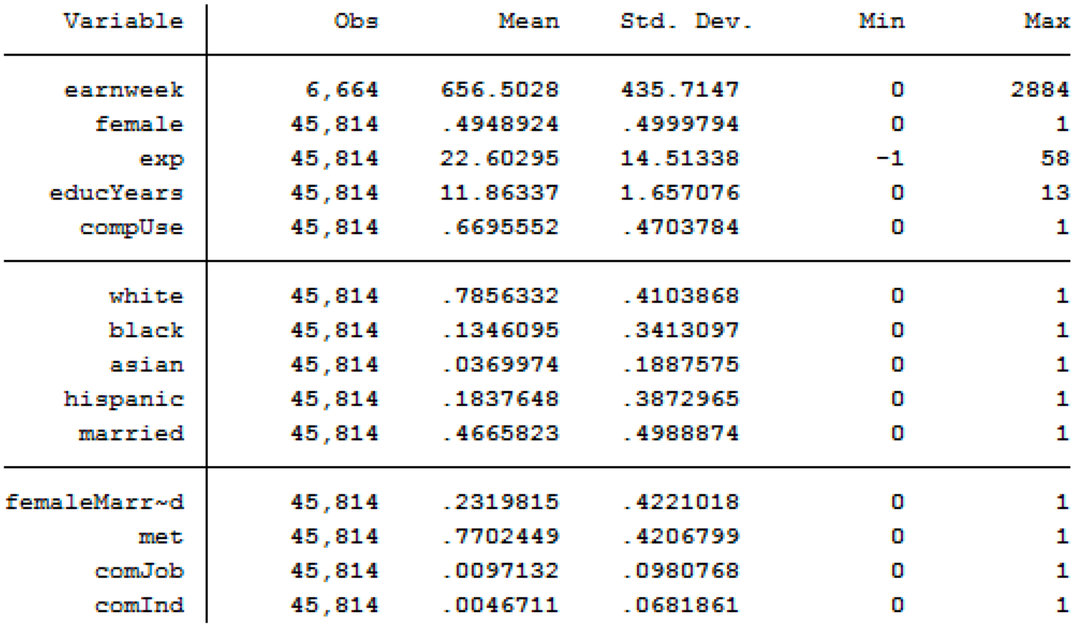
Another explanation for the positive association between Internet use and earnings approaches from looking at Internet use as a source of social capital. There are three kinds of social-capital enhancement. First, workers with Internet access can use the Internet as a means of searching for jobs. Online job search is becoming more and more popular. Among those who are unemployed and looking for work, 76.3% of them utilized the Internet to search for jobs in 2011 in contrast to 25.5% in 2000 (Faberman and Kudlyak, 2016). Such workers have more opportunities to apply for more jobs. Second, workers who use the Internet may be exposed to more networking opportunities. Websites such as linkedin.com enables them to connect with their potential employers and expand their professional network. The last improvement in social capital is that “employees with large, accessible professional networks may use technology to employ these networks in ways that benefit their employers: for example, getting useful information, contacting clients, or setting up collaborative ventures” (DiMaggio and Bonikowski, 2008, p. 232).

The last theory explains how cultural resource and signal can explain Internet use yields higher income. Especially in early 2000’s, which is a few years after the Internet became accessible to regular Americans, many of them “regarded the Internet as a transformative force that would ignite explosive economic growth” (DiMaggio and Bonikowski, 2008, p. 233). There is prior evidence in some employers using Internet job postings to filter out low-quality applicants, considering that Internet users are more able. Such evidence suggests that mere familiarity with the technology functioned as a signal where employers judge whether an applicant was competent and intelligent. Cultural resource theory describes that familiarity with high-status activities like the Internet (at least a few decades ago) puts a person into a favorable position in professional settings because he/she can form relations with high-status others easier than those who don’t use the Internet. However, this theory might not be applicable today because so many more people use the Internet, which is no longer considered high-status in 2017.

