Project

“Factors Affecting Household Income in the Northern Midlands and Mountains in Vietnam”

HaNoi, 2024

Table of contents

1 Introduction

Vietnam's economic development over the years has brought about significant improvements in living standards across different regions. However, disparities remain between urban centres and rural areas. The Northern Midlands and Mountains region, home to around 14 million people, relies on agriculture and forestry, industries that are less profitable than the service-oriented economies of urban areas. As a result, poverty rates in this region remain high, with limited access to infrastructure, markets and education, further widening the gap between rural and urban households (World Bank, 2021).This study used data from 7,417 households and applied multiple linear regression (MLR) and ordinary least squares (OLS) methods to analyze factors affecting household income. Factors such as age, education level and land status were analyzed to clarify the relationship between demographics, education level, land status and income. The main purpose of the study is to identify key influencing factors affecting household income, thereby providing a basis for appropriate economic policies. These policies will not only effectively support poverty reduction but also promote sustainable economic growth in Vietnam.

2 Literature Review

Factors affecting household income in the Northern Midlands and Mountains region, related to demographic, family and structural factors. Van Vu (2020) found that higher years of education of the household head were associated with higher household income. Educated individuals are better prepared to find higher-paying jobs and engage in money-generating activities, which increase household income, thereby increasing household income.

Age and gender are among other important factors affecting household income at the individual level. Older household heads tend to leverage their accumulated experience to increase their income; However, this advantage diminishes with age due to declining physical

capabilities (Tran & Vu, 2019). Gender also significantly influences income disparities. Nguyen et al. (2019) highlight that female-headed households often have lower incomes due to social barriers, which limit access to limited resources that generate income.

Household-level income is influenced not only by individual factors such as ethnicity or marital status but also by family groups (Barnard & Turner, 2011). Lam et al. (2019) highlight that language barriers, discrimination, limited access to education and financial services disproportionately affect ethnic minority households. These barriers often limit them to agricultural activities, which are less profitable than other sectors. Furthermore, ethnic minorities are less likely to own land or access credit, reducing their ability to invest in more profitable projects. This systemic inequality has created a significant income gap between ethnic minorities and the Kinh majority. In addition, family size is another influence. Larger families will face lower per capita incomes due to increased consumption demand. The additional effect is the dependency ratio. Barnard and Turner (2011) observed that households with high dependency ratios tend to experience lower income growth due to the additional financial pressure of supporting non-working dependents, such as children and the elderly.

Land ownership and efficient land use play an important role in income generation, especially in the northern midland and mountainous regions. Larger landholdings allow households to engage in more diversified agricultural activities, contributing to increased income potential. Furthermore, land ownership provides collateral for access to credit and financial support, which can further contribute to improving household income (Tuyen, 2019).

Factors such as province and urban-rural location also influence household income. Urban households often benefit from better infrastructure, access to higher-paying jobs, and diverse economic opportunities compared to rural households, who rely on agriculture. This can lead to higher household income (Nguyen Trong H, 2021). Tuyen (2015) notes that

provinces with stronger economic development, better infrastructure, and supportive government policies provide more opportunities to generate income.

3 Data and Research Design

3.1 Overview of the dataset

We are provided with a dataset of 7,417 households in the Northern Midlands and Mountains of Vietnam. Based on the data, we need to identify the factors that influence the average household income per capita. Information about the variables in the dataset:

Table 1 : Describe the variables of the data

	Variables	Explain
Explanatory variables	age	age of the household head (years)
	edu	the number of schooling years of the household head
	gender	gender of the household head: 1 = male; 0 = female
	married	marital status of the household head: "1=married; 0=otherwise"
	ethnicity	ethnicity of the household head: 1= Kinh; 0 = ethnic minority
	hhsize	total household members
	dep_ratio	the dependency ratio is calculated by dividing the number of dependents by the household size.
	log_aland	the natural log of the size of annual cropland
	log_pland	the natural log of the size of perennial cropland
	log_fland	the natural log of the size of forestland
	log_gland	the natural log of the size of garden land
	province	a categorical variable including 14 categories (name of provinces).
	urban	1=living in urban areas; 0=living in rural areas
The dependent variable	income	household income per capita/month (thousand VND)

3.2 Some characteristics of households in the northern midlands and mountains of Vietnam

Through survey analysis results, 47.04% of households surveyed are Kinh ethnic groups, meanwhile, the remaining 52.96% are ethnic groups other than Kinh. The number male of household heads is 79.52%, the remaining household head who is female is 20.48%.The average age of the household head is 48 years old with the youngest person is 18 years old and the oldest person is 99 years old. With such an age structure, we see that households have diversity in the age of the household head and can include many generations under one roof. With such an average age, the family breadwinner is usually middle-aged or elderly. This shows that the household may have had a stable economic life and spent a certain amount of time building and developing the family.

