NATIONAL UNIVERSITY OF ENGINEERING

COLLEGE OF ECONOMICS AND STATISTICAL ENGINEERING AND SOCIAL SCIENCES

STATISTICAL ENGINEERING PROGRAM

EF012 PROJECT WORKSHOP

Design of a Statistical Processing System for the Quantile Regression Analysis of the Distribution of Total Monthly Household Expenditures in Peru Regions.

by:

David Alex MILLA NOA

Professor

Ivan Victor SILVA GUILLEN

LIMA – PERU

2020- II

Abstract

In order to apply the quantile regression technique in the analysis of the total monthly expenses in education, health, home and luxury of the houses in Peru and to obtain what factors influence these expenses, a database was built with the information from the survey National Household (ENAHO). The data collected is for the annual period 2019. Using quantile regression for this type of study is convenient because the results are continuous and when the effects of the variables may differ in the distribution of expenditures in households in Peru. Linear regression methods estimate only the effects at the middle level, which can be an incomplete and biased summary of the effect of exposures for some ongoing expense outcomes. To estimate the total monthly expenses for each module of expenses, the method of minimum absolute deviations was used, which will allow us to obtain the estimated values of the total monthly expenses in education, health, home and luxury in housing in Peru for each quantile given in their distribution of expenses and the relationship with the explanatory variables. The model obtained shows that the influence of the explanatory variables on the total monthly expenses for each module varies considerably from one quartile to another, so that some of them are significant only in some quantiles, such as the variable stratum, type of employment and marital status. Thus, it is also observed that the variable housing situation in the total monthly expenses in education, health, home and luxury does not observe a significant difference between the effects estimated by the multiple linear regression and the effects estimated by the quantile regression. The results indicate that the data best fit through quantile regression is in the quantile 0.90 to 0.95 for the different aforementioned spending modules, which is consistent with the theoretical framework developed in this research work. In addition, the underestimation and overestimation of the quantile regression, its approximation by + - 20% with respect to the estimates, behaves in a better way in the Basic Expenses 2 and Household Expenses models for both quantiles (Q40 and Q50) compared to the multiple linear regression.

Keywords: Quantile regression, Quantile, Underestimation, Overestimation

General Index

Chapter 1

INTRODUCTION

1.1. Description of the problem situation

Currently, due to the COVID-19 pandemic, income in Peruvian families has been reduced and therefore expenses in Peruvian families. Thus, it is also interesting to analyze the expenses because the cash balance of each of the households can be known, since there is information on the income of the families, therefore, it is possible to have the ability to over-indebtedness that the companies have. Peruvian families, which is very relevant for different companies, such as requesting a consumer loan, card, mortgage or of a different nature. For this reason, it is of interest to know the expenses generated by Peruvian families, the factors that explain it: Number of members per household, the stratum to which the family belongs, the type of occupation of the head of the household, the age of the head of the household, housing situation, where the dwelling belongs, the marital status of the head of the household and income.

To reinforce this problem, according to data obtained from the National Institute of Statistics and Informatics (INEI) at the end of 2017, it obtained that the informal sector represents 70% of the economically active population, and this problem is notoriously prevalent and highlighted by former President Martin Vizcarra in the quarantine that was presented in 2020.

This research work will analyze the relationship of expenses in Peruvian households and income, members by household, age, stratum, marital status, housing situation, type of occupation and

location all with respect to the head of the household, this due to that we will work with the annual data from the national household survey (ENAHO) by means of a model. However, the expenses are very heterogeneous and with the presence of atypical observations and could overvalue the OLS estimate. Therefore, an estimate through quantile regression is proposed as an alternative.

It should be noted that some companies use the estimates of the long-term quantile regression as input, therefore, the productivity of the quantile regression model will be analyzed.

1.2. Formulation of the research problem (general and specific)

General problem:

Which regression model adequately explains the total monthly expenses in education, health, home and luxury of homes in Peru during the annual period of 2019?

Specific problems:

In which quantile of the quantile regression does the data best fit to estimate total monthly expenditures on education, health, home and luxury of homes in Peru?

What factors of the regression influence the total monthly expenses in education, health, home and luxury of the houses in Peru?

What factors are significant in both quantile regression and multiple linear regression in total monthly expenditures on education, health, home and luxury of homes in Peru?

What predictions is the best alternative for the total monthly expenses in education, health, home and luxury of the houses in Peru?

1.3. Objectives of the investigation

General objective:

Estimate the quantile regression model for the total monthly expenses for education, health, home and luxury of homes in Peru in the period 2019.

Specific objectives:

1. Evaluate and determine in the quantile regression, in which quantiles present the best fit in the total monthly expenses of education, health, home and luxury.

2. Determine what factors influence the total monthly expenses for education, health, home and luxury, for different quantiles.
3. Determine those factors that are significant in all quantiles of the quantile regression including the multiple linear regression.
4. Compare that the predictions by the quantile regression in certain quantiles is a better alternative than the multiple linear regression in the total monthly expenses of education, health, home and luxury.

1.4. Justification, scope and limitations of the research

Justification:

The reasons why the investigation will be carried out is that due to the situation we are going through due to COVID-19, expenses have varied due to different factors, the investigation work will serve as input for subsequent investigations, when annual information is available from the ENAHO 2020, since so far there is information from the first two quarters, however it is not enough, since there are not the different modules of expenses necessary for a better analysis. Likewise, the research work will serve for the analysis in which a person wants to estimate the quantile regression for a given quantile which is more convenient for their study.

Scopes:

The research covers only homes in Peru, taking as a sample through surveys in all departments of Peru, providing information from ENAHO in 2019.

Limitations:

The period of time for collecting information for an exhaustive analysis comprises one year, therefore, ENAHO data prior to 2020 will be analyzed, since due to the COVID-19 pandemic, no information has been collected since second quarter of 2020.

.

Chapter 2

Theoretical Framework

2.1. Investigative Background

QUANTY REGRESSION ANALYSIS FOR THE DISTRIBUTION OF TOTAL MONTHLY INCOME OF THE ECONOMICALLY ACTIVE EMPLOYED POPULATION OF METROPOLITAN LIMA (Richard Henry Huiman Morales)

In the research work, he proposes the quantile regression method to analyze the distribution of the total monthly income of the employed population of metropolitan Lima.

Problem formulation:

What is the relationship between sex, age, educational level, number of hours worked per week and total monthly income during the period August-September-October 2016?

General objective:

Determine if the sex, age, educational level, number of hours worked per week with the total monthly income during the period August-September-October 2016.

Estimation by quantile regression.

The results of the different estimates made are appreciated. Between the columns we can see the results for the quantiles 0.25, 0.5, 0.75. In all the estimates raised, it was determined that all the variables that are involved in the model obtained statistically significant parameters. Age, total hours worked per week and educational level positively influence total monthly income. Focusing on the results obtained, the female gender variable has a negative influence on income, that is, their income is reduced approximately from 12% to 13% with respect to the total monthly income of

men. The educational level variable is also relevant, because said variable with total monthly income increases with the group of people who receive their high monthly income (Q75), the employed population that has a higher university education level increases a total income monthly by 47% more than the employed population that has a primary educational level (reference), this value is lower than the group of people who have a low total monthly income (Q25), who manage to increase by 39.5% more compared to people with a primary educational level. The relationship between total hours worked per week and total monthly income tends to decrease as we move into segments of the employed population with high total monthly income, going from 0.89% to 0.54%, the opposite is true in relation to the age. Increasing age has a positive impact on total monthly income.

Table 1. *Quantile regression results (coefficients and standard errors)*

Quantile	0.25		0.5		0.75		OLS	
	coeff	Standard error	coeff	Standard error	coeff	Standard error	coeff	Standard error
Intercept	2.3624	0.0210	2.5842	0.0189	2.7594	0.0180	2.4922	0.0178
Gender (Female)	-0.1259	0.0080	-0.1232	0.0062	-0.1294	0.0065	-0.1369	0.0062
Age	0.0010	0.0000	0.002	0.0002	0.0031	0.0002	0.0022	0.0002
Total Hours Worked	0.0089	0.0002	0.0071	0.0002	0.0054	0.0002	0.0810	0.0002
Educational Level (High School)	0.1233	0.0113	0.0878	0.0118	0.0795	0.0102	0.1127	0.0109
Educational Level (Higher Non-University)	0.2257	0.0143	0.1923	0.0131	0.2069	0.0120	0.2376	0.0122
Educational Level (Higher University)	0.3958	0.0149	0.4138	0.0145	0.4769	0.0130	0.4660	0.0118

Source: Based on information from ENAHO 2016
Elaboration: Richard Henry Huiman Morales

Where it concludes the positive influence of the educational level and the male sex in the higher income generally perceive better income; and women receive, on average, a lower total monthly income than men. It is also concluded that the employed population of metropolitan Lima who receive high incomes do not necessarily need to have long hours worked per week for their income to be higher.