IV. Empirical Analysis

IV. i. The Data

Table 1. Summary Data Table



```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    earnweek |      6,664    656.5028    435.7147          0       2884
      female |     45,814    .4948924    .4999794          0          1
         exp |     45,814    22.60295    14.51338         -1         58
   educYears |     45,814    11.86337    1.657076          0         13
     compUse |     45,814    .6695552    .4703784          0          1
-------------+---------------------------------------------------------
       white |     45,814    .7856332    .4103868          0          1
       black |     45,814    .1346095    .3413097          0          1
       asian |     45,814    .0369974    .1887575          0          1
    hispanic |     45,814    .1837648    .3872965          0          1
     married |     45,814    .4665823    .4988874          0          1
-------------+---------------------------------------------------------
femaleMarr~d |     45,814    .2319815    .4221018          0          1
         met |     45,814    .7702449    .4206799          0          1
      comJob |     45,814    .0097132    .0980768          0          1
      comInd |     45,814    .0046711    .0681861          0          1
```

Source: Current Population Survey Computer and Internet Use Supplement July 2015, employed persons, ages 18 to 64.

Before comparing different regression models that analyzes Internet accessibility and income, some independent variables need to be described along with the summary data in *Table 1*. All the data were collected from "the Current Population Survey (CPS), a monthly household survey fielded continually by the Bureau of the Census and based on stratified probability samples of the non-institutionalized U.S. population" (DiMaggio and Bonikowski, 2008, p. 235). Since the data generation process by this complex survey design does not meet the requirements of Classical Econometric Model (simple random sampling), not only OLS (Ordinary Least Squares) but also PWLS (Probability Weighted Least Squares) regression was utilized in this analysis. The data were sampled from the latest available CPS Computer and Internet Use Supplement from July 2015. This paper selects variables according to DiMaggio and

Bonikowski (2008) and Krueger (1993). The theories behind the selection of some of the variables were explained in the previous section. Only sample with age 18 to 64 were selected because this is the typical labor force. This study omits those who have more than 13 years of education because the subject of this empirical analysis is the uneducated. The next section explains why 13 years was used as a threshold in detail. Those who were categorized as "Not In Universe" for Internet accessibility question were dropped as well because this is a key variable that will be included in every regression.

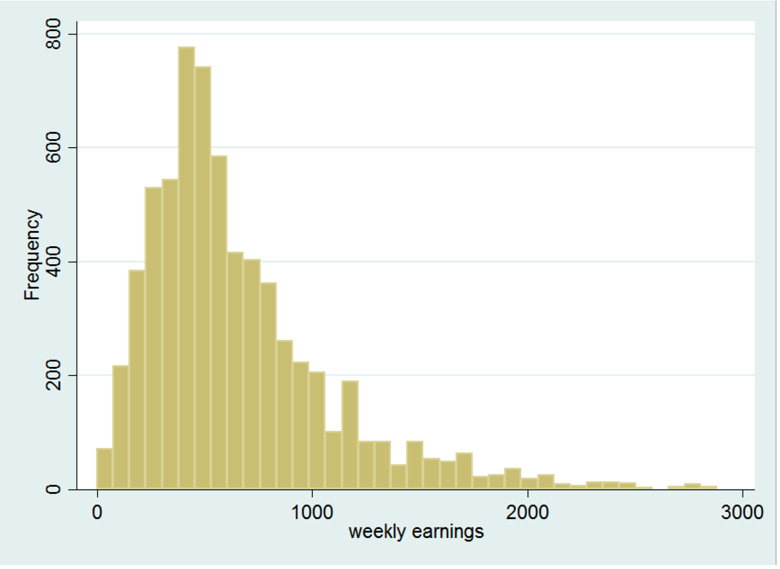
Table 1 summarizes the variables used in regression analysis. Occupation and industry variables are encoded for this table to show how much of the sample have computer- or Internet-related job or industry ("comJob" and "comInd" variables, respectively). Note that only a very small portion of the sample has computer or Internet related jobs or is in such field. It is reflected in the percentage of Internet accessibility as well: there are only 67% in the sample who use the Internet at home. The most skewed data among all the variables is years of education. As shown in *Figure 2.*, majority (88.77%) of the sampled individuals have a high school diploma or one year of college experience. Thus, this study's "uneducated" sample is heavily weighted toward high school graduates and college dropouts.

