The Effects of Education on Income

Kieran O'Hanlon

National College of Ireland

x21190020@student.ncirl.ie

Keywords - Age in years (age), Years of education (yrsed) Level of education (edcat), 1=Did not complete high school, 2=High school degree, 3=Some college, 4=College degree, 5=Postgraduate degree, Years with current employer (yrsempl), Credit card debt in thousands (creddebt), Other debt in thousands (othdebt), Ever defaulted on a bank loan (default) 0=no, 1=yes, Job satisfaction (jobsat) , 1=Highly dissatisfied, 2=Somewhat dissatisfied, 3=Neutral, 4=Somewhat satisfied, 5=Highly satisfied, Homeownership (homeown) 0=rent, 1=own Years at current address (address), Number of cars owned/leased (cars), Value of primary vehicle (carvalue)

Introduction

The scope of this assignment was to explore the effects YRSED, AGE, EDCAT, YRSEMPL, CREDDEBT, OTHERDEBT, DEFAULT, JOBSAT, HOMEOWN, ADDRESS, CARS, CARVALUE had on the dependent variable INCOME. The respondents were aged 18-79, with varying levels of education to try and determine if there is an effect of education on income earned. All things being equal, education in theory should have a positive effect on income as this is an increase in human capital [1]. It is possible YRSEMPL could have a negative impact on years of education as the respondent might have fewer years of formal education. Furthermore, the assumption of homoscedasticity is likely not to hold as many respondents with a higher level of education would be highly dispersed from the mean, which in turn would lead to the error term being highly varied. Because of the non-normal data, heteroskedasticity, and non-linear relationship a log-regression was decided as the best model to work with,

Literature Review

With an ever-increasing knowledge society does a higher education affect income within a household? Not only is an individual's standing in society affected by education but it is also affected by the labour market. Economic growth is driven by higher education [2] but according to Fields 1980, income inequality does not diverge substantially, dependent on how much each country invests in education for the public. A significant effect on income inequality is an investment in education with an ever-increasing gap growing as the education level rises. According to Alves 2012, the biggest impact on income inequality is an investment in early education [3]. Another study by Jamison et al, from a study of up to 62 countries did find a significant relationship between education and income growth rates [4]. The standard deviation of higher test performance yielded between .5-.9 percent higher income rates in one test and more open economies produced greater output and productivity because of this [4]. Yang 2016, analyses his research through innate ability, compulsory education which is grades 1 through 9, and non-compulsory which are grades from ten onwards, and wanted to study the effects of income on intergenerational mobility [5]. What Yang discovered was that investment in education had a substantial impact on income inequality and as education increased this income equality gap also increased.

Data & Methodology

This analysis was completed using SPSS, MS Excel, and Python. The first step of the process was to study all the data to identify variables of interest and omit any missing data which might have a significant impact on the study. The data was found sufficient and good to work with. The key variables of the study that were identified were income, years of education, the number of years employed, credit card debt, whether the respondent owned a home and the value of the respondent's car. I did run a z-test against all variables to determine the outliers but I decided to only remove four outliers from the dataset as their income was far greater than the general population. However, natural variation can produce outliers and this will be more prominent in large datasets and more likely to be of the natural population that is being studied. A quick look at the descriptive statistics illustrates there are 4,504 observations in the study, and the average age per respondent is 47 with a standard deviation of 17.6. The average respondent has almost 15 years of education with 50.6% educated to high school level and 29.1% with a college or postgraduate degree. The standard deviation for years of education is 3.2

indicating the data is centered around the mean. The average income is $54,000 with a standard deviation of almost $51,000 indicating their income is dispersed from the mean. 63% of the respondents own their home, while the average respondent has a credit card debt of $1,800 while the average value of the respondent's car is $26,000 with a standard deviation of almost $21,000. Again the standard deviation for both these variables is high similar to income. This would indicate some relationship as the averages are well dispersed from the general population.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
age	4504	18	79	46.92	17.671	312.263
yrsed	4504	6	23	14.53	3.283	10.776
edcat	4504	1	5	2.67	1.213	1.471
yrsempl	4504	0	52	9.71	9.652	93.153
income	4504	9	575	54.68	50.765	2577.086
creddebt	4504	.000000	48.704524	1.85002538	2.964188826	8.786
othdebt	4504	.000000	53.838400	3.62168443	4.651879770	21.640
default	4504	0	1	.24	.426	.182
jobsat	4504	1	5	2.96	1.378	1.898
homeown	4504	0	1	.63	.483	.234
address	4504	0	57	16.37	12.371	153.042
cars	4504	1	8	2.37	1.158	1.341
carvalue	4504	2.2	99.2	26.026	20.7855	432.038
Valid N (listwise)	4504					