EXPLANATORY MODEL OF THE EVOLUTION OF THE POPULATION OF THE MUNICIPALITIES OF EXTREMADURA THROUGH THE QUANTILE REGRESSION TECHNIQUE (Sanchez, 2011)

The objective of the research is to determine and quantify the variables that have influenced the evolution of the total population of Extremadura in the last ten years as a means of knowing the causes of the depopulation situation in which the Extremadura region is found.

To this end, a total of six least quadratic models are built in which the variation of the total population of all the municipalities of Extremadura between 2000 and 2010 is taken as an endogenous variable; and as explanatory variables different combinations of the following: the population of the municipality in the initial year, the average population age, a variable of spatial neighborhood, another of relative location Madrid-Portugal and a series of synthetic indices (of employment, of equipment of the home, characteristics of the home, accessibility and equipment and basic services of the municipality). The heteroscedasticity inherent in many models with great territorial disaggregation, such as this case, renders the least squares estimators inefficient, invalidating the results obtained. As an alternative to the OLS technique and in order to overcome the difficulty that the great heterogeneity present in the data brings to the model, a quantile regression is estimated that establishes the relationship between the population variation between the years 2000 and 2010 and a series of explanatory variables for the three quartiles of the endogenous variable with the following specifications:

$$y_i = \beta_{0;0.25} + \beta_{1;0,25} edadmedia_i + \beta_{3;0,25} ihogar_i + \beta_{5;0,25} xyctrd_i + u_{i;0,25}$$

$$y_i = \beta_{0;0.50} + \rho_{0,50} Wy_i + \beta_{1;0,50} edadmedia_i + \beta_{2;0,50} iacces_i + \beta_{3;0,50} ihog_i + u_{i;0,50}$$

$$y_i = \beta_{0;0.75} + \rho_{0,75} Wy_i + \beta_{1;0,75} edadmedia_i + \beta_{2;0,75} iacces_i + \beta_{3;0,75} ihog_i + \beta_{4;0,75} iviv_i + u_{i;0,75}$$

Where "y_i" is the variation of the population in the period 2000-2010 for municipality i; "$edadmedia_i$ " is the average age of the municipality; "$ihog_i$" is the municipal home equipment index that includes aspects such as the percentage of homes with heating, number of vehicles per home or telephone lines, among others; "$xyctrd_i$" indicates the relative distance of the municipality to Madrid; "$iacces_i$" is an index of accessibility of the municipality that includes variables such as the average travel time to the workplace, the percentage of homes with poor communications, etc.; "Wy_i" is the endogenous variable spatially displaced, so that it indicates the average variation of the population of the municipalities neighboring municipality i; "$iviv_i$" represents the index of

characteristics of the dwelling (street cleaning, green areas, crime ...); and finally "u_i" is the random disturbance term.

Table 2. Quantile regression results (coefficients and standard errors)

	Spatial quantile regression		
	(Q_1)	(Q_2)	(Q_3)
C	40,917*** (6,607)	29,854*** (10,148)	-0,522 (8,712)
edad_ media	-1,318*** (0,124)	-1,247*** (0,162)	-0,968*** (0,148)
Xyctrd	-6,28E-12*** (2.36E-12)		
Iviv			0,129** (0,059)
Iacces		0,084*** (0,038)	0,114** (0,046)
Ihog	0,117** (0,036)	0,103*** (0,035)	0,179*** (0,032)
Wvpob		0,211*** (0,066)	0,223** (0,100)
Pseudo R^2	0,279	0,280	0,271

Source: Based on data from Spain
Elaboration: Beatriz Sánchez Reyes

The model obtained shows that the influence of the explanatory variables on the endogenous varies considerably from one quartile to another, so that some of them are significant only in some quartiles, which is why we work with a different specification in each of them.

Thus, for the first quartile (where municipalities with greater population declines or lower increases - to a large extent the smallest -) have the greatest weight, the municipal average age and the variable of proximity to Madrid have a negative influence on the evolution of the population. population, while the household equipment index shows a positive sign; In the median regression (second quartile), the mean age and the household equipment index behave the same as in the first quartile, but there is also a positive relationship between the endogenous variable and the

accessibility index and the population evolution of the municipalities neighbors; Finally, in the third quartile (where the municipalities with the highest population gains have the greatest weight) the results have the same interpretation as in the median regression but adding the index of housing characteristics that appears with a positive sign as an explanatory variable. These results show the fact that a single least quadratic regression would not be sufficient to explain the peculiarities in the behavior of the population evolution for the different types of municipalities - classified according to the intensity of the depopulation to which they are subjected-; quantile regression allows this task to be carried out without being affected by the great heterogeneity present in the data.

DETERMINING FACTORS OF HOUSEHOLD ELECTRICITY DEMAND IN SPAIN: AN APPROACH THROUGH REGRESSION QUANTIL (Vicéns and Medina, 2011)

The objective of the study is to identify the determining factors of household electricity consumption that should be taken into account when defining energy saving policies. Specifically, it seeks to quantify the importance that income would have in the design of future energy saving policies, aimed at the introduction of tax figures that tax electricity consumption.

To do this, we work with data from the 2009 Family Budget Survey, published by the INE, annually and with a sample size of 24,000 households. An econometric model is proposed where the different patterns registered by household electricity expenditure are explained by:

- Socio-economic factors of household members: number of members and monthly household income.
- Characteristics of the home: size of the home in m2, year of construction, area of residence, type of house, and a variable that measures the use of electricity as a source of energy used to heat the home.
- Geography or location of the home within the national territory: through Autonomous Community dummy variables, all unobserved effects related to geographic location are collected, such as the climate that affects both the use of heating and air conditioning equipment.

$$\log(\text{electricity expense})_i = \beta_0 + \beta_1 \log(n^{\circ}\ \text{members}) + \beta_2 \log(\text{entry}) + \beta_3 \log(\text{surface}) + \beta_4 \text{antiquity} + \beta_5\ \text{Residence area} + \beta_6 \text{type of dwelling} + \beta_7\ \text{electric heating} + \beta_8\ \text{autonomous community} + \varepsilon i$$

As a first estimation alternative, the least square is proposed. However, a descriptive analysis of the variable to be modeled makes it possible to identify other more suitable alternatives.

Although an average Spanish household spends 49 euros a month on electricity, (2.4% of its income) there is a high asymmetry in the sample distribution, characterized by a greater concentration of households with low consumption. In turn, household electricity consumption patterns are very heterogeneous and with a high presence of atypical observations, especially in the upper part of the distribution. Specifically, the 1% of the households that consume the most, allocate 10% of their income to electricity consumption (202 euros per month), a figure that is very high. These characteristics of household electricity consumption could be overestimating the OLS estimate of income elasticity. As an alternative, an estimate is proposed through median quantile regression, which will allow more reliable results to be obtained insofar as it does not require compliance with the basic hypotheses required in the least squares estimate and estimates far from the median plane of regression will not affect directly. The results confirm the starting hypothesis, since although both estimates (OLS and median quantile regression) yield the same results in terms of the statistical significance of the parameters, they differ in the amount of the estimates obtained. The greatest differences are recorded in the estimate of income elasticity, which goes from 0.16% in the OLS estimate to 0.12% in the median regression.

The results of the study allow us to affirm that the size of the household (number of members) and of the dwelling (m2) are the variables that most influence the electricity bill. Thus, and under the "ceteris paribus" condition, a new member of the household represents an increase in the electricity bill of 13%, just over 6 euros per month, while ten additional meters of housing increase electricity costs by 4%, equivalent to 2 euros per month.

On the contrary, the low role of income stands out, which makes it possible to define electricity as a staple good, with an elasticity close to zero and where changes in income do not produce significant changes in household electricity consumption habits. Specifically, a 10% reduction in the household income level, which would mean an approximate drop of 200 euros per month, would generate savings in the electricity bill of only 0.6 euros per month, a figure that is insignificant.

This result provides evidence of the direction that energy saving policies should not take in their search for energy efficiency and environmental protection. The introduction of tax figures that tax consumption, or measures aimed at greater control of electricity pricing, seem to be doomed to failure in the short term, insofar as variations in income will not entail significant changes in electricity consumption habits of households. In this scenario, active demand management policies are becoming increasingly important, seeking to optimize the hourly curves of electricity consumption by reducing the value at the maximum power peaks by transferring them to off-peak hours.

