VIETNAM NATIONAL UNIVERSITY, HANOI
INTERNATIONAL SCHOOL

FINAL REPORT
ECONOMETRICS
GROUP 10: PROJECT 1

Course: INS304901
Lecturer: Dr. Tran Quang Tuyen - Master. Le Van Dao

Name: Le Thi Diem Quynh 20070974

Nguyen Minh Hien 20070928

Nguyen Thi Thu Thao 20071078

Phung Linh Chi 21070688

Hanoi, 2024/11/6

TABLE OF CONTENTS

I. Introduction

The economic landscape of Vietnam's Northern Midlands and Mountains region, characterized by a mix of socio-economic variables, serves as the focal point of our analysis. Leveraging data from a comprehensive survey of 7,417 households conducted across 13 provinces, this study aims to investigate the determinants of household income, emphasizing demographic factors, education, gender, marital status, ethnicity, household size, dependency ratio, urbanization, and various types of land ownership.

Our primary objective is to construct an econometric model that elucidates the factors intricately linked to household income. By examining the influence of province, gender, age, marital status, education, ethnicity, household size, dependency ratio, urbanization, and the natural log of different land types, we seek to discern their collective impact on income levels.

This research provides a detailed exploration of how individual and household-level characteristics, as well as structural factors like land ownership and geographic location, contribute to household income variations. Special attention is given to understanding the differences in income determinants between urban and rural areas, and how various land types affect income.

Our study aims to offer empirical evidence and theoretical insights that can inform policymakers and researchers. The subsequent sections will delve into detailed findings, policy implications, and potential future research avenues.

Our report is organized as follows: Literature Review, Data and Research Design, Results and Discussion, and Conclusion and Policy Implication.

II. Literature Review

In the Northern Midlands and Mountains region of Vietnam, individual-level factors significantly influence household income (Benin & Randriamamonjy, 2008). Human capital and education consistently emerge as vital predictors of income. Research indicates that higher levels of education, as measured by the number of school years completed by the household head, are strongly associated with higher household incomes (Nguyen & Nagase, 2019). Education equips individuals with the skills and qualifications needed to secure better-paying jobs and engage in income-generating activities, thereby enhancing household income.

Age and gender are also critical factors affecting household income at the individual level. Older household heads often accumulate more wisdom and experience, which can translate into higher earnings. Gender is crucial due to its significant impact on income disparities, as studies repeatedly show (Dau et al., 2022). Female-headed households typically earn less than their male counterparts, attributed to socio-cultural factors and limited access to resources or well-paying jobs (Thi Thu Huong et al., 2019).

Household income is further influenced by individual factors such as ethnic background and marital status, as well as family dynamics (Barnard & Turner, 2011). Married couples benefit from economies of scale and shared resources, leading to higher household incomes. Ethnic minorities, although a substantial part of the population, often face challenges in achieving financial stability due to discrimination and limited access to education and resources, compounded by language barriers (Lam et al., 2019). The number of dependents and family size also play a crucial role in determining household income, as larger families require more resources and labor (NVN Mbuya, SJ Atwood, PN Huynh - 2019).

In the Northern Midlands and Mountains region, land ownership and agricultural activities are major income sources. The size of various land types, including cropland, forestland, and garden land, directly affects household income (Tran & Vu, 2019). Larger landholdings enable households to diversify their agricultural activities, enhancing income potential. Land ownership also provides collateral for accessing credit and financial support, further boosting household income (Tuyen, 2019).

Structural factors such as the province of residence and urban-rural location also impact household income. Economic development, infrastructure, and resource availability vary across provinces, influencing income levels (Tuyen, 2015). Investment, market access, and government support differ throughout the region, affecting household income. The urban-rural divide is also significant. Urban areas offer more employment opportunities, higher-paying jobs, and a more diverse economy compared to rural areas, which are predominantly agriculture-based. This disparity leads to higher household incomes in urban locations (Wyss & Pawelzik, 2021)

III.Data and Research Design

1. Data

Data collected and extracted from an available data set was collected in 13 provinces located in the Northern midlands and mountainous regions of Vietnam with a variety of factors, including a sample of 7417 households. (5712 rural households and 1705 urban households). Both household data and commune data were surveyed according to the General Statistics Office survey (GSO, n.d.) with detailed population characteristics affecting the collection such as place of residence, gender, age Age of household head, marital status, education level, household size, Labor - Employment, Income, dependency ratio. Specific data will be collected for this study, specifically the size of annual cropland per person, the size of perennial crops per person, the size of forest area per person, and the size of garden area per person. Look at the relationships between factors to learn more about how household income is affected.

2. Research design

In this project to analyze, as well as learn about the correlation between variables and income, we will choose multiple linear regression using Stata as the main software for research and analysis

Economic Model:

The economic model aims to explain the variation in household income per capita based on several key factors. All explanatory variables have been carefully chosen for their potential impact on income, reflecting individual characteristics and regional conditions.