The data I will work with was supplied by my lecturer Himanshu Rathee. My initial analysis was to perform a log regression with INCOME as the dependent variable. I selected all variables and through SPSS I was able to split the categorical variables EDCAT, DEFAULT, JOBSAT, and HOMEOWN into dummy variables. The initial model had an R-square of 83.1% which indicated the model was a good fit. However, many of the variables in the model were proven to be statistically insignificant. These variables included all categories for the dummy variable EDCAT, ADDRESS, CARS, HOMEOWN as well as highly satisfied and highly dissatisfied both from the JOBSAT variable. However, at 18 degrees of freedom the analysis of variance (ANOVA) with a P-value of 0.00 satisfies the assumption, there is a statistically significant relationship between the dependent and independent variables. Although this might not seem intuitive it does imply the model is significant but just not enough evidence to prove any particular variable is significant.

Results

One area I wanted to focus on is the relationship between years of education and income. It would seem intuitive to conclude that income will rise the more years of education the respondent has on average. However, this might not necessarily hold. For example, years of experience are more likely to have a positive impact on income but on the other hand years of experience might lead to fewer years of education. As the income variable is highly skewed, I ran with a log-regression model. The R-squared and adjusted R-squared coefficients are both under 4% which would indicate a high variance meaning heteroscedasticity is evident in the model. This is not necessarily a problem, particularly with a statistically significant independent variable.

The following is the log equation, where y is the dependent variable income, a is the constant and β is the independent variable years of education.

$$\ln(y) = a + \beta x$$

$$ln(\widehat{income}) = 3.05 + .0447 * \text{YRSED}$$

The null hypothesis at an alpha level of .05 in this scenario will look like the following:

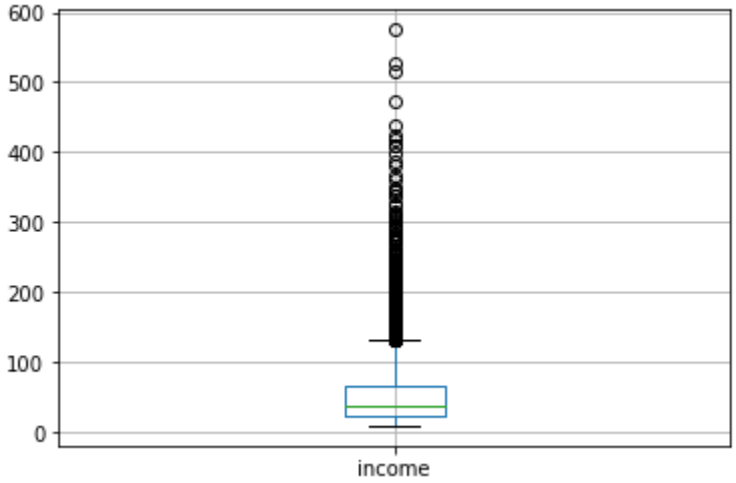
μ: years of education has no effect on income

μ1: years of education does have an effect on income

From the results with one degree of freedom, years of education is statistically significant at the 95% level and therefore we can reject the null hypothesis and state that years of education does have an association with income. The regression coefficient is .0447 and it implies that for one year increase in education income increases by 4.4%

Regression Statistics								
Multiple F	0.196005							
R Square	0.038418							
Adjusted I	0.038204							
Standard E	0.734602							
Observati	4504							
ANOVA								
	df	*SS*	*MS*	*F*	*gnificance F*			
Regressio	1	97.06412	97.06412	179.8681	3.04E-40			
Residual	4502	2429.461	0.53964					
Total	4503	2526.525						
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	3.05413	0.049665	61.49492	0	2.956763	3.151497	2.956763	3.151497
yrsed	0.044725	0.003335	13.41149	3.04E-40	0.038187	0.051263	0.038187	0.051263