Utilizing a statistical software package STATA, the raw data was encoded by creating dummy variables that could correlate with the most important independent variable in this study: Internet use. Control variables are experience, experience squared, years of education, female, race variables, marital status, female and married (interaction term), living in the city, occupation, industry, and accessibility to the Internet at home. Experience variable was created by age minus schooling minus 6 (the typical age to beginning school in the U.S.) and quadratic term was also included (Mincer, 1958). Income versus experience relationship is curved because too little experience would yield lower income, but too much experience or being too old would reduce one's earnings. Also, even though this study focuses on uneducated population, years of education still varies (SD = 1.65) and therefore it has to be controlled in regression analysis. For the independent variable, according to the Human Capital Theory, natural log should be applied to income (Mincer, 1958). The main point is that "[e]quilibrium pay differentials for jobs with different levels of training are constant *multiples* of each other" (Barreto and Howland, 2005). Using the semilog functional form also has an advantage of reducing heteroscedasticity in the data.

One of the limitations in this sample data is that the CPS Computer and Internet Use Supplement does not include the traditional annual income variable. The only available approximation of it is weekly earnings. There are two limitations to this particular variable. One is that it excludes non-wage/salary workers and self-employed persons. Some of the most innovative individuals such as Mark Zuckerberg are not part of the data. They do form their own career without college education and therefore this exclusion might affect the significance of this paper. Another limitation that could affect the result of this study is that it compresses all the individuals who make more than $2,884 per week into one category: 2884.61. The value was replaced with period (.), which signifies "missing value" in STATA. They were treated as missing values because dropping them would allow all the entire row that fit into the category in the earnings column. Thus, for example, some data will not be available even if income data is not used in a regression just because they happened to earn higher than $2,884 per week. This omitted variable bias is probably one of the biggest source of chance error in this data generation process.

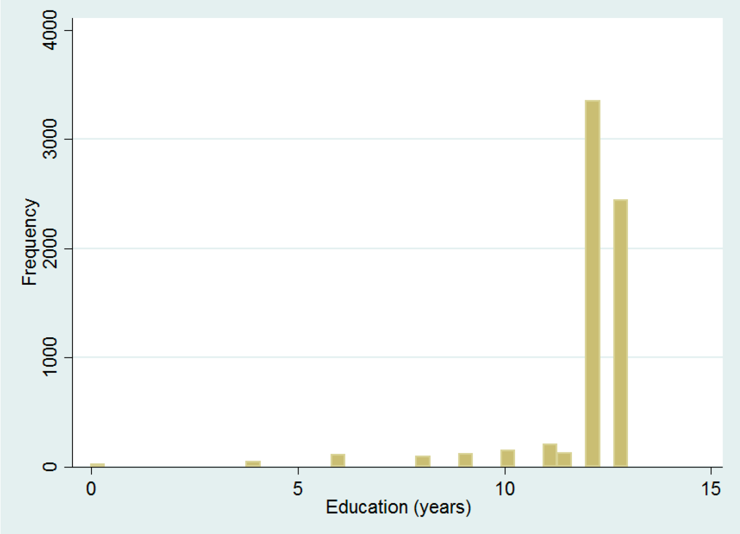
Figure 1. below shows the weekly income frequency distribution in the sample. The distribution is skewed to the right with skewness of 1.59 even though the right tail was cut off at $2,884 because of the limitations of data mentioned above. This fact shows that the sample used in regression analysis represents the general population well in terms of income because income distribution in general is right skewed. This is a valid concern to check because the CPS is, as mentioned, oversampled in some categories.

Figure 1. Weekly Income Frequency Distribution among sample with 13 or less years of education



Source: Current Population Survey Computer and Internet Use Supplement July 2015, ages 18 to 64.