Income per Capita (Age, Gender, Married, Edu, Ethnicity, HHSize, Dep_Ratio, Urban, Log_ALand, Log_PLand, Log_FLand, Log_GLand, Province)

Where:

- **Age**: Age of the household head (years)
- **Gender:** Whether or not the household head is male (male = 1; female = 0)
- **Married**: Marital status of the household head (married = 1; otherwise = 0)
- **Edu**: Years of schooling of the household head (years)
- **Ethnicity:** Whether or not ethnicity is Kinh & Chineses (Kinh & Chineses = 1; otherwise = 0)
- **HHSize:** Total household members (persons)
- **Dep_Ratio**: The proportion of dependents in the households
- **Urban**: Whether or not the household head lives in urban (urban = 1; rural = 0)
- **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 forest land
- **Log_GLand:** The natural log of the size of garden land
- **Province:** Provinces where they live (within the northern midland and mountainous provinces)

Econometric Model:

To estimate the relationship between the explanatory variables and household income per capita, an econometric model using ordinary least squares (OLS) can be employed. OLS is suitable for continuous dependent variables like income per capita. The model can be represented as follows:

$$\text{Income per Capita} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Gender} + \beta_3 \text{Married} + \beta_4 \text{Edu} + \beta_5 \text{Ethnicity} + \beta_6 \text{HHSize} + \beta_7 \text{Dep\_Ratio} + \beta_8 \text{Urban} + \beta_9 \text{Log\_ALand} + \beta_{10} \text{Log\_PLand} + \beta_{11} \text{Log\_FLand} + \beta_{12} \text{Log\_GLand} + \beta_{13} \text{Province} + \epsilon$$

In delving into the intricacies of multiple linear regression to understand the relationship between various variables and household income per capita, we recognize the pivotal role played by each factor. The decision to retain all variables results from careful consideration of multiple factors. Our choice is not solely contingent on p-values but involves leveraging subject matter expertise and practical insights. This approach is crucial as certain factors, though immeasurable, contribute significantly to ensuring the comprehensiveness and accuracy of our model. Retaining all variables reflects our commitment to encompassing all potential contributors for a profound understanding of income dynamics.

The table below shows the estimated coefficients, standard errors, t-values, p-values, and confidence intervals for each variable included in the model:

- The Regression of the Model:

```
. reg income province gender age married edu ethnicity hhsize dep_ratio urban log_aland log_pland log_fland log_gla
> nd

      Source |       SS           df       MS      Number of obs   =     7,417
-------------+----------------------------------   F(13, 7403)     =    223.21
       Model |  1.5075e+10         13  1.1596e+09   Prob > F        =    0.0000
    Residual |  3.8459e+10      7,403  5195002.42   R-squared       =    0.2816
-------------+----------------------------------   Adj R-squared   =    0.2803
       Total |  5.3533e+10      7,416   7218599.9   Root MSE        =    2279.3

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    province |  -14.18486   4.235259    -3.35   0.001    -22.48718   -5.882551
      gender |  -122.6192   88.27503    -1.39   0.165    -295.6634    50.42495
         age |   14.26567   2.232237     6.39   0.000     9.88985    18.64149
     married |   305.7927   99.80716     3.06   0.002     110.1422    501.4431
         edu |   151.2094   7.737031    19.54   0.000     136.0426    166.3762
   ethnicity |   977.0331   68.01238    14.37   0.000     843.7095    1110.357
      hhsize |  -163.4649   18.56086    -8.81   0.000    -199.8495   -127.0803
   dep_ratio |  -1674.061   100.0454   -16.73   0.000    -1870.178   -1477.944
       urban |   627.9007   81.94616     7.66   0.000     467.2629    788.5385
  log_aland  |  -95.05942   10.67954    -8.90   0.000    -115.9944   -74.12448
  log_pland  |  -4.143512   8.762985    -0.47   0.636    -21.32146    13.03443
  log_fland  |  -1.306673   7.097936    -0.18   0.854    -15.22065     12.6073
  log_gland  |  -11.94474   10.71368    -1.11   0.265     -32.9466    9.057116
       _cons |   2040.264   187.7571    10.87   0.000     1672.206    2408.321
------------------------------------------------------------------------------
```

- Explanatory Variables and Expected Signs:

We provide an overview of the explanatory variables used in our econometric model to analyze factors affecting household income per capita. Each variable is defined and measured, along with the expected sign of its coefficient based on economic theory and previous empirical studies.

Explanatory Variables	Definition and Measurement	Sign
Age	Age of the household head (years)	+/-
Province	Province where they live (within the northern midland and mountainous provinces)	+/-
Gender	Whether or not the household head is male (male = 1, female =0)	+/-
Marital Status	Marital status of the household head (married =1, otherwise =0)	+/-
Education	Years of schooling of the household head (years)	+
Ethnicity	Whether or not ethnicity is Kinh and Chinese (Kinh & Chinese =1, otherwise =0)	+/-
Household Size	Total household members (persons)	-
Dependency Ratio	The proportion of dependents in the households	-

Urban	Whether or not the household head lives in urban (urban =1, rural =0)	+/-
The Natural Log of Annual Crop Land	The natural log of the size of annual crop land	+
The Natural Log of Perennial Crop Land	The natural log of the size of Perennial crop land	+
The Natural Log of Forest Land	The natural log of the size of forest land	+
The Natural Log of Garden Land	The natural log of the size of garden land	+

IV. Discussion and results

1. Assumptions of Multiple Linear Regression (MLR) for Unbiased Coefficient Estimates

a, Linearity

The econometric model satisfies the assumption MLR.1 - Linearity in Parameters. This assumption implies that the relationship between the dependent variable (Household Income) and the independent variables (Province, Gender, Age, Marital Status, Education, Ethnicity, Household Size, Dependency Ratio, Urban, Natural Log of Annual Crop Land, Natural Log of Perennial Crop Land, Natural Log of Forest Land, Natural Log of Garden Land) is linear. In this model, each variable has a linear coefficient (β1, β2, β3, β4, β5, β6, β7, ...).