Although the multiple linear regression (MLR) did bring up some statistically significant results, the fact the dependent variable INCOME was highly skewed the model would have proven to be problematic. By creating a log function of the dependent variable I was able to get a linear model. From the results, we can conclude there is a statistically significant relationship between the independent variables and the dependent variable and we can reject the null hypothesis at an alpha level of .05 with five degrees of freedom. The R-square in this model again is 82% with only 18% of the variation not explained by the model. By looking at the results, it is evident the P-values again are statistically significant at an alpha level of .05%. By using the log-linear specification, the model is telling us that for each additional year of education, income increases by 2.9%. The value of the observations car for each additional year of education increased the value by 88%. The negative coefficient for YRSEMPL is indicating for each additional year of experience income decreases by 3.4%. This is possible for a number of reasons. It is possible that experience alone won't help increase income but instead the most important determinant in income increasing is education. There are also some limitations to the model with the first being the respondents with no years of experience but respondents aged up to 79 who are most likely retired. Although, it might not seem intuitive as income rises credit card debt also rises. The rationale behind this according to Mann 2008 is that the more affluent use credit cards for convenience [6]. Being a homeowner is also statistically significant at the 95% level as higher income will most likely lead to the respondent owning their own home by 19%. The variance inflation factor (VIF) stat is below 2.0 for all coefficients indicating no multicollinearity in the model and illustrated in the appendix is the area under the curve (AUC) model. The figure of 73.2% is indicating the model is a good fit.



$$\ln(y) = \beta_0 + \beta_{1x} + \beta_{2x} + \beta_{3x} + \beta_{4x} + \beta_{5x}$$

$\widehat{\ln(y)}$ = 2.765 + .029 * YRSED + .0880 * CARVALUE -.034 * YRSEMPL + .055 CREDDEBT + .019 HOMEOWN

μ: There is no relationship between the independent variables and the dependent variable

μ1: There is a relationship between the independent variables and the dependent variable

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.908[a]	.824	.824	.31404	1.673

a. Predictors: (Constant), carvalue, homeown, yrsed, yrsempl, creddebt
b. Dependent Variable: ln_x

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2082.915	5	416.583	4223.954	.000[b]
	Residual	443.610	4498	.099		
	Total	2526.525	4503			

a. Dependent Variable: ln_x
b. Predictors: (Constant), carvalue, homeown, yrsed, yrsempl, creddebt

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	2.765	.024		115.743	.000		
	yrsempl	-.003	.001	-.034	-4.876	<.001	.802	1.247
	yrsed	.007	.002	.029	4.303	<.001	.882	1.134
	creddebt	.014	.002	.055	7.450	<.001	.715	1.398
	homeown	.030	.010	.019	3.014	.003	.976	1.025
	carvalue	.032	.000	.880	110.785	.000	.619	1.615

a. Dependent Variable: ln_x

Kruskal-Wallis

After determining the data was non-normal a Kruskal-Wallis test which is an alternative to the ANOVA was chosen. This is a rank-based non-parametric test to see if there is a statistical difference between two or more groups of an independent variable on a continuous or ordinal dependent variable. Similar to the regression models I will check the null hypothesis at an alpha level of .05.

The Kruskal-Wallis equation:

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)$$

μ_0: population medians are equal.

μ_1: population medians are not equal.

With four degrees of freedom and at the 95% level I can reject the null hypothesis that one or more of the population medians are not equal. This is because the significance level is less than .001. By looking at the output of the pairwise comparison of the education category the only categories that are not statistically significant are did not complete a school-high school degree and college degree-postgraduate degree.

Pairwise Comparisons of edcat

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
Did Not Complete School-High School Degree	-68.752	56.216	-1.223	.221	1.000
Did Not Complete School-Some College	-312.623	61.739	-5.064	<.001	.000
Did Not Complete School-College Degree	-565.450	60.642	-9.324	.000	.000
Did Not Complete School-Postgraduate Degree	-838.876	85.057	-9.863	.000	.000
High School Degree-Some College	-243.871	55.098	-4.426	<.001	.000
High School Degree-College Degree	-496.698	53.866	-9.221	.000	.000
High School Degree-Postgraduate Degree	-770.124	80.367	-9.583	.000	.000
Some College-College Degree	-252.827	59.607	-4.242	<.001	.000
Some College-Postgraduate Degree	-526.253	84.322	-6.241	<.001	.000
College Degree-Postgraduate Degree	-273.427	83.522	-3.274	.001	.011

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .050.