The survey results also show the education level of the household head, especially, a primary school education has 26.95% of the household heads (0-5 years), a secondary school education has 40.38% of the household heads (6-9 years),a high school education has 16.92% of the household heads (10–12 years) and higher education has 15.75% of the household heads (14-22 years). So, we can see that the main education level of the majority of the household head is secondary school education.

Besides, the average household member is about 4 people, at least 1 person with and at most 15 people in one household. The dependency ratio averages at 0.36, indicating that each working individual supports approximately 0.36 dependents, suggesting a moderate burden on breadwinners. The average monthly per capita income is 2759 thousand VND, with a wide range from 129 thousand VND to 73,086 thousand VND, reflecting substantial income disparities among households. These factors, including household size and dependency ratios, are likely to have a significant impact on household economic stability.

***Table 2**: Measure the concentration trends of variables.*

```
. summarize age edu gender married ethnicity hhsize dep_ratio log_aland log_pland log_fland log_glan
> d province urban income
```

Variable	Obs	Mean	Std. dev.	Min	Max
age	7,417	48.40326	13.12027	18	99
edu	7,417	8.722529	3.907178	0	22
gender	7,417	.7952002	.4035824	0	1
married	7,417	.8496697	.357419	0	1
ethnicity	7,417	.4704058	.4991571	0	1
hhsize	7,417	3.937711	1.57264	1	15
dep_ratio	7,417	.3684073	.2766372	0	1
log_aland	7,417	5.670269	3.572751	0	12.06682
log_pland	7,417	1.643519	3.097642	0	11.69526
log_fland	7,417	2.698598	4.159105	0	12.84793
log_gland	7,417	2.993104	2.770599	0	10.46313
province	7,417	14.8676	7.068493	2	25
urban	7,417	.2298773	.4207821	0	1
income	7,417	2759.546	2686.745	129.5833	73086.46

```
.
```

3.3 Research design

3.3.1 Model

- **Economic model**

An economic model is constructed to identify the factors influencing household income per capita. The general relationship is specified as follows:

Income = f(age, education level, gender, married, ethnicity, household size, dependency ratio, the natural log of annual cropland, the natural log of perennial cropland, the natural log of forest land, the natural log of garden land , province, urban area)

***Table 3**: The expected signs of the variables*

Explanatory variables	**Expected sign**
Age	+/-

Education level	+
Gender	+/-
Married	+/-
Ethnicity	+/-
Household size	-
Dependency ratio	-
The natural log of Annual cropland	+
The natural log of Perennial cropland	+
The natural log of Forest land	+
The natural log of Garden land	+
Province	+/-
Urban	+/-

- **Econometric model**

*Log of per capita household income = β0 + β1*Age + β2*Edu + β3*Gender + β4*Married + β5*Ethnicity + β6*Hhsize + β7*Dep_ratio + β8*Log_aland + β9*Log_pland + β10*Log_fland + β11*Log_gland + β12*Province + β13*Urban + ε*

3.3.2 Assumptions of multiple linear regression

To obtain unbiased coefficient estimates, we need to satisfy 4 assumptions about linearity in parameters, random sampling, no perfect multicollinearity, and unconditional means.

Assumption MLR.1: Linear in parameters

The relationship between Y and X is linear in parameters. This implies that a one-unit change in X has the same effect on Y regardless of X's initial values. βj appears with a power of 1 only and is not multiplied or divided by any other parameter. Our model satisfied this assumption.

Assumption MLR.2: Random sampling

The data is a random sample drawn from the population. It means that every member of the population has an equal chance of being selected for the sample. This is a dataset of random 7,417 households in 14 provinces of the Northern Midlands and Mountains region of Vietnam, which enables our model to satisfy Assumption MLR.2.

Assumption MLR.3: No perfect multicollinearity

A regression model's level of multicollinearity is assessed using the Variance Inflation Factor (VIF) test. When the independent variables in the model have a strong linear relationship with one another, this is known as multicollinearity, and it can make it more difficult to estimate the model's coefficients and lower its accuracy.

```
. vif
```

Variable	VIF	1/VIF
log_aland	2.08	0.481178
married	1.82	0.550475
gender	1.81	0.551919
urban	1.70	0.589175
ethnicity	1.65	0.607807
edu	1.30	0.766551
province	1.28	0.781632
log_gland	1.26	0.795045
log_fland	1.24	0.803808
age	1.22	0.816677
hhsize	1.22	0.822169
dep_ratio	1.09	0.914536
log_pland	1.05	0.950713
Mean VIF	1.44	

All VIF values below 10 as a rule of thumb, indicating that the regression model has no perfect multicollinearity and the coefficient estimates are considered reliable.