The assumption of linearity in the parameters is met as the model is structured in a linear combination of these coefficients for each independent variable. This means that any change in an independent variable result in a proportional change in the dependent variable, assuming other variables remain constant.

b, No Perfect Multicollinearity

- Assumption: None of the independent variables is a perfect linear function of any other predictors.
- Check: Use Variance Inflation Factor (VIF).
- Reasoning: Multicollinearity occurs when two or more independent variables are highly correlated, making it difficult to estimate the regression coefficients accurately.
- How to Check: If VIF values are greater than 10, it indicates a problematic level of multicollinearity.

- Results: The VIF table below shows that all variables in the model have VIF values less than 10, indicating no severe multicollinearity issues.

```
. 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
```

Explanation:

- All variables in the model have an average VIF of 1.44.
- The variable with the highest VIF is log_aland with a value of 2.08, still below the threshold of 10.
- This result confirms that there is no perfect multicollinearity among the independent variables in the model.

2. Assumptions for Precise Estimates and Hypothesis Testing

a, Homoscedasticity

To ensure the validity of our regression model's estimates, it is crucial to test for heteroscedasticity, which refers to the non-constant variance of error terms. Heteroscedasticity violates the assumption of homoscedasticity required for Ordinary Least Squares (OLS) estimates to be efficient and unbiased.

We have tested Brreusch-Pagan / Cook-Weisberg to check the homogeneous variance and the test results are shown in the output below:

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of income

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

The chi-square statistic is 3251.23 with a p-value of 0.0000. Since the p-value is less than the conventional significance level (0.05), we reject the null hypothesis. This indicates the presence of heteroscedasticity in our model.

- Adjusting for Heteroskedasticity

Due to the heterogeneous variance, we proceed to change the logarithm to the dependent variable (Income). After changing, we ran over and checked the homogeneous variance.

```
. gen log_income = ln(income + 1)

. 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	Coef.	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

- Check the uniformity of the variance

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of log_income

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

After changing the logarithm to the dependent variable (log_income), we proceed to check the homogeneous variance by the Brreusch-Pagan / Cook-Weisberg test. The test results after adjustment showed that the value of chi-squared was 0.51 with a P-value of 0.4758, indicating that the homogeneous variance was guaranteed.

b, The standard of errors

The assumption of the standard of the error requires the residual part to follow the standard distribution with an average of 0 and the standard deviation is constant. This assumption can be evaluated through residual charts, Q-Q charts, or tests such as Shapiro-Wilk testing.

In our analysis, do not directly check the standard due to the large sample size (n = 7,417), this ensures that the central limit theorem is applied, making the standard of less important residues than.

To determine heteroscedasticity and multicollinearity, one can examine plots of residuals versus predicted values and review the correlation matrix or corresponding VIF (Variance Inflation Factor) values. If heteroscedasticity is detected, it can be handled by using robust standard errors or transforming the dependent variable. Multicollinearity can be managed by eliminating or combining correlated variables or applying techniques such as Principal Component Analysis (PCA).

3. Create an interaction term between the urban and ethnic variables

To create an interaction term between the urban and ethnicity variables, we create a new variable named "urban_ethnic" with the following groups:

- Group 1: Kinh households in urban areas (Urban 1, Ethnic 1)
- Group 2: Kinh households in rural areas (Urban 0, Ethnic 1)
- Group 3: Ethnic minority households in urban areas (Urban 1, Ethnic 0)
- Group 0: Ethnic minority households in rural areas (base group) (Urban 0, Ethnic 0)

```
. gen urban_ethnic = urban * ethnicity

. replace urban_ethnic = 1 if (urban == 1 & ethnicity == 1)
(0 real changes made)

. replace urban_ethnic = 2 if (urban == 0 & ethnicity == 1)
(2,263 real changes made)

. replace urban_ethnic = 3 if(urban == 1 & ethnicity == 0)
(479 real changes made)