CHANGES IN THE US SALARY STRUCTURE BETWEEN 1963 AND 1987: AN APPLICATION WITH QUANTY REGRESSION (Buchinsky, 1994)

The objective of the study is to analyze salary changes due to differences in levels of education and experience for different points of the distribution of wages in the United States between 1964 and 1988. Precisely because what is interesting are differences throughout the distribution of salaries, the quantile regression is presented as the best alternative since it will allow us to know how the levels of education and experience affect the different salary levels present in the sample.

The data for this purpose come from the US Population Surveys (March Current Population Survey) carried out between the years 1964 and 1988 of around 1,000,000 people.

The sample for the study contains between 10,000 and 34,000 observations per year that includes all men (both black and white) between the ages of 18 and 70 and who meet the following conditions:

a) That you work for at least one week in the year prior to the survey.
b) That they work in both the public and private sectors, excluding self-employed workers and those who do not receive compensation for their work.
c) Make them at least $ 50 a week in 1982 prices

The author constructs two explanatory models of the variation of wages:

a) **Model 1:** The endogenous variable (log of weekly wages) is explained by education, experience, experience squared, and a binary variable that determines the person's race. The model is obtained

for five quantiles of the wage distribution and the 25 years for which the study is carried out (1963-1987), so that a regression per year and quantile is obtained.

b) **Model 2:** In this second model, the author segments the sample into a total of 16 groups attending to different degrees of education and experience. In addition, he adds explanatory variables on model 1: a binary variable for part-time work, another for the metropolitan area, and several dummies related to the different levels of education and residence in different regions. Here, for each of the 16 groups, five regressions (corresponding to the five quantiles) are obtained for each of the years considered (1963-1987).

In general lines, by way of conclusion of the two models, Buchinsky concludes that the changes in wages due to education and experience follow quite similar patterns in all quantiles, although with substantial differences between the different levels considered, thus, the Changes in wages due to education are greater in the higher quantiles, while changes due to experience are greater in the lower quantiles.

2.2. Conceptual framework.

In general, empirical studies are interested in analyzing the behavior of a dependent variable given the information contained in a set of regressors or explanatory variables.

A standard approach is to specify a linear regression model and estimate its unknown parameters using the OLS (Ordinary Least Squares) method or the LAD (Absolute Least Deviation) method.

- The OLS method estimates the parameters by minimizing the sum of the squared errors and leads to an approximation of the mean function of the conditional distribution of the dependent variable.
- The LAD method minimizes the sum of the absolute errors and leads to an approximation of the conditional median function.

Although the mean and median are two important measures of location that represent the average behavior or central tendency of a distribution, they account for very little about the behavior at the tails of the distribution.

A new point in regression analysis is the quantile regression approach proposed by Koenker and Basset (1978).

- This approach allows estimating different quantile functions of the conditional distribution, including the median function as a special case.
- Each quantile function characterizes a particular point in the conditional distribution.

Thus, by combining different quantile regressions, we have a more complete description of the underlying conditional distribution.

Quantile function:

For any t e (0,1) and for any random variable Y (continuous and discrete):

The t-th quantile of Y can be defined as:

$$\mathcal{E}_t \in \mathbb{R} \mid P(Y < \mathcal{E}_t) \leq t \leq P(Y \leq \mathcal{E}_t)$$

At least t percent of the probability mass of "y" is less than or equal to $\mathcal{E}t$ and at least (1-t) percent of the probability mass of "y" is higher than $\mathcal{E}_t$.

If "Y" is a random variable it is defined as:

$P(Y \leq y) = F_Y(y)$ Continuous function to the right

$Q_Y(t) = F_Y^{-1}(y) = inf\{y | F_Y(y) \geq t\}$ Continuous function to the left

Q.....F

$t \in (0,1)$ $\Re$ Inverse of accumulation $\Re$ (0,1)

For any $t \in (0,1)$, $Q_Y(t)$ gives us the t-th quantile of Y not conditional

- Increasing monotonous.
- g continuous function to the left $\forall\ t \in (0,1)$ $P(Y \leq Q_Y(t)) = P\ (g(Y) \leq g(Q_Y(t))) = t$
- For a continuous random variable Y, the probability density function is defined:

$$f_Y(y) = \frac{d\ F_Y(y)}{dy}$$

- Similarly, for the quantile function we have:

$$S_Y(t) = \frac{dQ_Y(t)}{dt}$$ Quantile density function

Some additional formulas

$$\frac{dF_T(F_Y^{-1}(t))}{dt} = f_Y(F_Y^{-1}(t))\frac{dF_Y^{-1}(t)}{dt} = 1$$

$$\frac{dF_Y^{-1}(t)}{dt} = \frac{1}{f_Y(F_Y^{-1}(t))}$$

$$\frac{dQ_Y(t)}{dt} = \frac{1}{f_Y(F_Y^{-1}(t))}$$

Reciprocal of the density function evaluated at the quantile of interest.

$$\mathrm{S_Y(t)} = \frac{1}{f_Y(F_Y^{-1}(t))}$$

Empirical quantiles

Sea $y_1, y_2, \dots, y_n$ a random sample, its empirical distribution function is defined by the ratio between the number of observations less than or equal to the value of interest and the total number of observations.

$$F_Y(y) = \frac{\#\,(y_i \le y)}{n}$$

Similarly, the empirical quantile function can be defined as:

$$Q_Y(t) = F_Y^{-1}(y) = inf\left\{y | \frac{\#\,(y_i \le y)}{n} \ge t\right\}, 0 < t < 1$$

In order to obtain the desired quantile:

- Order sample
- Check in which observation the threshold is reached
- Method for calculating quantiles

$$Q_Y(t) = argmin_{\varepsilon_t \in \Re}\left[\sum_{i \in \{i | y_i \ge \varepsilon_t} t|y_i - \varepsilon_t| + \sum_{i \in \{i | y_i < \varepsilon_t} (1-t).|y_i - \varepsilon_t|\right]$$

Now, the verification function allows to reformulate the objective function

$\rho_t(u) = u(t - I(u < 0))$ where I: random variable

- Si u < 0 -> $\rho_t(u) = u(t-1)$
- Si u ≥ 0 -> $\rho_t(u) = ut$

  Where 0 < t < 1

$Q_Y(t) = argmin_{\varepsilon_t \in \Re} \sum_i (y_i - \varepsilon_t) \rightarrow$ Provides the desired quartile

Hence, the optimum of the loss function provides the desired quantile. Expected value of the loss function:

$$E[\rho_t(y-\varepsilon_t)] = t\int_{\varepsilon_t}^{\infty}(y-\varepsilon_t)dF(y) - (1-t)\int_{-\infty}^{\varepsilon_t}(y-\varepsilon_t)dF(y)$$

Taking the derivative with respect to $\mathcal{E}_t$ we have:

$$\frac{\partial E[\rho_t(y-\varepsilon_t)]}{\partial \mathcal{E}_t} = t\frac{\partial \int_{\varepsilon_t}^{\infty}(y-\varepsilon_t)\partial dF(y)}{\partial \mathcal{E}_t} - (1-t)\frac{\partial \int_{-\infty}^{\varepsilon_t}(y-\varepsilon_t)\partial dF(y)}{\partial \mathcal{E}_t}$$

$$= -t\int_{\varepsilon_t}^{\infty}\partial F(y) + (1-t)\int_{-\infty}^{\varepsilon_t}\partial F(y)$$

$$= -t\big(1-F(\varepsilon_t)\big) + (1-t)(-0+F(\varepsilon_t))$$

$$= -t + tF(\varepsilon_t) + F(\varepsilon_t) - tF(\varepsilon_t)$$

Making the derivative equal to zero we have: $F(\mathcal{E}_t) = t$

Therefore, the expected loss function is convex and is minimized only if we have that the function is "t".

Quantile regression

Once the point of how to determine empirical quantiles has been studied, the question that arises is: How could this new formulation be used in regression analysis?