Assumption MLR.4: Zero conditional mean

This assumption is used to detect omitted variables in the regression model. To detect whether our model is violating variable omission errors, we conducted the Ramsey reset test.

```
. ovtest

Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of income

H0: Model has no omitted variables

F(3, 7400) =  53.60
  Prob > F = 0.0000
```

- Null hypothesis (H0): The model has no omitted variables.

- F(3,7400) = 53.60: This value is high, indicating that the powers of the fitted values of the dependent variable (income) significantly explain variation in the dependent variable that is not captured by the original model.

- Prob > F = 0.0000: A p-value of 0.0000 means that the result is statistically significant at any conventional level (e.g., 1%, 5%, 10%). Thus, we can reject the null hypothesis (H0) that there are no omitted variables in the model.

Since the null hypothesis is rejected, this strongly suggests that the model suffers from omitted variable bias. There are variables not included in the model that could explain variations in the dependent variable.

Assumption MLR.5: Homoscedasticity

The homogeneity of variance (homoscedasticity) assumption in the regression model's error terms is tested using the Breusch-Pagan/Cook-Weisberg test. If violated, we'll use Log-transformation to adjust for heteroskedasticity in this model.

```
. estat hettest

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of income

H0: Constant variance

    chi2(1) = 3251.23
Prob > chi2 =  0.0000
```

p-value = 0.00 < 0.05 => reject Ho

There is a phenomenon where the error variance changes. Consequently, the model meets Heteroscedasticity.

It is vital to handle this problem when the homoscedasticity assumption is violated. Here are several methods for handling heteroscedasticity:

1. Log-transformation.

2. Using the “robust” option provided by Stata.

3. Using other linear regression estimators

We will use Log-transformation to address this issue.

```
. reg log_income province gender age married edu ethnicity hhsize dep_ratio urban log_aland log_pland l
> og_fland log_gland
```

Source	SS	df	MS	Number of obs	=	7,417
				F(13, 7403)	=	489.56
Model	2075.70991	13	159.669993	Prob > F	=	0.0000
Residual	2414.48378	7,403	.326149369	R-squared	=	0.4623
				Adj R-squared	=	0.4613
Total	4490.19368	7,416	.605473798	Root MSE	=	.57109

log_income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
province	.0040569	.0010612	3.82	0.000	.0019767	.0061371
gender	-.0254533	.0221184	-1.15	0.250	-.0688116	.017905
age	.0067345	.0005593	12.04	0.000	.0056381	.0078309
married	.0659726	.0250079	2.64	0.008	.01695	.1149952
edu	.0561575	.0019386	28.97	0.000	.0523573	.0599577
ethnicity	.361966	.0170413	21.24	0.000	.3285602	.3953718
hhsize	-.0592282	.0046506	-12.74	0.000	-.0683448	-.0501116
dep_ratio	-.5616252	.0250676	-22.40	0.000	-.6107648	-.5124856
urban	.2416663	.0205326	11.77	0.000	.2014165	.281916
log_aland	-.0342871	.0026759	-12.81	0.000	-.0395326	-.0290416
log_pland	.0021598	.0021957	0.98	0.325	-.0021443	.006464
log_fland	.0030165	.0017785	1.70	0.090	-.0004698	.0065028
log_gland	.0011293	.0026844	0.42	0.674	-.004133	.0063915
_cons	7.101819	.0470448	150.96	0.000	7.009598	7.19404

Afterwards, we assess whether this new model addresses the issue of heteroscedasticity (varying residuals).

```
. estat hettest

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of log_income

H0: Constant variance

    chi2(1) =   0.51
Prob > chi2 = 0.4758
```

p-value = 0.4758 > 0.05 => do not reject H0

=> The new model does not violate the assumption of heteroscedasticity, indicating that the model has more precisely and successfully addressed the issue.

Assumption MLR.6 (Normality)

Unobserved factors (***ui***) are assumed to be normally distributed around the population regression function. Residuals should follow a normal distribution with zero mean and constant standard deviation. We use the Kolmogorov-Smirnov Test.

```
. ksmirnov income=normal((income- 2759.546)/ 2686.745)

One-sample Kolmogorov-Smirnov test against theoretical distribution
  normal((income- 2759.546)/ 2686.745)

Smaller group             D     p-value
-------------------------------------------
income               0.1327       0.000
Cumulative          -0.1789       0.000
Combined K-S         0.1789       0.000

Note: Ties exist in dataset;
      there are 7166 unique values out of 7417 observations.
```

p-value = 0 < 0.05 → violation the assumption

We decided to use log-transformation again to handle the violation: convert income to log_income and use Q-Q plots to test the model.