. replace urban_ethnic = 0 if (urban == 0 & ethnicity == 0)
(0 real changes made)
```

Figure 3. 1. Interaction term between the urban and ethnicity

a. What is the proportional differential in income between these groups?

To analyze the proportional differences in income between the groups mentioned above, we need to regress log of income on these groups, controlling for all other explanatory variables in the model *(Figure 3. 2).*

Statistical significance: All groups have p-values = 0, specifically this value is < 0.05 so it is completely statistically significant. This indicates the presence of heteroscedasticity in our model.

The base group (Ethnic minority households in rural areas): The coefficient for this region is set as the reference category, and its coefficient equals 0. By comparing the coefficients of the interaction variable "urban_ethnic" in the regression results, we can see the difference in income between groups as follows:

- With group 1, holding other factors, we can predict the average income of **Kinh households in urban** areas is nearly **110% higher** than Ethnic minority households in rural areas. The figure of 110% represents a sharp contrast between these two groups, pointing to the strong economic advantage that belongs to Kinh people living in big cities.
- With group 2 , holding other factors, we can predict that the average income of **Kinh households in rural areas** is **69% higher** than Ethnic minority households in rural areas. The income gap still exists even in the same region. The Kinh people have an advantage, possibly due to education and geographical location (the delta).

- With group 3, holding other factors, we can predict that the average income of **Ethnic minority household in urban areas** is **81% higher** than Ethnic minority household in rural areas. Higher than 81% emphasizes the positive impact of urbanization on income levels even among ethnic minorities.*Figure 3. 2.*

```
. reg log_income i.urban_ethnic

      Source |       SS           df       MS      Number of obs   =     7,417
-------------+----------------------------------   F(3, 7413)      =   1137.82
       Model |  1415.7044          3  471.901467   Prob > F        =    0.0000
    Residual |  3074.48928     7,413  .414742922   R-squared       =    0.3153
-------------+----------------------------------   Adj R-squared   =    0.3150
       Total |  4490.19368     7,416  .605473798   Root MSE        =    .64401

------------------------------------------------------------------------------
  log_income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
urban_ethnic |
           1 |   1.103539   .0214135    51.53   0.000     1.061562    1.145515
           2 |   .6892302   .0174219    39.56   0.000     .6550784     .723382
           3 |   .8163371   .0314023    26.00   0.000     .7547797    .8778944
             |
       _cons |   7.174682   .0109659   654.27   0.000     7.153185    7.196178
------------------------------------------------------------------------------
```

Figure 3. 2. Regression log of income on urban ethnicity groups

b. What is the proportional differential in income between the base province (Ha Giang) and other provinces?

First, to determine the proportional difference in income between the base province (Ha Giang) and the other provinces, we need to determine which and how many provinces there are by using the command *"tabulate province"* in STATA *(Figure 3. 3).*

There are a total of 14 provinces and cities with the first province - Ha Giang. By default, Stata uses the first category of the variable as the base category. From there we can ensure that Ha Giang is the base province in this case without any change.

```
. tabulate province
```

provinces of the North Midlands and Moutains	Freq.	Percent	Cum.
Tỉnh Hà Giang	369	4.98	4.98
Tỉnh Cao Bằng	434	5.85	10.83
Tỉnh Bắc Kạn	478	6.44	17.27
Tỉnh Tuyên Quang	538	7.25	24.52
Tỉnh Lào Cai	420	5.66	30.19
Tỉnh Điện Biên	355	4.79	34.97
Tỉnh Lai Châu	350	4.72	39.69
Tỉnh Sơn La	559	7.54	47.23
Tỉnh Yên Bái	503	6.78	54.01
Tỉnh Hoà Bình	562	7.58	61.59
Tỉnh Thái Nguyên	726	9.79	71.38
Tỉnh Lạng Sơn	531	7.16	78.54
Tỉnh Bắc Giang	817	11.02	89.55
Tỉnh Phú Thọ	775	10.45	100.00
Total	7,417	100.00	

Figure 3. 3. List of 14 provinces

Statistical significance: Only in Lang Son province, p-value = 0.936 > 0.05 is not statistically significant. The remaining provinces all have very small p-values, specifically this value is < 0.05 so it is completely statistically significant. This indicates the presence of heteroscedasticity in our model.

We can see the proportional difference in income between Ha Giang province and the remaining provinces based on the coeficient of the variable "province" in *Figure 3. 4*.

- Some regions such as **Cao Bang, Bac Can, Dien Bien, Lai Chau, Son La** have **lower income** than Ha Giang province with a difference from -11% to -30%.
- The remaining regions such as **Tuyen Quang, Lao Cai, Yen Bai, Hoa Binh, Thai Nguyen, Bac Giang, Phu Tho** have **higher income** levels 10% - 43% than Ha Giang.