Conclusion

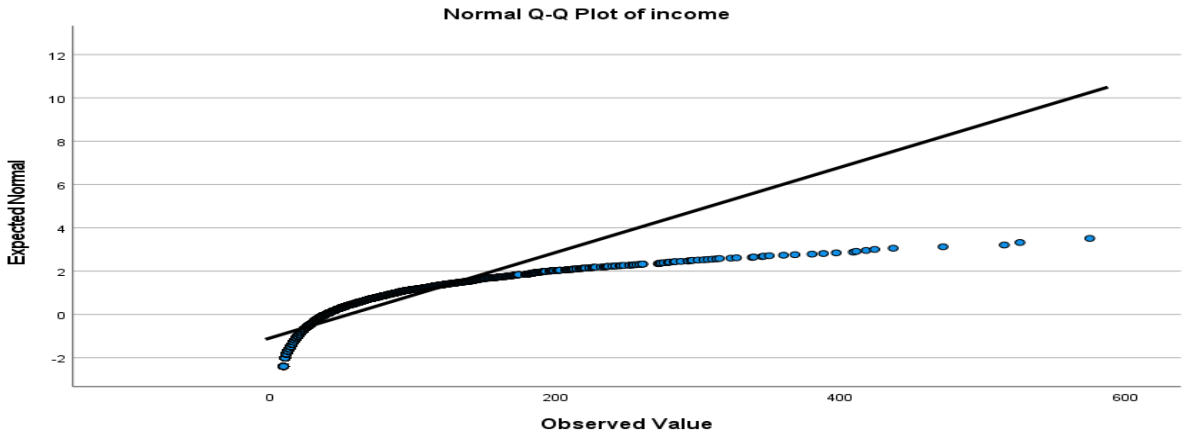
The violations of the Gauss-Markov assumptions particularly the non-normal data and the heteroskedasticity may have biased some of the results. For example, the negative coefficient on YRSEMPL likely occurred because of the observations with zero years of work experience but also the observations who had retired. However, this paper has found evidence that more years of education affects income. Although, the linear regression models were not appropriate because they failed the assumptions. The log regression was able to correct for non-linearity. The SLR model determined income rose by 4.4% for every one-year increase in education and the multiple log regression determined it to be 2.9%. Although a 4.4% increase seems high a 2.9% increase per extra year of education does seem like a rational assumption. Furthermore, an 82.4% R-square coefficient and the ANOVA less than the alpha level of .05% suggests the log-regression model is a good fit. There is no multicollinearity evident in the model also with VIF figures all below 2.0 for the independent variables.

Appendices

Tests of Normality

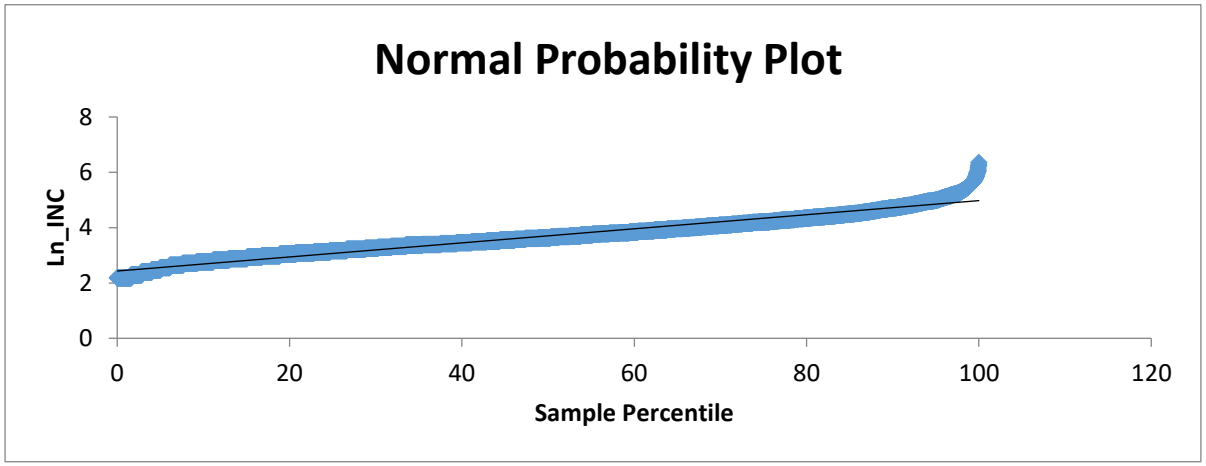
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
yrsed	.071	4504	<.001	.985	4504	<.001
yrsempl	.157	4504	<.001	.865	4504	<.001
age	.073	4504	<.001	.956	4504	<.001
income	.184	4504	.000	.718	4504	<.001