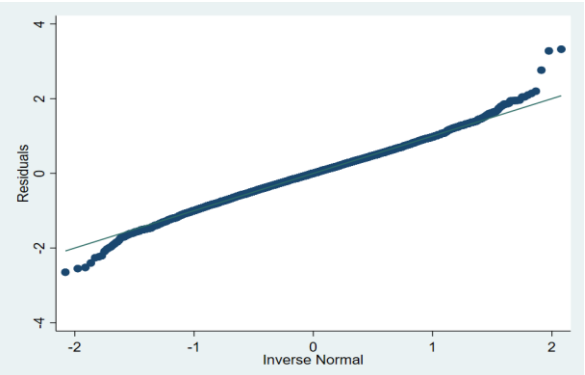

***Figure 3**: Q-Q plots*

The plot shows that the model satisfies the above assumption.

4 Results and Discussion

4.1 Examining the interaction term

4.1.1 The proportional differential in income between these groups

```
. reg log_income i.urban##i.ethnicity province gender age married edu hhsize dep_ratio log_aland log
> _pland log_fland log_gland
```

Source	SS	df	MS	Number of obs	=	7,417
				F(14, 7402)	=	467.80
Model	2107.85027	14	150.560734	Prob > F	=	0.0000
Residual	2382.34341	7,402	.321851312	R-squared	=	0.4694
				Adj R-squared	=	0.4684
Total	4490.19368	7,416	.605473798	Root MSE	=	.56732

log_income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
1.urban	.4617034	.0300145	15.38	0.000	.4028665	.5205404
1.ethnicity	.4343314	.0184125	23.59	0.000	.3982377	.4704252
urban#ethnicity						
1 1	-.3454039	.0345644	-9.99	0.000	-.41316	-.2776478
province	.0037514	.0010546	3.56	0.000	.0016841	.0058188
gender	-.0289449	.0219749	-1.32	0.188	-.0720221	.0141322
age	.0063978	.0005566	11.49	0.000	.0053067	.007489
married	.0729286	.0248523	2.93	0.003	.024211	.1216463
edu	.0551014	.0019287	28.57	0.000	.0513206	.0588822
hhsize	-.0569035	.0046258	-12.30	0.000	-.0659713	-.0478357
dep_ratio	-.5580936	.0249044	-22.41	0.000	-.6069133	-.5092739
log_aland	-.0364805	.0026672	-13.68	0.000	-.041709	-.0312519
log_pland	.0016983	.0021816	0.78	0.436	-.0025783	.005975
log_fland	.0039888	.0017694	2.25	0.024	.0005203	.0074573
log_gland	.0004871	.0026675	0.18	0.855	-.0047419	.0057161
_cons	7.103244	.046734	151.99	0.000	7.011632	7.194856

To consider the difference in income between the two variables, urban and ethnicity, with the base group being Group 4 (Ethnic minority households in rural areas), we will convert the dependent variable income to a logarithmic function to improve the income ratio. Improve the quality of regression models and increase the accuracy of analysis results. The regression table illustrates the differences in income between the groups as follows:

- Group 1: Kinh households in urban areas

Proportional differential = 1.urban.(0.461) + 1.ethnicity.(0.434) + 1.urban##ethnicity(-0.345) = 0.55

The coefficient of 0.555 shows that the change in logarithmic income of urban households is higher than that of the base group. From there, we can conclude that Kinh households in urban areas are about 55% higher than ethnic minority households in rural areas.

- Group 2: Ethnic minority households in urban areas

Proportional differential = 1.urban.(0.461) + 0.ethnicity.(0.434) + 0.urban##ethnicity(-0.345) = 0.461

➔ In group 2, urban ethnic minority households are about 46.1% higher than the base group.

- Group 3: Kinh household in rural areas

Proportional differential = 0.urban.(0.461) + 1.ethnicity.(0.434) + 0.urban##ethnicity(-0.345) = 0.434

➔ From the coefficient of 0.434, we can conclude that rural Kinh households are higher than the base group (rural ethnic minority households).

- Group 4: Ethnic minority households in rural areas (base group)

Proportional differential = 0.urban.(0.461) + 0.ethnicity.(0.434) + 0.urban##ethnicity(-0.345) = 0

The regression results show that, ceteris paribus, Kinh households in urban areas have the highest expected income among the four incomes, reflecting the significant influence of both location and ethnicity on income differences. In rural areas, Kinh households continue to earn more than ethnic minority households, showing that ethnicity plays an important role in rural income levels. These results underline how both ethnicity and location together affect income, suggesting that specific policies are needed to reduce these income differences.