```
. reg log_income i.province
```

Source	SS	df	MS		
Model	390.228093	13	30.0175456	Number of obs	= 7,417
Residual	4099.96559	7,403	.553824881	F(13, 7403)	= 54.20
				Prob > F	= 0.0000
				R-squared	= 0.0869
				Adj R-squared	= 0.0853
Total	4490.19368	7,416	.605473798	Root MSE	= .74419

log_income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
province						
Tỉnh Cao Bằng	-.181252	.052697	-3.44	0.001	-.2845531	-.0779509
Tỉnh Bắc Kạn	-.1618873	.0515704	-3.14	0.002	-.2629801	-.0607946
Tỉnh Tuyên Quang	.169834	.050302	3.38	0.001	.0712277	.2684403
Tỉnh Lào Cai	.2404089	.0530991	4.53	0.000	.1363196	.3444982
Tỉnh Điện Biên	-.1329767	.0553259	-2.40	0.016	-.2414311	-.0245222
Tỉnh Lai Châu	-.1119086	.0555269	-2.02	0.044	-.2207571	-.0030601
Tỉnh Sơn La	-.3018293	.0499162	-6.05	0.000	-.3996791	-.2039794
Tỉnh Yên Bái	.1342342	.0510091	2.63	0.009	.0342419	.2342264
Tỉnh Hoà Bình	.1055595	.0498632	2.12	0.034	.0078136	.2033055
Tỉnh Thái Nguyên	.4323005	.0475786	9.09	0.000	.3390329	.5255682
Tỉnh Lạng Sơn	.0040291	.0504367	0.08	0.936	-.0948413	.1028994
Tỉnh Bắc Giang	.4261842	.0466772	9.13	0.000	.3346837	.5176847
Tỉnh Phú Thọ	.2174122	.047069	4.62	0.000	.1251436	.3096809
_cons	7.520235	.0387412	194.11	0.000	7.444291	7.596179

Figure 3. 4. Regression log of income on the province

c. Interpret the effect of various types of land on income.

We consider the impact of 4 types of land on income based on the regression results in *Figure 3.5.*

1. Land for growing annual crops (log_aland)

- With coefficient: -0.100 and p-value: 0.000
- Explanation: A 1% increase in annual crop land area is associated with a **0.1% decrease in household income**, holding other factors. This result is statistically significant at the 1% significance level ($p < 0.01$).

2. Perennial crop land (log_pland)

- With coefficient: -0.004 and p-value: 0.085
- Explanation: A 1% increase in land area planted with perennial crops is associated with a 0.004% decrease in household income, holding other factors. However, this result is **not statistically significant** at the 5% significance level ($p > 0.05$).

3. Forest land (log_fland)

- With coefficient: -0.007 and p-value: 0.000
- Explanation: Forest land area increases by 1%, **household income decreases by 0.007%**, holding other factors. This result is statistically significant at the 1% significance level ($p < 0.01$).

4. Garden land (log_gland)

- With coefficient: -0.008 and p-value: 0.007
- Explanation: Garden land area increases by 1%, **household income decreases by 0.008%**, holding other factors. This result is statistically significant at the 1% significance level ($p < 0.01$).

```
. regress log_income log_aland log_pland log_fland log_gland

      Source |       SS           df       MS      Number of obs   =     7,417
-------------+----------------------------------   F(4, 7412)      =    596.89
       Model |  1093.98964         4  273.49741   Prob > F        =    0.0000
    Residual |  3396.20404     7,412  .45820346   R-squared       =    0.2436
-------------+----------------------------------   Adj R-squared   =    0.2432
       Total |  4490.19368     7,416  .605473798   Root MSE        =    .67691

------------------------------------------------------------------------------
  log_income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  log_aland |  -.1007317   .0025185   -40.00   0.000    -.1056686   -.0957947
  log_pland |  -.0044395   .0025734    -1.73   0.085     -.009484     .000605
  log_fland |    -.00741   .0020218    -3.67   0.000    -.0113733   -.0034467
  log_gland |  -.0084822   .0031293    -2.71   0.007    -.0146165   -.0023478
      _cons |    8.24396   .0153701   536.36   0.000     8.21383    8.274089
------------------------------------------------------------------------------
```

Figure 3. 5. Regression log of income on 4 types of land

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

To quantify and compare the relative importance of each individual explanatory variable with the dependent variable (income), we need to standardize the regression coefficient. *(Figure 3. 6)*