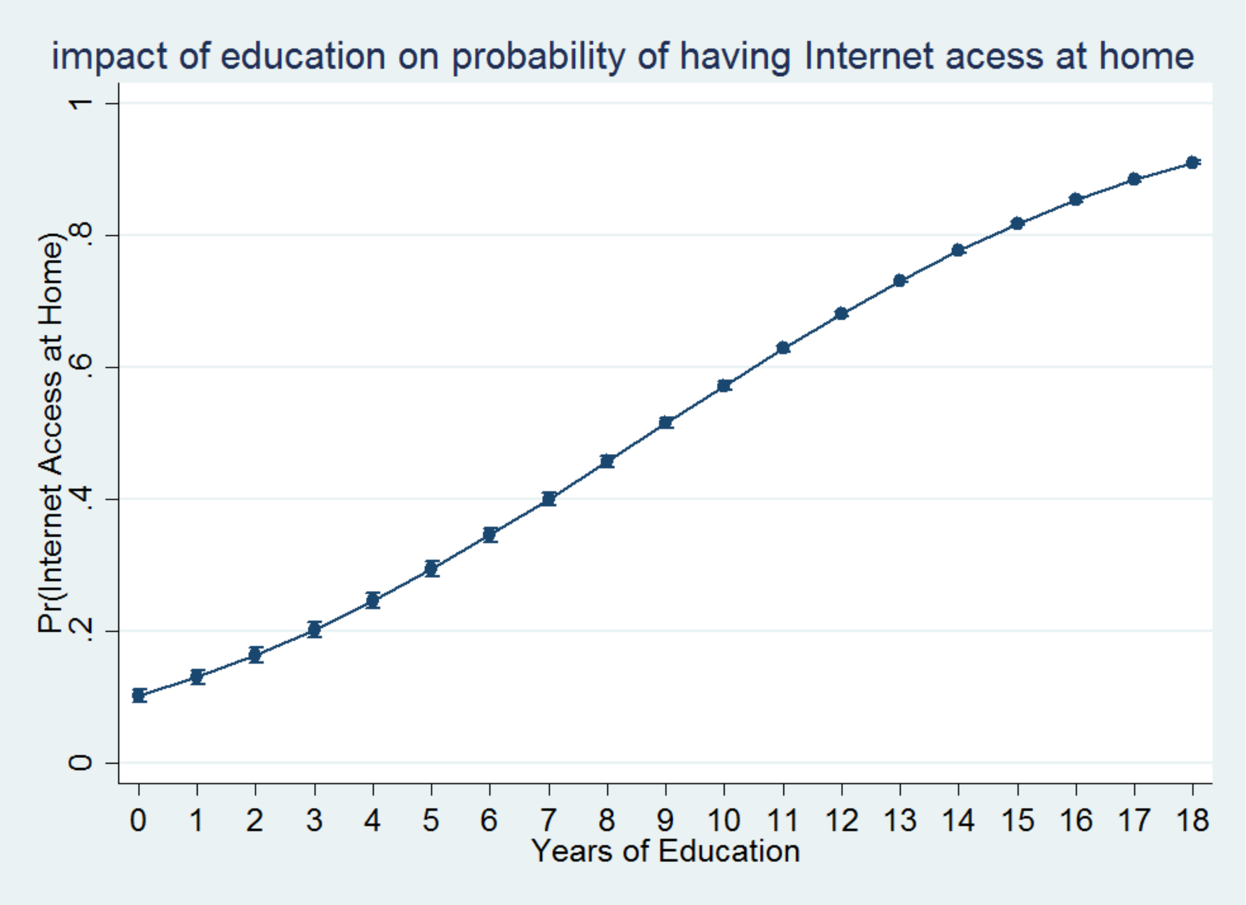
Figure 2. Frequency Distribution of Years of Education



Source: Current Population Survey Computer and Internet Use Supplement July 2015, ages 18 to 64.

IV. ii. Who Doesn't Have Internet Access at Home?

Figure 3. The Impact of Years of Education on the Probability of Having Internet Access at Home



Source: Current Population Survey Computer and Internet Use Supplement July 2015, employed persons, ages 18 to 64.