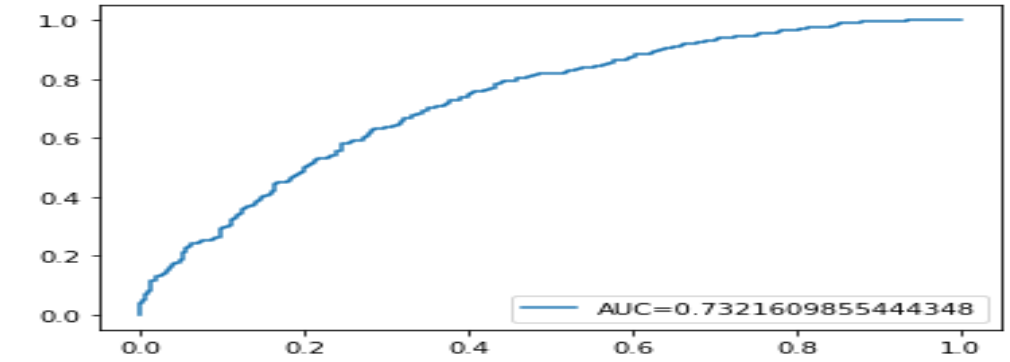
a. Lilliefors Significance Correction

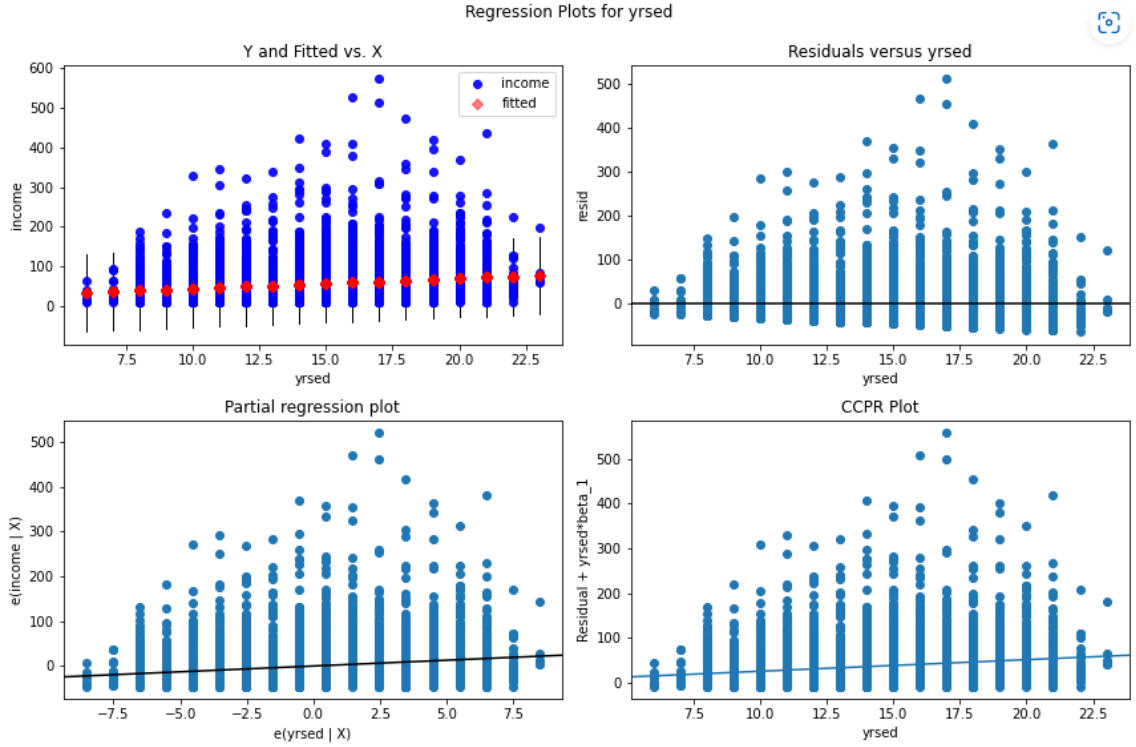


Correlations

		ln_x	age	yrsed	yrsempl	creddebt	othdebt	address	cars	carvalue	edcat	jobsat	homeown
Pearson Correlation	ln_x	1.000	.150	.196	.271	.520	.617	.215	.056	.905	.197	.260	.157
	age	.150	1.000	-.095	.701	.132	.155	.822	-.022	.208	-.081	.415	-.020
	yrsed	.196	-.095	1.000	-.219	.102	.119	-.058	.009	.174	.964	-.091	.044
	yrsempl	.271	.701	-.219	1.000	.214	.246	.598	-.004	.341	-.204	.470	.018
	creddebt	.520	.132	.102	.214	1.000	.590	.152	.026	.532	.096	.168	.084
	othdebt	.617	.155	.119	.246	.590	1.000	.180	.028	.632	.118	.183	.103
	address	.215	.822	-.058	.598	.152	.180	1.000	.000	.257	-.044	.351	.137
	cars	.056	-.022	.009	-.004	.026	.028	.000	1.000	.044	.004	.014	.018
	carvalue	.905	.208	.174	.341	.532	.632	.257	.044	1.000	.176	.266	.151
	edcat	.197	-.081	.964	-.204	.096	.118	-.044	.004	.176	1.000	-.077	.045
	jobsat	.260	.415	-.091	.470	.168	.183	.351	.014	.266	-.077	1.000	.022
	homeown	.157	-.020	.044	.018	.084	.103	.137	.018	.151	.045	.022	1.000







OLS Regression Results

Dep. Variable:	income	R-squared:	0.838
Model:	OLS	Adj. R-squared:	0.838
Method:	Least Squares	F-statistic:	5832.
Date:	Tue, 15 Nov 2022	Prob (F-statistic):	0.00
Time:	17:37:55	Log-Likelihood:	-19975.
No. Observations:	4504	AIC:	3.996e+04
Df Residuals:	4499	BIC:	3.999e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8.3509	1.516	-5.508	0.000	-11.323	-5.378
yrsed	0.3255	0.099	3.297	0.001	0.132	0.519
carvalue	1.9649	0.018	106.446	0.000	1.929	2.001