When using a linear regression model:

$$y_i = X_i'\beta + u_i, i = 1,2,\ldots.T$$

Where is it supposed to $E[u_i|X_i] = 0 \Rightarrow E[y_i|X_i] = X_i'\beta$

The vector of parameters β can be estimated by:

$$\text{OLS: } \beta = argmin_{\beta\in\mathbb{R}^k} \sum_i (y_i - X_i'\beta)^2$$

Suppose

$$y_i = \mathrm{X_i'}\beta + u_{i,T} \quad i = 1,2,\ldots..T$$

The "t" -th quantile of the error term conditional on the regressors is zero, not the expected value.

$$Q_t(u_{i,t}|X_i) = 0 \quad \rightarrow \quad Q_t(u_{i,t}|X_i)$$

$$Q_{u_{i,t}}(t|X_i)$$

$$Q_{u_{i\,(t)}}(t|X_i)$$

The t-th conditional quantile of y_i with respect to X_i can be written

$$Q_{u_{i\,(t)}}(t|X_i) = \mathrm{X_i'}\beta_t$$

Putting together the following equations:

$$Q_Y(t) = argmin_{\mathcal{E}_t \in \Re}\left[\sum_{i\,\in\{i|y_i \geq \mathcal{E}_t} t|y_i - \mathcal{E}_t| + \sum_{i\,\in\{i|y_i < \mathcal{E}_t} (1-t).|y_i - \mathcal{E}_t|\right]$$

$$Q_Y(t) = argmin_{\mathcal{E}_t \in \Re}\sum_i \rho(y_i - \mathcal{E}_t)$$

$$Q_t(y_i|X_i) = \mathrm{X_i'}\beta_t$$

For any $t \in (0,1)$ the vector of parameters β can be estimated as follows:

$$\beta_t = argmin_{\beta_t \in \mathbb{R}^k}\left[\sum_{i\,\in\{i|y_i \geq \mathrm{X_i'}\beta_t} t|y_i - \mathrm{X_i'}\beta_t| + \sum_{i\,\in\{i|y_i < \mathrm{X_i'}\beta_t} (1-t).|y_i - \mathrm{X_i'}\beta_t|\right]$$

$$= argmin_{\beta_t \in \mathbb{R}^k}\sum_i \rho(y_i - \mathrm{X_i'}\beta_t)$$

All observations above the hyperplane estimated by Xβ_t,, that is, the absolute difference between yi and X′β_t are weighted by t and all observations below are weighted by (1-t). The conditional median is obtained when t = 0.5

$$\beta_t = argmin_{\beta_{0.5} \in \mathbb{R}^k}\sum_i |y_i - \mathrm{X_i'}\beta_{0.5}|$$

The estimator of β of the t-th quantile can be obtained by minimizing its sample counterpart. That is, it can be understood as the asymmetric weighted average of the absolute errors, with weights t on positive errors and (1-t) on negative errors:

$$V_T(\beta;T) = \frac{1}{T}\left[t \sum_{i:\, y_i \geq X_i'\beta_t} |y_i - X_i'\beta| + (1-t) \sum_{i:\, y_i < X_i'\beta_t} |y_i - X_i'\beta|\right]$$

Using the function ρt we have:

$$V_T(\beta;t) = \frac{1}{T}\sum_{i=1}^{T} \rho_t(y_i - X_i'\beta)$$

$$= \frac{1}{T}\sum_{i=1}^{T} (t - 1_{\{y_i - X_i'\beta < 0\}})(y_i - X_i'\beta)$$

The first-order condition of the minimization of $V_T(\beta; t)$ is: $= \frac{1}{T}\sum_{i=1}^{T}(t - 1_{\{y_i - X_i'\beta < 0\}}) = 0$

Except in $y_i = X'\beta$ since the derivative is undefined

Solving for β gives β - the t-th quantile regression estimator for β

Once β is obtained, the quantile regression hyperplane and the residuals are estimated:

$$X_i'\beta$$

$$e_i(t) = y_i - X_i'\beta$$

The more quantile regressions that are estimated, the better the shape of the conditional distribution can be understood.

If the median regression line differs significantly from that obtained through OLS (mean), the distribution is skewed.

The conditional distribution is skewed to the left if the upper quantile lines are very close to each other compared to the lower quantile lines.

In general, it can be found that the estimated quantile regressions differ from each other across the quantiles. Which suggests that the explanatory variables may have different impacts on the dependent variable. That is, the impact depends on the location of the conditional distribution.

Calculation of the estimator

The quantile regression estimator is not easy to calculate because the objective function is not differentiable, therefore standard numerical optimization methods are not easily applicable.

In practice, quantile regression estimation is usually carried out by solving a linear programming problem.

$$y_i = \mathrm{X_i'}\beta + e_i = \sum_{j=1}^{k} x_{i,j}\left(\beta_j^+ - \beta_j^-\right) + \left(e_i^+ - e_i^-\right)$$

Where βj is the j-th coefficient of β such that $\left(\beta_j^+ - \beta_j^-\right)$

$\beta_j^+ = \max(\beta_j, 0)$ Positive part

$\beta_j^- = -\min(\beta_j, 0)$ Negative part

In the same way, we have ⟶ $e_i = (e_i^+ - e_i^-)$

Let e^+ ⟶ The $vector$ e_i^+

Let e^- ⟶ The $vector$ e_i^-

$Z = [\beta^+, \beta^-, e^+, e^-]$ ⟶ Of dimensions 2 $(K + T)$ of non-negative elements

$$A = [X, -X, I_T, -I_T] \longrightarrow T * 2(K + T)$$

From the above we have the following non-linear specification:

$$Y = X(\beta^+ - \beta^-) + (e^+ - e^-)$$

Y=AZ

Where:

$Y = \{yt\} \longrightarrow dimension\ TX1$

X ⟶ dimension $T * K$ t- th row ⟶ $X't$

It defines:

$$c = [0', 0', ti', (1-t)i']'$$

$$0 \longrightarrow k - dimensional$$

$$i \longrightarrow T - dimensional$$

The objective function $V_T(\beta; t) \longrightarrow \frac{1}{T} c'Z$

Minimize $V_T(\beta; t)$ is equivalent to minimizing $c'Z$ with respect to Z, subject to the restriction that Y = AZ and that Z does not contain negative elements.

To solve the linear programming problem, Barrodale and Roberts (1974) designed an algorithm based on the simplex method for LAD estimation. This was later extended by Koenker and d'Orey (1987) for quantile estimation.

Finally, Hao and Naiman (2007) citing Koenker and Machado (1999) suggested measuring the goodness of fit by comparing the sum of the weighted distances for the model of interest with the sum in which only the intercept of the parameter appears. In quantile regression, to obtain the goodness of fit of the model, the Pseudo R^2 is used as a measure equivalent to the R^2 of the OLS, which measures the relative success of the quantile regression model. It can be interpreted as a measure of local goodness of fit, for a particular quantile.

$$Pseudo\ R^2 = 1 - \frac{\sum_{i=1}^{n} |Y_i - \hat{Y}_i|}{\sum_{i=1}^{n} |Y_i - Y_\theta|}$$

Where:

$\hat{Y}_i$: Estimation observation (i) of the quantile line

Y_i: actual observation

Y_θ: Percentile given to actual observation

Multiple Linear Regression

This model is an extension of the simple linear regression model, where it was assumed that the dependent variable influences only one independent variable. In general, this case in practice is too scarce, so a generalization is made for the case where there is more than one independent or predictor variable that influences a dependent or predicted variable, which we will call as multiple linear regression. In general, the response variable Y can be related to k independent variables x_1, x_2, x_3, x_k, in this case the model is represented by:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \ldots \ldots \ldots . + \beta_k x_k + \epsilon$$

Where the coefficients β_j, where j=0,1,2,….k are unknown constants and are the parameters of the model. Each βj, represents the expected change in response Y times the unit change in x_j when all other independent variables x_j $(i \neq j)$ remain constant, furthermore ϵ is a random error component.

In the case of multiple regression models, it is preferable to use the matrix notation, since this form allows to express the model in a more compact way and that with a little knowledge of matrix algebra the results are obtained very quickly.

$$Y = X\beta + \epsilon$$

Finally, the goodness of fit of the multiple linear regression model can be used by using the R^2 statistic or coefficient of determination expressed by.

$$R^2 = \frac{SC_{regresion}}{SC_{total}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where:

y_i: Real observation

$\bar{y}$: $El\ promedio\ de\ las\ observaciones$

$\hat{y_1}$: Regression observation estimation

Which measures the proportion of the total variability explained by the proposed regression model, or the total proportion that is due to the regression. This proportion is expected to be high and close to 100% and only a small part is due to error.

Chapter 3

Methods and Materials

3.1. Methods

3.1.1 Research method.

The type of research will be quantitative, the research level is statistical, and the design will be non-experimental and observational, due to the fact that the independent variables are not deliberately manipulated.