```
. regress log_income age edu gender married ethnicity hhsize dep_ratio log_alan
> d log_pland log_fland log_gland province urban urban_ethnic, beta
```

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	Coef.	Std. Err.	t	P>|t|	Beta
age	.0063978	.0005566	11.49	0.000	.1078766
edu	.0551014	.0019287	28.57	0.000	.2766803
gender	-.0289449	.0219749	-1.32	0.188	-.0150126
married	.0729286	.0248523	2.93	0.003	.0334987
ethnicity	.2616295	.0196823	13.29	0.000	.1678326
hhsize	-.0569035	.0046258	-12.30	0.000	-.1150061
dep_ratio	-.5580936	.0249044	-22.41	0.000	-.1984129
log_aland	-.0364805	.0026672	-13.68	0.000	-.1675002
log_pland	.0016983	.0021816	0.78	0.436	.0067608
log_fland	.0039888	.0017694	2.25	0.024	.0213202
log_gland	.0004871	.0026675	0.18	0.855	.0017345
province	.0037514	.0010546	3.56	0.000	.0340782
urban	.2026505	.0207672	9.76	0.000	.1095866
urban_ethnic	.086351	.0086411	9.99	0.000	.1124969
_cons	7.103244	.046734	151.99	0.000	.

Figure 3. 6. Standardized the coefficients

We should not compare conventional regression coefficients to consider the importance of each independent variable because this coefficient only reflects the change in income when x increases by 1%. Furthermore, the units of the variables are different so they cannot be directly compared. So we need to transform the coefficients so that they are based on the same scale (mean of 0 and a standard deviation of 1) so that it is easy to compare them directly. The beta coefficients represent the average change in response given a change in standard deviation in the predictor.

After fitting the regression model using standardized beta coefficients, the predictor variable with the largest standardized coefficient is the most important. To see the results faster and more clearly, we created a beta coefficient table arranged in descending order (Excel) in *Figure 3. 7* and a bar chart (STATA) in *Figure 3. 8*.

Explainatory Variables	Beta coefficients
edu	0.276
dep_ratio	0.198
ethnicity	0.1678
log_aland	0.1675
hhsiz	0.115
urban_ethnic	0.1124
urban	0.1095
age	0.107
province	0.034
married	0.033
log_fland	0.021
gender	0.015
log_pland	0.006
log_gland	0.0017

Figure 3. 7. Beta coefficient - sorted in descending order

- **Education**, the *edu* variable, has the **largest standardized** effect - the most important explanatory variable. A one standard deviation increase in education increases income by 0.276 standard deviations. We can conclude that the increase in income of household heads is closely related to their years of schooling.
- The **second** important variables in this model are: The **proportion of dependents in the household**, as *dep_ratio* variable (0,198). This means that a higher dependency ratio (i.e. more people dependent on the head of household) is associated with higher household income.
- The ***ethnicity*** variable (0.1678) and ***log_aland*** variable (0.1675). The ethnicity is also an important predictor when there is a large difference between the income of the Kinh and ethnic minorities *(part a)*. And this shows that owning a lot of land to grow annual crops has a positive effect on income.
- **Moderate predictors** such as: Household size, urban_ethnic interaction, living area: urban or rural, and age have a moderate impact on income. The beta coefficients of these variables range from 0.115 - 0.107.
- **Garden land** area, the *log_gland* variable, has the **smallest standardized** effect. Garden land area has a very small positive impact on income, so it can be ignored.

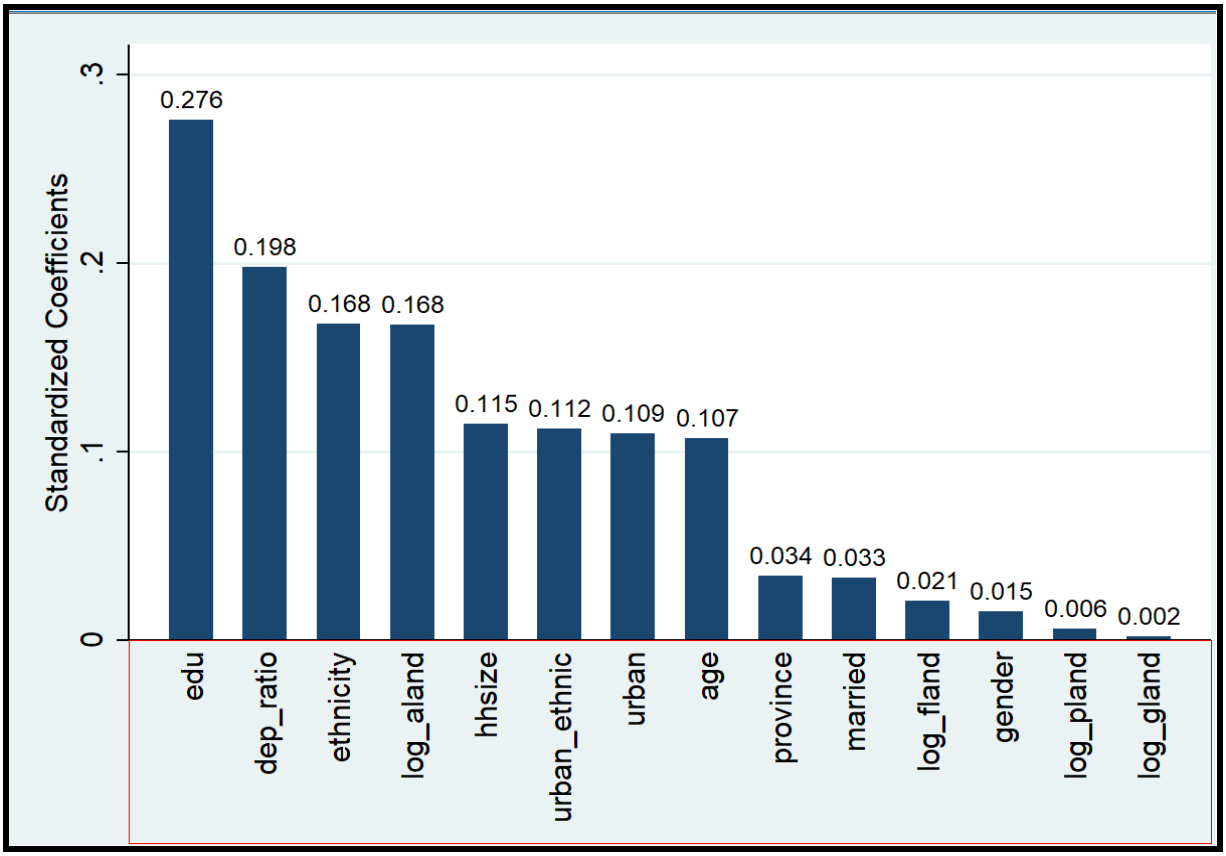


Figure 3. 8. Bar chart of eta coefficient

4. Controlling for other explanatory variables in the model

a, Check the hypotheses about the influence of education

We conduct an instead of the hypothesis that the effect of education on income is greater than 5% and 7%. The regression results of the model include independent variables (Province, Gender, Age, Married, Edu, Ethnicity, Hhsize, Dep_ratio, Urban, Log_aland, Log_pland, Log_Fland, Log_Gland) are presented as follows:

```
// test edu
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	Coef.	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

The regression results show that the education variable has a coefficient of 0.0561575 with a standard deviation of 0.0019386.