4.1.2 The proportional differential in income between the base province (Ha Giang) and other provinces.

Source	SS	df	MS	Number of obs	=	7,417
				F(25, 7391)	=	272.93
Model	2155.42716	25	86.2170862	Prob > F	=	0.0000
Residual	2334.76653	7,391	.315893185	R-squared	=	0.4800
				Adj R-squared	=	0.4783
Total	4490.19368	7,416	.605473798	Root MSE	=	.56204

log_income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
province						
Tỉnh Cao Bằng	-.2040218	.0411011	-4.96	0.000	-.2845916	-.1234519
Tỉnh Bắc Kạn	-.1409443	.0396101	-3.56	0.000	-.2185914	-.0632971
Tỉnh Tuyên Quang	.0641294	.0385137	1.67	0.096	-.0113684	.1396272
Tỉnh Lào Cai	.0685877	.0404713	1.69	0.090	-.0107475	.147923
Tỉnh Điện Biên	-.1582368	.0427586	-3.70	0.000	-.2420559	-.0744177
Tỉnh Lai Châu	-.0257858	.0424479	-0.61	0.544	-.1089958	.0574241
Tỉnh Sơn La	-.1910236	.0388976	-4.91	0.000	-.2672741	-.1147732
Tỉnh Yên Bái	-.0588109	.0390329	-1.51	0.132	-.1353264	.0177046
Tỉnh Hoà Bình	.0148491	.0381797	0.39	0.697	-.059994	.0896921
Tỉnh Thái Nguyên	.0286457	.0376311	0.76	0.447	-.0451219	.1024134
Tỉnh Lạng Sơn	-.0556621	.0385713	-1.44	0.149	-.1312729	.0199486
Tỉnh Bắc Giang	.2021771	.0373744	5.41	0.000	.1289127	.2754415
Tỉnh Phú Thọ	-.0984688	.0372485	-2.64	0.008	-.1714865	-.025451
ethnicity	.2964827	.0184146	16.10	0.000	.2603848	.3325807
urban	.2765191	.0205541	13.45	0.000	.2362272	.316811
gender	-.0429977	.0219822	-1.96	0.050	-.0860891	.0000936
age	.0068945	.0005593	12.33	0.000	.0057981	.0079909
married	.075437	.0246853	3.06	0.002	.0270469	.1238272
edu	.0589454	.0019288	30.56	0.000	.0551644	.0627264
hhsize	-.0597608	.0046012	-12.99	0.000	-.0687806	-.0507411
dep_ratio	-.5588769	.0247424	-22.59	0.000	-.6073791	-.5103747
log_aland	-.031252	.0026827	-11.65	0.000	-.0365108	-.0259931
log_pland	.0023883	.0022329	1.07	0.285	-.0019887	.0067654
log_fland	.0021722	.0018953	1.15	0.252	-.0015432	.0058876
log_gland	.0035462	.0027688	1.28	0.200	-.0018814	.0089738
_cons	7.165053	.0527917	135.72	0.000	7.061567	7.26854

Controlling for other factors, there are differences in income when comparing households living in Ha Giang with households living in other provinces. However, there are some cases that are not statistically significant when the p-value > 0.1 results, such as Lai Chau, Yen Bai, Hoa Binh, Thai Nguyen, Lang Son provinces. In addition, the per capita income is about 6.41%, 6.86%, and 20.2% higher for Tuyen Quang, Lao Cai, Bac Giang respectively than for the base group (Ha Giang province). As for the provinces of Cao Bang, Bac Kan, Dien Bien, Son La and Phu Tho, the difference rate is lower than the base province, with differences of 20.4%, 14.09%, 15.82%, 19.1%, and 9.85% respectively. The results show a clear differentiation in household Income in the Northern Midlands and Mountains in Vietnam. This may reflect the disparity in economic development and living conditions between regions. There is a need for improved policies and necessary measures to narrow the income gap between provinces.

4.1.3 Interpret the effect of various types of land on income

Firstly, the coefficient of negative log_aland shows that if we increase the annual cropland area by 1%, the dependent variable income will decrease by 3.43%, holding other factors constant. With a coefficient of 0.002 for log_pland indicates that a 1% increase in perennial cropland area will increase the income variable by 0.2%. Furthermore, the results reveal that log_fland has a negative impact on income, with a 1% increase in forestland area resulting in a 0.3% decrease in average household income per capita. Finally, the coefficient demonstrates that if the garden land area is increased by 1% per year, the average income per capita of the household will increase by 0.01%. , with other factors remaining constant.

4.1.4 How do you quantify and compare the relative importance of each individual explanatory variable to the dependent variable (income)?