3.1.2 Research design

The ENAHO sample for the periods analyzed is probabilistic, of areas, stratified, multistage and independent in each department of study, below, the detail of the sample size per period.

3.1.3 Population and sample

The population is all private dwellings and their occupants residing in the urban and rural areas of the country.

The sample in the period 2019 was 36,994 private dwellings, corresponding to 12,514 dwellings in urban areas and 24,480 dwellings in rural areas throughout the country.

3.1.4 Inclusion and exclusion criteria Inclusion criteria

- The members of the family home
- Domestic workers with bed inside, whether or not they receive payment for their services.
- The members of a family pension that has a maximum of 9 pensioners
- People who are not members of the family household, but who were present in the household in the last 30 days.

Exclusion criteria

- The members of a family pension that has 10 or more pensioners
- Domestic workers with bed outside

3.2 Materials

A) Informant

The informant for the research work will be the head of the household or his / her spouse.

B) Equipment

The interview is direct, where the collection of information from the surveys is using mobile equipment for data capture (TABLET).

C) Statistical study

The data obtained are subjected to a statistical analysis using the Rstudio program, which will allow distributions for each variable, as well as their univariate and bivariate analysis.

As well as the different analyzes of statistical models such as quantile and linear regression.

Chapter 4

Analysis and Results

4.1. Database

Rstudio software was used for the construction process, where the information in question was taken as a source from the national household survey (ENAHO), for said construction the following bases were taken.

***Table 3**. ENAHO surveys for the construction of the database*

ENAHO - Original name	Definition	Name in the Script Rstudio
Enaho01-2019-100.sav	Characteristics of the home and the Home	VIVIENDA
Enaho01-2019-200.sav	Characteristics of the members of the home	Enaho01-2019-200
Enaho01A-2019-300.sav	Education	EDUC
Enaho01A-2019-400.sav	Health	SALUD
Enaho01a-2019-500.sav	Employment and Income	EMP_ING
Enaho01-2019-601.sav	Expenses on Food and Beverages	GAST_ALIM
Enaho01-2019-603.sav	Housing Maintenance	MANT_VIV
Enaho01-2019-604.sav	Transport and Communications	TRANS_COM
Enaho01-2019-605.sav	Housing Services	SERV_VIV
Enaho01-2019-606.sav	Recreation, Fun and Culture Services	ESP_DIV_CULT
Enaho01-2019-606D.sav	Care Goods and Services Personal	BN_SV_CUID

Enaho01-2019-607.sav	Dress and Footwear	VES_CAL
Enaho01-2019-609.sav	Transfer Expenses	GAST_TRNSF
Enaho01-2019-610.sav	Furniture and fixtures	MUEB_ENS
Enaho01-2019-611.sav	Other Goods and Services	BNS_SERV
Sumaria-2019.sav	Summary (Calculated variables)	Sumaria_2019

Source: Based on information from the ENAHO 2019
Elaboration: Own

Table 3 shows that each survey contains information on the expenses in each of the segments to be reviewed for the statistical model in question.

For the construction of family expenses, each family spending module was taken into account, which are segmented in table 4.

Table 4. *List of data sources for the construction of the model*

Expenses	Expenses concept
Basic Expenses 1	Food expenditure
	Spending on clothing and footwear
Basic Expenses 2	Spending on health care and medical services
	Spending on transport and communications
Household Expenses	Expenditure on fuel, electricity and maintenance of the home
	House maintenance expenditure
Luxury Expenses	Spending on recreation, fun, cultural and teaching services
	Spending on other goods and services

Source: Based on information from the ENAHO 2019
Elaboration: Own

From gathering the sources, the necessary data was constructed, which differs from the number of respondents for each dwelling indicated by ENAHO in its technical file, more detail below.

Table 5. *Difference of construction and technical sheet ENAHO*

N°	Department	Real base - Housing	Technical sheet - Housing	Differences
1	**Amazonas**	1,229	1,340	111
2	**Ancash**	1,421	1,456	35
3	**Apurímac**	959	994	35
4	**Arequipa**	1,560	1,696	136
5	**Ayacucho**	1,169	1,244	75
6	**Cajamarca**	1,442	1,562	120
7	**Callao**	1,009	DOES NOT EXIST	
8	**Cusco**	1,278	1,364	86
9	**Huancavelica**	1,032	1,088	56
10	**Huánuco**	1,277	1,332	55
11	**Ica**	1,563	1,618	55
12	**Junín**	1,571	1,620	49
13	**La libertad**	1,585	1,666	81
14	**Lambayeque**	1,428	1,442	14
15	**Lima**	4,514	6,248	1,734
16	**Loreto**	1,468	1,538	70
17	**Madre de dios**	642	696	54
18	**Moquegua**	971	1,104	133
19	**Pasco**	889	934	45
20	**Piura**	1,655	1,696	41
21	**Puno**	1,202	1,342	140
22	**San Martin**	1,335	1,394	59
23	**Tacna**	1,348	1,508	160
24	**Tumbes**	854	910	56
25	**Ucayali**	1,164	1,202	38
	Total	**34,565**	**36,994**	**2,429**

Source: Based on information from the ENAHO 2019
Elaboration: Own

4.2 Identification of variables:

Dependent variable or Response:

The response or dependent variable used in the quantile regression model is of the continuous quantitative type.

Y_1= Total Monthly Spending on Basics 1 (PEN) (Continuous quantitative variable)

Y_2= Total Monthly Spending on Basics 2 (PEN) (Continuous quantitative variable)

Y_3= Total Monthly Household Spending (PEN) (Continuous quantitative variable)

Y_4= Total Monthly Luxury Spending (PEN) (Continuous quantitative variable)

It is the total monthly expenditure per family, coming from the different expenses in health, home, luxury and others.

Independent or predictor variables

X_1= Stratum (Nominal qualitative variable)

X_2= Number of members per household (Discrete quantitative variable)

X_3= Type of occupation (Nominal qualitative variable)

X_4= Age (discrete quantitative variable)

X_5= Ubigeo (Nominal qualitative variable)

X_6= Income (continuous quantitative variable)

X_7= Housing situation (Nominal qualitative variable)

X_8= Marital Status (Nominal qualitative variable)

4.3 Descriptive analysis of the variables:

Because our data set does not comply with the OLS assumptions, that is, it requires a previous hypothesis about the randomness of the relationship expressed in terms that the errors (residuals) follow a normal distribution with zero mean and sigma-squared variance (Homoscedasticity), then quantile regression is very useful to visualize the changes in the conditional distribution of the data set. Lee, cited by John and Nduka (2009) established that the benefit of applying quantile regression is that, when using quantiles, they have the property of being robust in estimating outliers and therefore quantile regression inherits its robustness property. In short, a quantile regression model

was proposed to evaluate the effect of the factors that influence the distribution of the total monthly expenditure of the population of Peru.

The model to be estimated has the following general form:

$$Log(Y_i) = \beta_{0,t} + \beta_{1,t} * \text{Stratum} + \beta_{2,t} * \text{Number of members per household} + \beta_{3,t} * \text{Occupation type} + \beta_{4,t} * \text{Age} + \beta_{5,t} * Ubigeo + \beta_{6,t} * \text{Income} + \beta_{7,t} * \text{Situation of the house} + \beta_{8,t} * \text{Civil status}$$

Where Y_i the total monthly expense, X_i to i=1,2,3,4 are the covariates or independent variables and finally e represents the error.

To establish the correct transformation of the total monthly expenditure, the logarithm function has been used frequently, since it approximates a normal distribution, some studies have also shown that this transformation is the best class of Box-Cox transformations, however, for our analysis goes more for its interpretation in the analysis as percentage change, which is very convenient.

Next, more detail of the variables to be used for the models.

***Table 6**. Detail of the variable to be incorporated into the models*

DIMENSION	VARIABLES	CATEGORY	DESCRIPTION
Dependent variable			
	Basic 1 monthly household expenditure Basic 2 monthly household expenditure Monthly household expenditure of the household Monthly luxury household expenditure	Quantitative	Numerical
Independent variables			
ENTRY	Monthly income of the Head of the Household	Quantitative	Numerical
	Ubigeo (Department)	Qualitative	01=Amazonas, 02=Ancash, 03=Apurimac,

GEOGRAPHICAL			……..25=Ucayali
	Stratum	Quantitative	1= From 500 000 to more inhabitants 2=From 100 000 to more inhabitants …..
			8= Rural registration area
DEMOGRAPHICS	Age of the head of the household	Quantitative	Numerical
	Number of people in the household	Quantitative	Numerical
	Civil status	Qualitative	1= Cohabiting, 2= Married 3= Widower ….6= Single
WORKING MARKET	Occupation type	Qualitative nominal	1= Employer, 2= Independent worker,3= Employee, ……. 7= Other
HOME FEATURE	Housing situation	Qualitative nominal	1= Rented, 2= Own, fully paid, 3= Own, by invasion, 7= Another way

Source: Based on information from ENAHO 2019
Elaboration: Own

4.4 Univariate Analysis

The univariate analysis consists of carrying out a study to assess the consistency of the information and determine the adjustments and filters necessary to fine-tune the database, at the level of each of the variables.