In order to determine who should be considered "uneducated" in this study, a probit (probability unit) model was utilized on probability of having Internet access at home. As seen in *Figure 3.*, the fitted line is slightly curved. Until 13 years, additional year of education increases the probability of having Internet access at home by similar amount. After 13 years, the rate of increase in probability declines. Additional year of education does not make people install Internet access at home as much as those who are less educated. Therefore, in the context of Internet access at home, it is reasonable to draw a line between uneducated and educated at 13 years of education.

IV. iii. OLS and PWLS Regression Analysis of Internet Use on Earnings

Table 2. Regression Results for Internet Use at Home

	(1)	(2)	(3)	(4)	(5)
	OLS	OLS	OLS	OLS	PWLS
	ln(earnings)	ln(earnings)	ln(earnings)	ln(earnings)	ln(earnings)
Internet Access at home (1 = yes)	0.0498**	0.0195	0.0596***	**0.0331***	**0.0454***
	(0.0188)	(0.0192)	(0.0169)	(0.0163)	(0.0186)
Years of Education		0.0413***	0.0628***	0.0399***	0.0404***
		(0.00551)	(0.00534)	(0.00534)	(0.00668)
Experience			0.0479***	0.0329***	0.0344***
			(0.00227)	(0.00225)	(0.00263)
Experience^2			-0.000766***	-0.000523***	0.000551***
			(0.0000469)	(0.0000458)	(0.0000551)
Female			-0.295***	-0.188***	-0.165***
			(0.0207)	(0.0225)	(0.0253)
Female and Married			-0.152***	-0.0981***	-0.0897**
			(0.0298)	(0.0293)	(0.0339)
24 Race dummies included	No	No	Yes	Yes	Yes
16 Hispanic dummies included	No	No	Yes	Yes	Yes
7 Marital Status dummies included	No	No	Yes	Yes	Yes
5 Metropolitan dummies included	No	No	Yes	Yes	Yes
568 Occupation dummies included	No	No	No	Yes	Yes
245 Industry dummies included	No	No	No	Yes	Yes
_cons	6.239***	5.766***	5.224***	6.447***	6.491***
	(0.0159)	(0.0651)	(0.104)	(0.241)	(0.258)
Adjusted R^2	0.000902	0.00911	0.254	0.3816	0.3952
N	6662	6662	6662	6662	6662

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In total, five regressions were completed. The model (1) was run to measure raw differential. The model (2) includes Internet use and years of education. The model (3) adds core demographic control variables to it. The model (4) includes 828 more variables on occupation and industry. Lastly, the model (5) utilizes PWLS regression instead of OLS. As shown in the bottom row of *Table 2.*, the model (5) has the highest adjusted R^2.

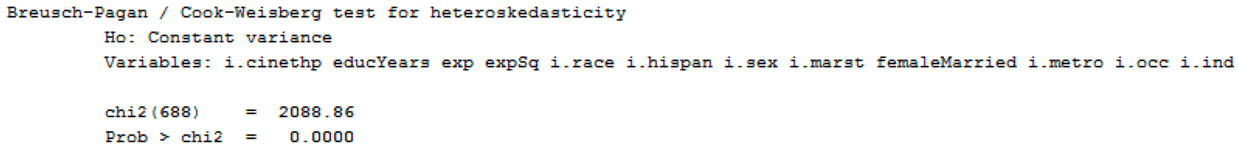
Adjusted R^2 measures the goodness-of-fit of a regression by penalizing the number of predictors. The more variables in a regression, the heavier the penalty so as to prevent so called "kitchen sink regression." Using many variables in regression analysis will increase the goodness-of-fit measured by a regular R^2 because a number of predictors can be used to explain its dependent variable. That is one of the problems that R^2 has; It does not decrease as you increase the number of predictors. However, such regressions often times over-fit the sample. Over-fitting sample data is a problem because the model is not flexible enough to predict trends in population data. It produces misleadingly high R^2 values with less ability to predict. The model 5 outperforms other 4 regressions despite its having so many variables. The regular R^2 value was 0.458, which was heavily penalized in the adjusted R^2. Thus, as expected, the model (5) is the best model in this paper, and the values will be referred to. It is also important to note that the model (4) is just as good a regression as the model (5) because their adjusted R^2's are close. In this regression, PWLS and OLS seem to perform equally well. It is interesting to see that the model (5) has slightly positive experience squared term where, in theory, it is supposed to be negative (Mincer, 1958).