Source	SS	df	MS	Number of obs	=	7,417
				F(13, 7403)	=	489.56
Model	2075.70991	13	159.669993	Prob > F	=	0.0000
Residual	2414.48378	7,403	.326149369	R-squared	=	0.4623
				Adj R-squared	=	0.4613
Total	4490.19368	7,416	.605473798	Root MSE	=	.57109

log_income	Coefficient	Std. err.	t	P>|t|	Beta
province	.0040569	.0010612	3.82	0.000	.0368531
ethnicity	.361966	.0170413	21.24	0.000	.2321974
urban	.2416663	.0205326	11.77	0.000	.130685
gender	-.0254533	.0221184	-1.15	0.250	-.0132016
age	.0067345	.0005593	12.04	0.000	.1135529
married	.0659726	.0250079	2.64	0.008	.0303035
edu	.0561575	.0019386	28.97	0.000	.2819833
hhsize	-.0592282	.0046506	-12.74	0.000	-.1197044
dep_ratio	-.5616252	.0250676	-22.40	0.000	-.1996685
log_aland	-.0342871	.0026759	-12.81	0.000	-.1574292
log_pland	.0021598	.0021957	0.98	0.325	.008598
log_fland	.0030165	.0017785	1.70	0.090	.0161234
log_gland	.0011293	.0026844	0.42	0.674	.0040209
_cons	7.101819	.0470448	150.96	0.000	.

To quantify and compare the relative importance of each individual explanatory variable in the data with the dependent variable of income, we use standardized coefficient. The result allows us to compare the strength of the effect of each explanatory variable on the dependent variable.The most important factors to determine household income are education, ethnicity, and the dependency ratio with beta coefficients of 0.282, 0.232, and -0.199, respectively. In

addition, variables are not too important for the dependent variable of income, including gender, log_aland, log_pland, log_gland. However, we do not have enough evidence to conclude that non-zero correlations exist between the size of forest land and household income at significance level of 5%. But at significance level of 1%, the impact of forest land on income can be considered statistically significant.

4.2 Controlling for other explanatory variables in the model

The OLS regression model is used to analyze the effect of education on income, where income is the dependent variable and education and other factors are independent variables. This method helps to determine the relationship between education and income, and excludes the influence of other factors.

```
. reg log_income province gender age married edu ethnicity hhsize dep_ratio urban log_aland log_pland log_fland log_gland
```

Source	SS	df	MS	Number of obs	=	7,417
				F(13, 7403)	=	489.56
Model	2075.70991	13	159.669993	Prob > F	=	0.0000
Residual	2414.48378	7,403	.326149369	R-squared	=	0.4623
				Adj R-squared	=	0.4613
Total	4490.19368	7,416	.605473798	Root MSE	=	.57109

log_income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
province	.0040569	.0010612	3.82	0.000	.0019767	.0061371
gender	-.0254533	.0221184	-1.15	0.250	-.0688116	.017905
age	.0067345	.0005593	12.04	0.000	.0056381	.0078309
married	.0659726	.0250079	2.64	0.008	.01695	.1149952
edu	.0561575	.0019386	28.97	0.000	.0523573	.0599577
ethnicity	.361966	.0170413	21.24	0.000	.3285602	.3953718
hhsize	-.0592282	.0046506	-12.74	0.000	-.0683448	-.0501116
dep_ratio	-.5616252	.0250676	-22.40	0.000	-.6107648	-.5124856
urban	.2416663	.0205326	11.77	0.000	.2014165	.281916
log_aland	-.0342871	.0026759	-12.81	0.000	-.0395326	-.0290416
log_pland	.0021598	.0021957	0.98	0.325	-.0021443	.006464
log_fland	.0030165	.0017785	1.70	0.090	-.0004698	.0065028
log_gland	.0011293	.0026844	0.42	0.674	-.004133	.0063915
_cons	7.101819	.0470448	150.96	0.000	7.009598	7.19404

Results from the regression model show that the coefficient of education (edu) is 0.0561575, with a 95% confidence interval from 0.0523573 to 0.0599577. This can be interpreted as when the education level (edu) increases by 1 unit (one extra year of study), the income log increases to 0.0561575 units. In other words, every year of increase in education level, real income will increase by about 5.6%. Thus, the number of years of education of the head of the family has a significant positive impact on household income.

Test the alternative hypothesis that the effect of education on income is larger than 5% and 7%.

- **Hypothesis Testing with 5% Effect**

Hypotheses:

H_0: $\beta\_edu \leq 0.05$

H_a: $\beta\_edu > 0.05$

The **t-statistic** is calculated as:

$$t = (\beta\_edu - 0.05)/ SE(\beta\_edu) = (0.0561575 - 0.05)/ 0.0019386 = 3.17626122$$

At a 5% significance level (one-sided test) and with 7403 degrees of freedom, the critical value is 1.6450595.

```
. display invttail(7403, 0.05)
1.6450595
```

Given that the t-statistic (3.17626122) exceeds the critical value (1.6450595), we can reject the null hypothesis. This indicates that there is sufficient evidence to support the claim that the effect of education on income is statistically greater than 5%.