Table 7. Predictor variables of the model

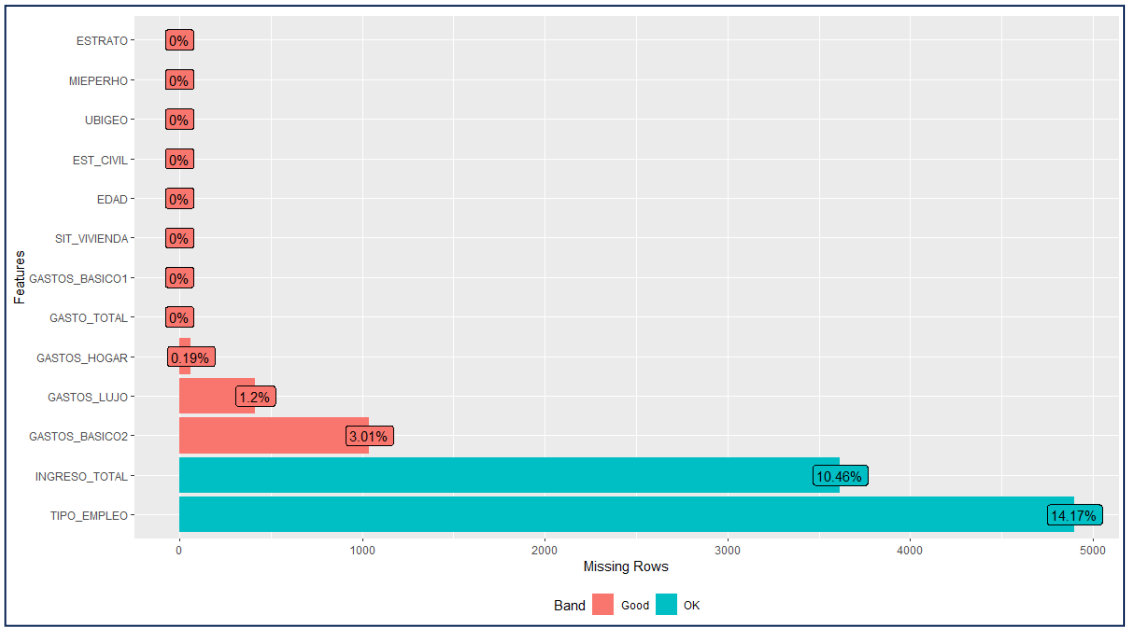
Variable	Variable concept
Stratum	Geographic stratum
Mieperho	Number of members per household
P507	Occupation type
Age	Age of the head of the household
Ubigeo	Department where it belongs
Income	Sum of fixed and variable income

P105A	Housing situation
P209	Civil status

Source: Based on information from the ENAHO 2019
Elaboration: Own

Now, each predictor variable of the mentioned models is found missing values (not reported values), as well as the predicted variables, in this regard it is shown below in the following table

Graph 1. *Missing for each predictor and predicted variable*



Source: Based on information from ENAHO 2019
Elaboration: Own

It can be seen that the percentage of missing in general is not so significant, except in the predictor variables "Total income" (10.49%) and "Type of employment" (14.17%) which exceed 10%, to complete the values It will be imputed by the median and mode in the quantitative and qualitative variables respectively. The median will be used as the imputation method because there are heterogeneous data for the same reason of the study and that families in Peru have different expenses to them, and extreme values are also located, therefore, the median will reflect more information. Thus, the mode will also be used as the imputation method for the variable “Type of employment”.

For the construction of the model, information from the ENAHO 2019 was used and under certain thresholds, below, more detail.

***Table 8**. Universe of the model*

Criteria for filtering	Falls	% Falls	Remnant base
			34,565
Income: >= S/60 y <= S/ 6,600	2,404	6.96%	32,161
Basic Expenses 1: > S/ 0 y <= S/2,490	320	0.99%	31,841
Basic Expenses 2: > S/ 0 y <= S/716	743	2.33%	31,098
Household expenses: > S/ 0 y <= S/1,238	1,193	3.84%	29,905
Luxury Expenses: > S/ 0 y <= S/3,972	239	0.80%	29,666
Number of members per household <=14	299	1.01%	29,367
Age >= 18	2	0.01%	29,365

Source: Based on information from the ENAHO 2019
Elaboration: Own

- For the criterion that income is greater than S / 60 and less than S / 6,600, for this purpose 3Q + 3 RIC (interquartile range) was used to eliminate outliers as the largest data, which is considered maximum outliers.

- For the criterion that Basic Expenses 1 are greater than S / 0 and less than S / 2,490, for this purpose, it was used for the elimination of outliers as greater data to 3Q + 3.RIC (interquartile range), which was consider maximum outliers.
- For the criterion that Basic Expenses 2 are greater than S / 0 and less than S / 716, for this purpose, it was used for the elimination of outliers as greater data at 3Q + 3.RIC (interquartile range), which was consider maximum outliers.
- For the criterion that Household Expenditures are greater than S / 0 and less than S / 1,238, for this, it was used for the elimination of outliers as greater data to 3Q + 3.RIC (interquartile range), which is considered maximum outliers.
- For the criterion that Luxury Expenses are greater than S / 0 and less than S / 3,972, for this purpose, the elimination of outliers was used as higher data at 3Q + 3.RIC (interquartile range), which is considered maximum outliers.
- For the criterion that the Number of members per household is less than 14, for this, it was used for the elimination of outliers as greater data at 3Q + 3.RIC (interquartile range), which is considered maximum outliers.
- For the criterion that the age is greater than 18, for this it was put by investigation since normally one could be in charge of a household at the age of 18 in Peru.

Next, it is shown that the average expenses of each module of basic expenses 1, 2, home and luxury are 553.8, 127.2, 250.5, 651.3 respectively. It can also be observed that the standard deviation is high, which shows that the variables have a lot of variability, and a not less detail is that the "Luxury expenditure" is one of the variables with greater variability and the one with a higher average even than some most relevant expenses for a family. Thus, it is also observed that for the dependent variables the mean is greater than the median, therefore, the distribution is asymmetric with a tail to the right (skewed to the right), and a positive asymmetry is observed for all mutes.

Table 9. *Statistical description of total monthly spending in Peru 2019*

Dependent variable	Mean	Standard deviation	Median	Asymmetry
Basic Expenses 1	553.8	401.4305	443.8	1.454866
Basic Expenses 2	127.2	129.5605	86.9	1.767227
Household Expenses	250.5	221.9287	181.5	1.78812
Luxury Expenses	651.3	715.5675	430.4	1.758291

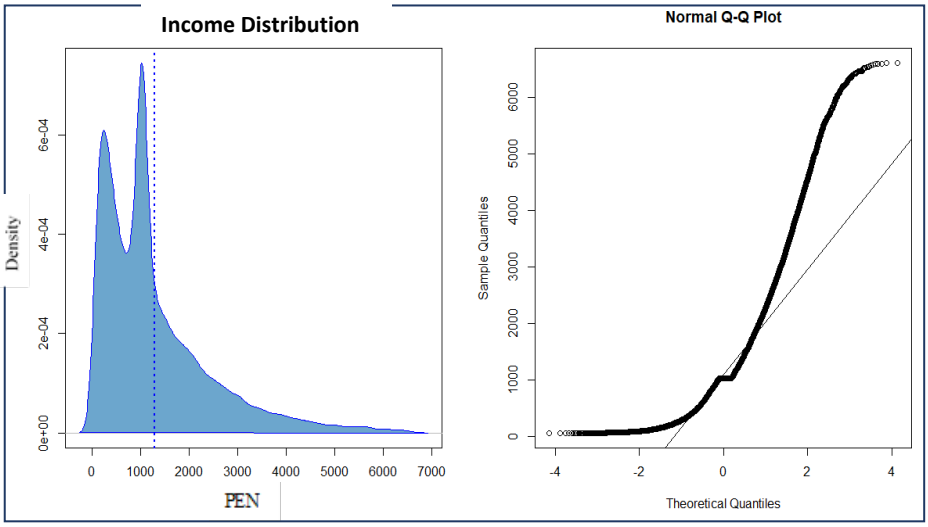
Source: Based on information from ENAHO 2019
Elaboration: Own

Next, the univariate analysis is shown for each of the predicted variables, which are the four expenditure modules and a predictor variable.