Another process in the analysis that should be noted about these regression results is that STATA omitted 15 industry variables because of collinearity in the model (4) and (5). Omitted variables include had several categories that have similar features such as beauty salons and nail salons.

According to the model (5), the standard error of the estimated coefficient for Internet access at home is 0.0186. The exact coefficient estimate is $e^{0.0454} - 1 \approx 0.046$ because the dependent variable is wrapped with natural log. Anti-log operation has to be applied on the coefficient estimate in order to appropriately interpret this result. Since 0.0186 * 1.96 = 0.0365, 95% of confidence interval constructed as 4.6 ± 3.65% will cover the true parameter value about the impact of Internet access at home on weekly earnings. Even though the precision of the coefficient estimate is not as good, it is important to check that having Internet access at home increases earnings more than an additional year of education does on average. This result signifies how crucial it is to have Internet access at home in terms of income, holding years of education, experience, core demographics, occupation, and industry constant. Again, however, keep in mind that these results exclude the richest group in the sample because of the limitations mentioned in the previous section. Thus, the true parameter might be different from these results.

Lastly, Breusch-Pagan test was run on the model (4) in order to check the presence of heteroscedasticity. STATA returns an extremely small P-value, which means that the null hypothesis (homoscedasticity) was rejected with significant evidence. Since heteroscedasticity is a violation of identical distribution of error term in Classical Econometric Model, the standard errors reported in *Table 2.* should be broken. Thus, the results need to report robust standard errors. Yet, STATA reports almost identical standard errors possibly because of the large sample size and the use of natural log on earnings variable. Therefore, robust standard errors will not be reported on this paper.

Breusch-Pagan test on STATA



```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: i.cinethp educYears exp expSq i.race i.hispan i.sex i.marst femaleMarried i.metro i.occ i.ind

         chi2(688)    =  2088.86
         Prob > chi2  =   0.0000
```

Some might argue that significance of having Internet access at home depends on the line between the educated and the uneducated. In this paper, because of the aforementioned reasons from the probit analysis, 13 years of education was chosen as the borderline. Yet, this criticism is valid because the argument for choosing 13 years is not necessarily a strong one. Thus, additional five regressions were completed on the same sample but with the borderline at 12 years of education.

As seen in *Table 3.*, limiting the sample to individuals with 12 or less years of education increases the significance of having Internet access at home. Similar to the previous regression results, the model (5) performs best. The estimated coefficient is $e^{0.0736} - 1 \approx 0.0764$. Thus, having Internet access at home predicts 7.64% point higher earnings on average, holding other variables constant. This time, the level of significance increased because of the smaller standard error. Note that, since the data is more heteroscedastic, robust standard errors were reported for these regression results. The 95% confidence interval is [3.25%, 12.03%].

Compared to the previous regression results when the sample included those with 13 years of education, it reports much higher impact on weekly earnings. Therefore, it seems that the impact of having Internet access at home is more significant for those with less years of education.