To test whether the impact of education is larger than 5% and 7%, we use the following formulas to calculate the t value:

```
. gen t_edu5 = (_b[edu] - 0.05) / _se[edu]

. gen t_edu7 = (_b[edu] - 0.07) / _se[edu]

. display ttail(e(df_r), t_edu5)
.00074902

. display ttail(e(df_r), t_edu7)
1
```

The p-value value corresponds to the theory that the effect of education is greater than 5% and 7% are 0,00074902 and 1. This shows that we rejected the hypothesis for no more than 5%, but did not reject the hypothesis for no more than 7% effect.

b, Check the quadratic relationship between age and log of income

To check the quadratic relationship between age and income log, we take the following steps:

Step 1: Create the square of age

We create a new variable, Age2, the square of age (Age). This allows us to test the quadratic relationship between age and income log.

Step 2: Regress log of income on age and age2

We performed a regression of log income (log_income) on the independent variables age and age2. The regression results indicate a quadratic relationship between age and log income.

```
. gen age2 = age*age

. 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	Coef.	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

Regression results including age and age2 variables

The model revealed a significant relationship between age and age squared with log_income (F(2, 7417) = 269.39, p = 0.000). However, it only explains about 6.78% of the variability in log_income (R-squared = 0.0678), suggesting that other factors also influence income levels. Age positively affects log_income (coefficient 0.074788, p = 0.000), but this effect diminishes as age increases. Age squared has a negative coefficient (-0.0006221, p = 0.000), indicating a declining income increase with age.

Step 3: Test the Shape of the Quadratic Relationship and Identify the Extreme Point

We used the utest command to determine if the relationship between age and log income has an inverse U shape and to identify the extreme point.