- Income
- Basic expenses 1
- Basic expenses 2
- Household expenses
- Luxury expenses

It can be seen in Graph 2 the distribution of total monthly income in Peru in 2019 is very variable in its distribution, in addition, the variable does not have a normal behavior.

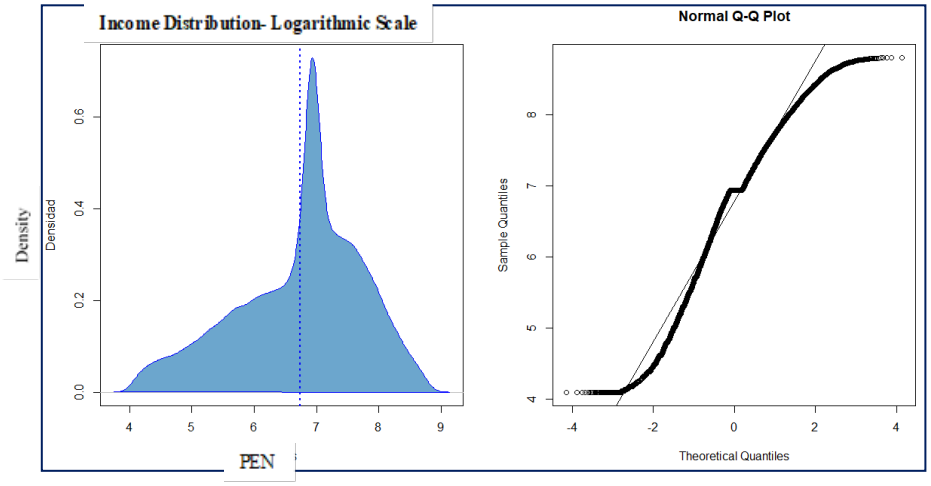
Graph 2. *Distribution of total monthly income in Peru 2019*



Source: Based on information from the ENAHO 2019
Elaboration: Own

Due to the fact that the total monthly income in Peru in 2019 is very variable in its distribution, in addition to being distributed in a bimodal way, for this reason the variable "Income" was transformed into the logarithm transformation where a both the distribution change, then more detail.

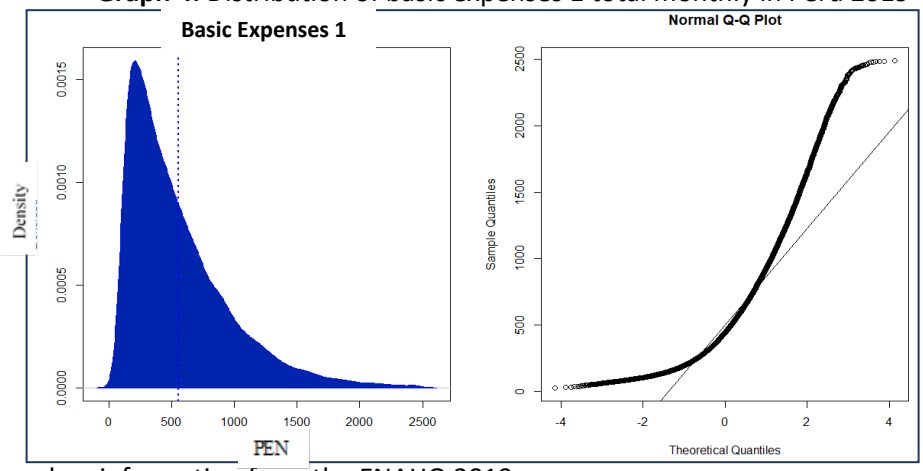
Graph 3. *Distribution of total monthly transformed income in Peru 2019*



Source: Based on information from the ENAHO 2019
Elaboration: Own

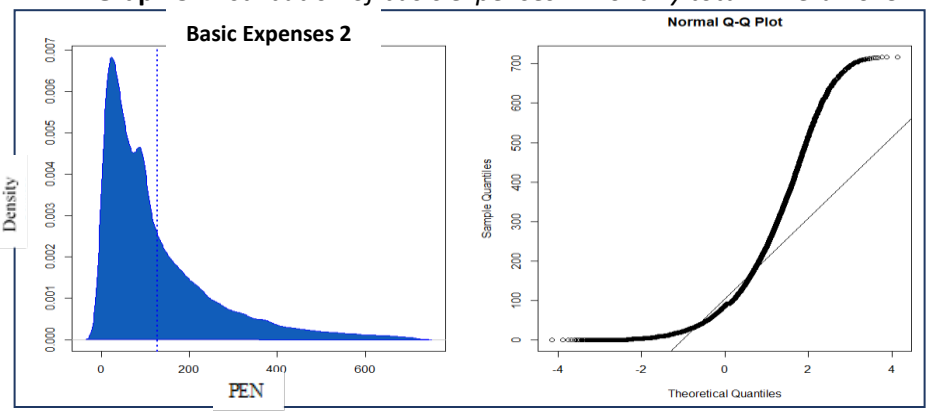
Next, we will proceed to review each of the distributions of the dependent variable, which are each expenditure module.

Graph 4. Distribution of basic expenses 1 total monthly in Peru 2019



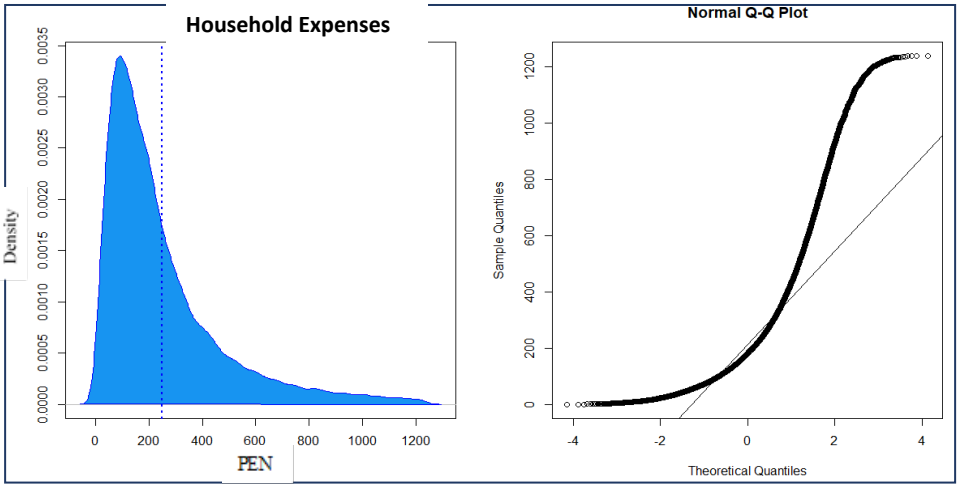
Source: Based on information from the ENAHO 2019
Elaboration: Own

Graph 5. *Distribution of basic expenses 2 monthly total in Peru 2019*



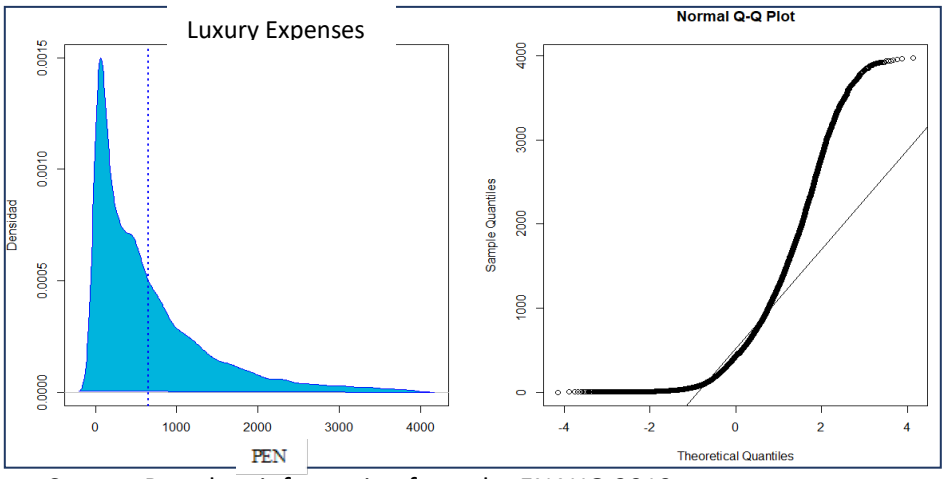
Source: Based on information from the ENAHO 2019
Elaboration: Own