Table 3. Regression Results for People without College Education

	(1)	(2)	(3)	(4)	(5)
	OLS	OLS	OLS	OLS	PWLS
	ln(earnings)	ln(earnings)	ln(earnings)	ln(earnings)	ln(earnings)
Internet Access at home (1 = yes)	0.0746***	0.0420	0.0729***	0.0537**	**0.0736****
	(0.0218)	(0.0221)	(0.0199)	(0.0196)	(0.0224)
Years of Education		0.0487***	0.0551***	0.0387***	0.0407***
		(0.00611)	(0.00611)	(0.00608)	(0.00791)
Experience			0.0431***	0.0275***	0.0294***
			(0.00284)	(0.00287)	(0.00350)
Experience^2			-0.000696***	-0.000438***	-0.000470***
			(0.0000570)	(0.0000570)	(0.0000708)
Female			-0.337***	-0.196***	-0.159***
			(0.0266)	(0.0298)	(0.0331)
Female and Married			-0.0784*	-0.0317	-0.0354
			(0.0371)	(0.0373)	(0.0424)
24 Race dummies included	No	No	Yes	Yes	Yes
16 Hispanic dummies included	No	No	Yes	Yes	Yes
7 Marital Status dummies included	No	No	Yes	Yes	Yes
5 Metropolitan dummies included	No	No	Yes	Yes	Yes
568 Occupation dummies included	No	No	No	Yes	Yes
245 Industry dummies included	No	No	No	Yes	Yes
_cons	6.215***	5.682***	5.376***	6.440***	6.542***
	(0.0177)	(0.0691)	(0.124)	(0.300)	(0.270)
Adjusted R^2	0.00252	0.171	0.232	0.368	0.384
N	4218	4218	4218	4218	4218

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

V. Conclusion and Further Research

The purpose of this econometric analysis of Internet use among uneducated population was carried out with the intention of determining the impact of having Internet at home on earnings in 2015.

The empirical analysis using CPS data from 2015 predicts that having Internet access at home does have a statistically significant positive impact on earnings. This finding aligns with prior studies introduced in the Literature Review section. However, considering the variability in the coefficient estimate and Krueger's (1993) finding that there was 10-15% increase in 1984, Internet use at home does not have much economic significance any more. One of the background factors behind this conclusion might be the advent of smartphones. Having data plan on one's smartphone might be sufficient to get enough beneficial online information.

Thus, further research can be conducted with data on possession of smartphones and how it fixes unequal distributions of Internet access. I believe that this is a part of omitted variable biases that I did not have control over in this paper given the dataset does not contain any information about smartphones.

In order to respond to a possible objection against the choice of 13 years of education as a line dividing the uneducated and the educated, the second set of regressions were completed by the same data except for those who had 13th year of education being dropped. The analysis found that Internet access at home had much larger significance on those with high school diploma or less. This finding also highlights the importance of some college education even for a year. It can be inferred from this information that the less educated individuals could have higher probability of having Internet access only through the Internet.

VI. References

Barreto, H. and Howland, F. M. (2005). *Introductory Econometrics: Using Monte Carlo Simulation with Microsoft Excel®*. New York: Cambridge University Press

DiMaggio, P. and Bonikowski, B. (2008). "Make Money Surfing the Web? The Impact of Internet Use on the Earnings of U.S. Workers." *American Sociological Review*, Vol. 73(2), pp. 227-250.

Faberman, J. R. and Kudlyak, M. (2016). "What does online job search tell us about the labor market?" *Economic Perspectives*, Federal Reserve Bank of Chicago, No. 1, January 2016.

Fernandez, R. M. (2001). "Skill-Biased Technological Change and Wage Inequality: Evidence from a Plant Retooling." *American Journal of Sociology*, Vol. 107(2), pp. 273-320.

Flood, S., King, M., Ruggles, S., and Warren, R. J. *Integrated Public Use Microdata Series, Current Population Survey: Version 5.0* [dataset]. Minneapolis, MN: University of Minnesota, 2017. https://doi.org/10.18128/D030.V5.0

Jackson, L. A., Eye, A., Biocca, F. A., Barbatsis, G., and Zhao, Y. "Does Home Internet Use Influence the Academic Performance of Low-Income Children?" *Developmental Psychology*, Vol. 42(3), pp. 429-435.

Krueger, A. B. (1993). "How Computers Have Changed the Wage Structure: Evidence from Microdata, 1984-1989." *The Quarterly Journal of Economics*, Vol. 108(1), pp. 33- 60.

Mankiw, G. N. (2013). *Principles of Economics*. Stamford: Cengage Learning

Mincer, J. (1958). "Investment in Human Capital and Personal Income Distribution." *The Journal of Political Economy*, Vol. 66(4), pp. 281–302.

Young, B. (2006). "A Study on the Effect of Internet Use and Social Capital on the Academic Performance." *Development and Society*, Vol. 35(1), pp. 107-123.