creddebt	2.6253	0.121	21.628	0.000	2.387	2.863
yrsempl	0.2379	0.035	6.759	0.000	0.169	0.307

Omnibus:	5368.347	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1180083.010
Skew:	6.058	Prob(JB):	0.00
Kurtosis:	81.367	Cond. No.	185.

References

[1] Nuno, A. (2012). The impact of education on household income and expenditure inequality.Applied Economics Letters, 19(10), 915-919.

[2] Talley, Q., Wang, T. and Zaski, G. (n.d.). "*Effect of Education on Wage Earning*. [online] Available at: https://smartech.gatech.edu/bitstream/handle/1853/60543/talley_wang_zaski_effect_of_education_on_wage_earning.pdf?sequence=1&isAllowed=y.

[3]Alves, N (2012) The impact of education on household income and expenditure inequality, Applied Economics Letters, 19:10, 915-919,

[4] Jamison, E. A.; Jamison, D. T.; Hanushek, E. A. (2007) The effects of education quality on income growth and mortality decline. Economics of Education Review, *[s. l.]*, v. 26, n. 6, p. 771–788,. DOI 10.1016/j.econedurev.2007.07.001. Accessed 12 November 2022.

[5] YANG, J.; QIU, M. (2016) "The impact of education on income inequality and intergenerational mobility". China Economic Review, *[s. l.]*, v. 37, p. 110–125, Accessed: 12 November 2022.

[6] Mann R. J. (2000) "Patterns of Credit Card Use Among Low and Moderate Income Households" Available at: https://scholarship.law.columbia.edu/faculty_scholarship/1528.

University of Virginia Library Research Data Services + Sciences, Research Data Service=Sciences. Available at: https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/ (Accessed: November 16, 2022).