- **Hypothesis Testing with 7% Effect**

Hypotheses:

H_0: $\beta\_edu \leq 0.07$

H_a: $\beta\_edu > 0.07$

The **t-statistic** is calculated as:

$$t = (\beta\_edu - 0.07)/ SE(\beta\_edu) = (0.0561575 - 0.07)/0.0019386 = -7.14046219$$

At a 7% significance level (one-sided test) and with 7403 degrees of freedom, the critical value is 1.4759494.

```
. display invttail(7403, 0.07)
1.4759494
```

Given that the t-statistic (-7.14046219) less the critical value (1.4759494), we can not reject the null hypothesis. This indicates that there is not sufficient evidence to support the claim that the effect of education on income is statistically greater than 7%.

In conclusion, the effect of education on income is **larger than 5%** and **smaller than 7%**.

4.3 The quadratic relationship between age and the log of income

4.3.1 Quadratic relationship between age and the log of income

To determine if a quadratic relationship exists between age and the logarithm of income, we will start by creating a new variable, age2= age*age. Then, we will perform a regression of "log_income" using "age" and "age2" as independent variables.

```
. reg log_income age age2
```

Source	SS	df	MS	Number of obs	=	7,417
				F(2, 7414)	=	269.39
Model	304.195226	2	152.097613	Prob > F	=	0.0000
Residual	4185.99846	7,414	.564607291	R-squared	=	0.0677
				Adj R-squared	=	0.0675
Total	4490.19368	7,416	.605473798	Root MSE	=	.7514

log_income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	.0747369	.0040173	18.60	0.000	.0668619	.0826118
age2	-.0006217	.0000387	-16.05	0.000	-.0006976	-.0005458
_cons	5.566183	.1000637	55.63	0.000	5.37003	5.762336

To check the U-shaped pattern in the relationship between age and the logarithm of income, we use the Utest function in Stata.

```
. utest age age2

Specification: f(x)=x^2
Extreme point:  60.10623

Test:
     H1: Inverse U shape
 vs. H0: Monotone or U shape
```

	Lower bound	Upper bound
Interval	18	99
Slope	.0523554	-.048361
t-value	19.74162	-12.83719
P>|t|	6.85e-85	1.26e-37

```
Overall test of presence of a Inverse U shape:
     t-value =     12.84
     P>|t|   =  1.26e-37
```

Since the p-value is less than 5%, the null hypothesis is rejected, indicating a quadratic relationship between age and the logarithm of income at a 95% confidence level. This means income increases with age up to a certain point and then starts to decrease. The extreme point, calculated to be approximately 60.10623 years, represents the age at which income is highest. After this age, income tends to decline, possibly due to factors like reduced productivity, career

shifts, or preference for less demanding roles. This pattern highlights the nonlinear nature of income growth over time and the importance of considering age when analyzing income trends.

4.3.2 Graphs showing the quadratic relationship between age and the log of income

We use outreg2 to get the regression outputs

VARIABLES	(1) Without age2	(2) With age2
age	0.0112***	0.0747***
	(0.000676)	(0.00402)
age2		-0.000622***
		(3.87e-05)
Constant	7.080***	5.566***
	(0.0339)	(0.100)
Observations	7,417	7,417
R-squared	0.035	0.068

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4: *Quadratic relationship between age and household income*

We found a quadratic relationship between the natural log of income and age, peaking at 60. From 18 to 60, income increases with age due to greater work experience. After 60, income declines as health deteriorates, affecting earning ability, and 60 is also Vietnam's male retirement age. The coefficients for age and age squared are highly significant ($p < 0.01$) with minimal standard errors, showing a strong relationship. Adding age squared increased the explained variation in income from 3.5% to 6.8%.