```
. 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
```

The utest results show that the quadratic relationship between age and log income has an inverse U shape, with an extreme point at 60.10623 years. This indicates that log income increases up to around 60 years of age and then decreases as age increases further.

Quadratic relationship:

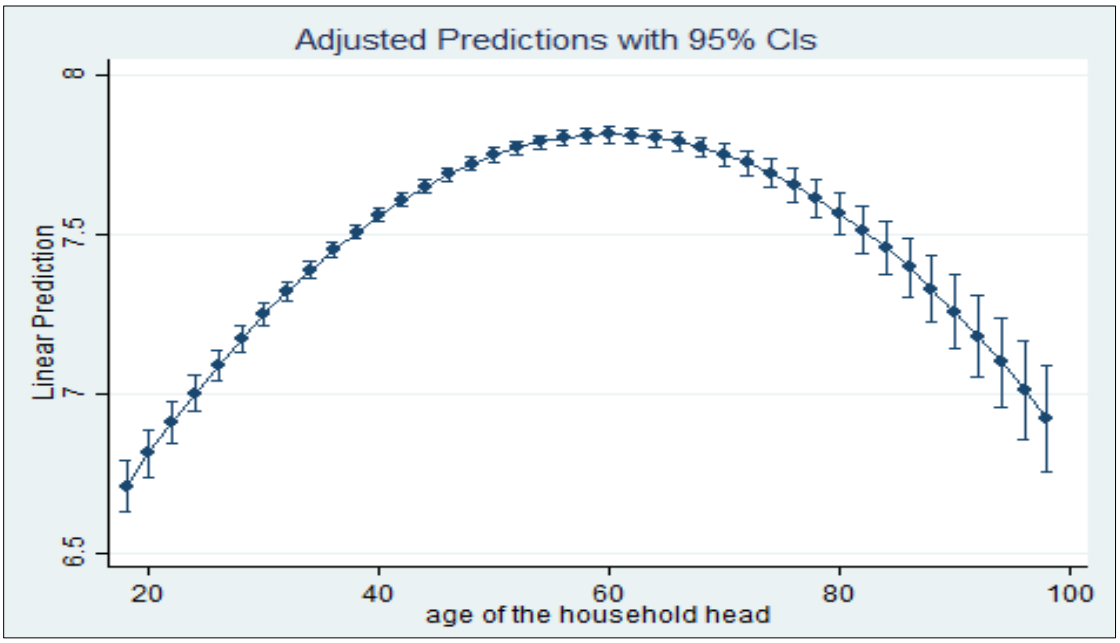
reg log_income c.age##c.age

The dependent is log_income is used to represent income in natural logarithm form. In this context, log_income is the variable we aim to predict or explain.' c.age##c.age' denotes the interaction between c.age and itself (age squared), aiming to model the non-linear relationship between age and log_income.

margins,at(age=(18(2)99))

After the model has been estimated, the margins command, at(age=(18(2)99)) calculates the predicted mean value of log_income at values of age from 18 to 99, with a step of 2 year. This command assumes that all other variables in the model are held at their mean values.

Finally, creates a graph based on the predictions calculated in the previous step. The chart shows the y-axis as predicted values of log income (log_income), and the x-axis as values for age. Use: 'marginsplot'.



Quadratic relationship between age and log_income

The age of the household head significantly influences income, as illustrated by the graph of log income by age. When the household head is younger, their income tends to increase with age. This can be attributed to the accumulation of experience, skills, and career advancement over time.Around the age of 60, household heads reach their highest income level. This is likely the point at which they peak in their careers, achieve financial stability, and benefit from years of hard work and dedication.However, after the age of 60, income begins to decline. The primary reasons for this decrease may be retirement or a reduction in workload, leading to a drop in earnings.

Use **outreg2** to get the regression outputs and then present the results professionally.

running a regression: *reg log_income age*

Use the command *outreg2 using age1.doc* to export the regression results into a document named age1.doc.

Then, uses the **outreg2** command to export the results of the regression to a document called age1.doc. The appen option appends the new results to the existing document if it already exists.

	v1	v2	v3	Notes_Titles
1		(1)	(2)	
2	VARIABLES	log_income	log_income	Standard errors in parentheses
3				*** p<0.01, ** p<0.05, * p<0.1
4	age	0.0112***	0.0748***	
5		(0.000677)	(0.00402)	
6	age2		-0.000622***	
7			(3.88e-05)	
8	Constant	7.079***	5.564***	
9		(0.0339)	(0.100)	
10				
11	Observations	7,417	7,417	
12	R-squared	0.035	0.068	

V.Summary and policy implications

Summary

This study delves into the multifaceted determinants of household income in the Northern Midlands and Mountains region of Vietnam, using data from a survey of 7417 households across 13 provinces. Through a comprehensive econometric model, we examined how various factors, including demographics, education, gender, marital status, ethnicity, household size, dependency ratio, urbanization, and different types of land ownership, influence household income. We can see significant differences in average income between different household groups now in this model. These findings highlight the multifaceted nature of income inequality, influenced by both ethnicity and geography, and suggest targeted policy interventions to close these gaps.

Key Findings:

1. **Education and Income:** Our analysis confirms that education significantly impacts household income, with higher levels of education correlating with increased income. Specifically, the effect of education on income is more than 5% but less than 7%.
2. **Quadratic Relationship between Age and Income:** The relationship between age and income follows an inverse U-shape, peaking around the age of 60. This indicates that income increases with age up to a certain point before declining.
3. **Urban and Ethnic Interaction:** There are substantial income disparities between urban and rural areas, and between Kinh and ethnic minority households. Kinh households in urban areas have significantly higher incomes compared to ethnic minority households in rural areas.

4. **Effect of Land Types on Income:** Different types of land ownership have varied impacts on household income. Annual cropland, forestland, and garden land show negative relationships with income, while perennial cropland shows an insignificant effect.
5. **Regional Disparities:** Income levels vary significantly across different provinces, highlighting the impact of regional economic development and infrastructure on household income.

Policy Implications

1. **Education Policies:** Enhancing access to education and improving the quality of education in rural and ethnic minority areas can help bridge the income gap. Policies should focus on increasing school attendance and providing vocational training to equip individuals with skills for better-paying jobs.

2. **Support for Elderly Workforce**: Given the peak income age of around 60, policies should support the elderly workforce through continuous training and health programs to maintain their productivity and earnings.

3. **Urban-Rural Development Programs**: To reduce the income disparity between urban and rural areas, investments in rural infrastructure, such as transportation, healthcare, and education, are essential. Encouraging businesses to invest in rural areas can also create more job opportunities.

4. **Land Use and Agricultural Policies**: Revisiting land use policies and providing support for diversified agricultural activities can enhance income from land. Access to credit and financial support should be facilitated for farmers to improve productivity and income.

5. **Regional Economic Development**: Targeted economic development programs are needed for lower-income provinces. Government support and incentives can attract investments, improve infrastructure, and create job opportunities in these regions.
6. **Gender and Ethnic Equality**: Addressing socio-cultural barriers that limit income opportunities for women and ethnic minorities is crucial. Programs that promote gender equality and support ethnic minorities in accessing education and resources can help mitigate income disparities.

VI. References

[1] Minitab Blog Editor. (2016). How to Identify the Most Important Predictor Variables in Regression Models. Retrieved June 10, 2024, from Minitab.com website:

https://blog.minitab.com/en/adventures-in-statistics-2/how-to-identify-the-most-important-predictor-variables-in-regression-models%20

[2] Benin, S., & Josee Randriamamonjy. (2008). Estimating Household Income to Monitor and Evaluate Public Investment Programs in Sub-Saharan Africa. Retrieved June 10, 2024, from ResearchGate website: https://www.researchgate.net/publication/5056700_Estimating_Household_Income_to_Monitor_and_Evaluate_Public_Investment_Programs_in_Sub-Saharan_Africa

[3] Socio-Economic Determinants of Household Income among Ethnic Minorities in the North-West Mountains, Vietnam. (2015). *Croatian Economic Survey*.

VII. Contribution

Name	Student ID	Role	Contribution
Le Thi Diem Quynh	20070974	Leader	Write report part 1, 2, 4, 6
Nguyen Minh Hien	20070928	Member	Write report part 1, 2, 4, 6
Nguyen Thi Thu Thao	20071078	Member	Write report part 1, 3, 5, 6
Phung Linh Chi	21070688	Member	Write report part 1, 3, 5, 6