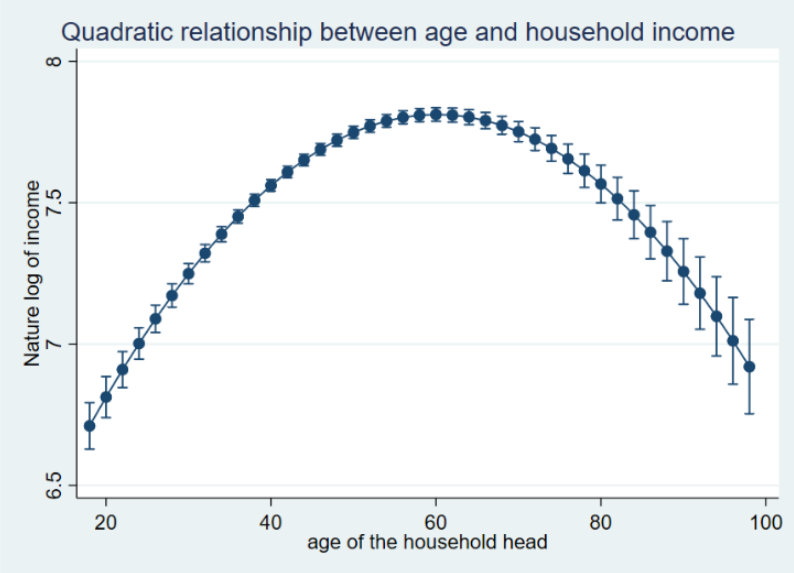

Figure 4: *Quadratic relationship between age and household income*

5 Conclusions and Policy

5.1 Conclusion

This study investigated the impact of different socioeconomic factors on household income in the Northern Midlands and Mountains of Vietnam, based on a data set including information on 7417 households. The study identified significant income disparities based on ethnicity and region of residence (urban vs. rural). Kinh households in urban areas have the highest income, followed by ethnic minority households in urban areas, Kinh households in rural areas, and finally ethnic minority households in rural areas.These findings underscore deep and systemic income inequalities among population groups, reflecting differences not only in location but also in socio-cultural dynamics.

Additionally, the analysis identified the impact of land use on household income. An increase in perennial cropland and garden land is associated with higher income, while the expansion of forestland, annual cropland has a negative impact. This highlights the importance of sustainable land management practices in fostering economic development.

Furthermore, the results also show that education plays an important role and has a positive impact on household income. Specifically, each additional year of education spent by the head of the household significantly increases household income, underscoring the importance of achieving high levels of education in improving economic outcomes.

Besides, the study found that income has a quadratic relationship with age, gradually increasing until a certain point (about age 60), and then gradually decreasing. This mirrors the typical earnings life cycle, where earnings increase with experience and productivity in the early years but decline as productivity declines with age.

5.2 Policy implication

There are some policy recommendations to improve household income in the Northern Midlands and Mountains of Vietnam based on the results analyzed. Some insights are provided on various factors that affect Vietnam citizens' outcomes, guiding decision-making and future activities. Based on the findings, the following policy measures are proposed to address these challenges.

Firstly, focusing on education is extremely important. Those people who have a high level of education provide a high income, and vice versa. The analysis reveals that to address income disparities among citizens, it is recommended to implement policies that promote equal access to education. Policy could be to expand educational infrastructure in remote areas, especially those with large ethnic minorities. Training programs should be developed based on the local cultural context. Financial support and scholarship programs are needed to support continuing education for these groups. In addition, vocational education programs should be established in line with local economic opportunities and market needs, allowing for the development of practical skills that can generate income.

Secondly, support programs should be developed for different age groups. Middle-aged workers need support to increase their productivity and improve social security measures. Creating programs and flexible work arrangements in the workplace are needed to help maintain older workers' income levels.

Furthermore, to lower household dependency ratios and boost income, effective policies should be put in place. This includes family planning to control birth rates, improving reproductive health education, and providing reliable contraceptive methods. Additionally, investing in education, promoting workforce participation, supporting entrepreneurship, and creating new job opportunities will help improve income and reduce dependency ratios.

Finally, addressing the urban-rural gap requires a comprehensive approach to rural development. Economic development zones should be developed around urban centers to create employment opportunities closer to rural areas.

References

1. Van Vu, H. (2020). The impact of education on household income in rural Vietnam. International Journal of Financial Studies, 8(1), 11.
2. Đình, T. T., Mai, N. N., Yến, N. N., Chung, N. V., & Thanh, H. N. T. (2022). Các yếu tố ảnh hưởng đến thu nhập của hộ gia đình tại một số tỉnh Tây Bắc Việt Nam. Tạp chí Kinh tế và Phát triển, (305 (2)), 69-78.
3. Nguyen, T. L. H., & Nagase, K. (2019). The influence of total quality management on customer satisfaction. *International journal of healthcare management*, *12*(4), 277-285.
4. Huong, N. T. T., Hoa, N. Q., & Lien, N. T. T. Gender Inequality In Education In Northern Midland And Mountainous Area In Vietnam.
5. Barnard, H., & Turner, C. (2011). Poverty and ethnicity: A review of evidence. *York: Jrf*.
6. Lam, B. T., Hop, H. T. M., Burny, P., Dogot, T., Cuong, T. H., & Lebailly, P. (2019). Impacts of credit access on agricultural production and rural household's welfares in northern mountains of Vietnam. *Asian Social Science*, *15*(7), 119.
7. Tran, T. Q., & Van Vu, H. (2019). Land fragmentation and household income: First evidence from rural Vietnam. *Land use policy*, *89*, 104247.
8. Tuyen, T. Q. (2015). Socio-economic determinants of household income among ethnic minorities in the North-West Mountains, Vietnam. *Croatian Economic Survey*, *17*(1), 139-159.
9. Nguyen Trong, H. (2021). *Forest Landscape Restoration and Ecosystem Services in A Luoi District, Thua Thien Hue Province, Vietnam* (Doctoral dissertation, Dissertation, Göttingen, Georg-August Universität, 2021).