Graph 6. Distribution of total monthly household expenses in Peru 2019



Source: Based on information from the ENAHO 2019
Elaboration: Own

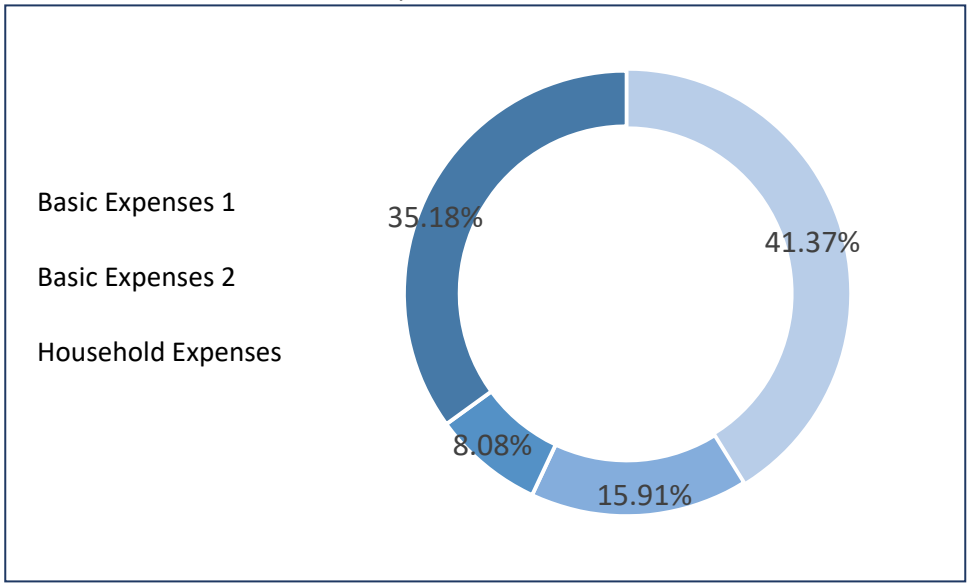
Graph 7. Distribution of monthly luxury expenses in Peru 2019



Source: Based on information from the ENAHO 2019
Elaboration: Own

From what is observed in the previous graphs with respect to the distribution, it can be concluded that they have an asymmetric behavior towards the right which tends to be more sustained than the income variable, however, it does not mean that the expenditure modules will not be transformed, in addition It can be noted that for the "Luxury Expenses" module there is a particular behavior. To do this, it will be analyzed how representative it is between the different expense modules.

Graph 8. *Distribution of expenditure modules with respect to total expenditure*



Source: Based on information from ENAHO 2019
Elaboration: Own

From what is observed in Graph 8, it is known that Luxury Expenses are made up of: recreation, entertainment, cultures and other goods and services. In addition, it contains one of the largest contributions of expenses, represented by 35.18% of total expenses.

4.5 Bivariate Analysis

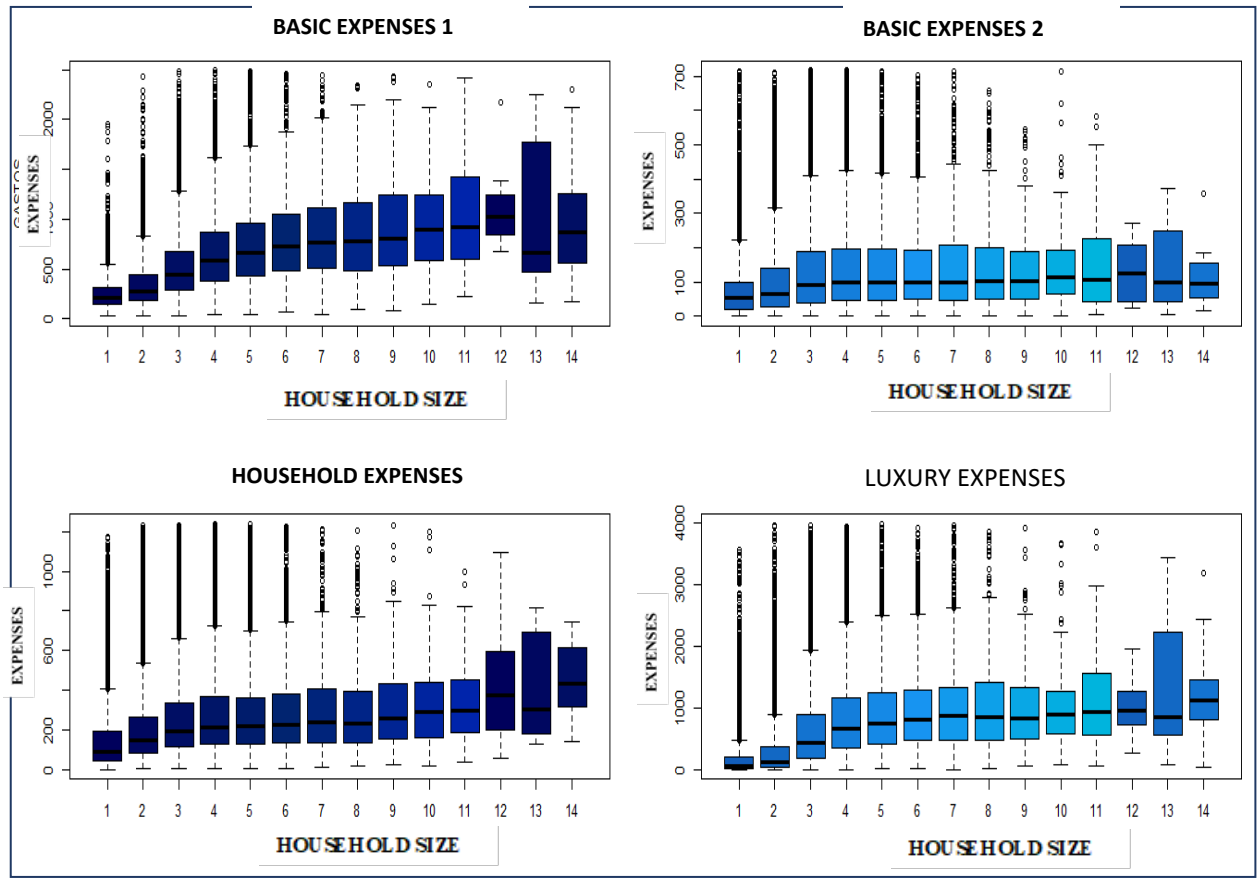
The bivariate analysis consists of evaluating the predictive capacity of each variable considered with respect to the variable to be predicted. In this regard, the methodology develops how each of the selected variables behave.

- Age
- Number of members per household
- Stratum
- Job type
- Ubigeo
- Civil status
- Income
- Housing situation

On this, it is shown by means of a graphic statistical analysis of each expenditure module that the predicted variables versus the predictor variables already mentioned are shown, in order to have a more reduced list of variables that are included in the modeling, which are develop below:

a) Number of members per household versus the different spending modules:

Graph 9. Expenditure modules versus Number of members per household



Source: Based on information from ENAHO 2019
Elaboration: Own

- The relationship of the variable Number of members per household and the Basic Expenses 1 module is not linear and the dispersion of the variable basic expenses 1 varies according to the number of members per household, therefore, it is necessary to carry out some transformation of the variable.
- The relationship of the variable Number of members per household and the Basic Expenses 2 module is not linear and the dispersion of the variable basic expenses 2 varies according to the number of members per household, therefore, it is necessary to carry out some transformation of the variable.
- The relationship of the variable Number of members per household and the Household Expenses module is not linear and the dispersion of the household expenditures variable varies according to the number of members per household, therefore, it is necessary to carry out some transformation of the variable.

- The relationship of the variable Number of members per household and the Luxury Expenses module is not linear and the dispersion of the luxury expenses variable varies according to the number of members per household, therefore, it is necessary to carry out some transformation of the variable.

It is considered that it is correct to transform the variable Number of members per household for its respective inclusion in the different models, and to include it as a qualitative variable to reflect the elasticity of the expenditure modules with respect to the size of the household, as can be seen the variable "More than 6 members" encompasses all the subsequent ones due to the fact that a similar behavior is observed.

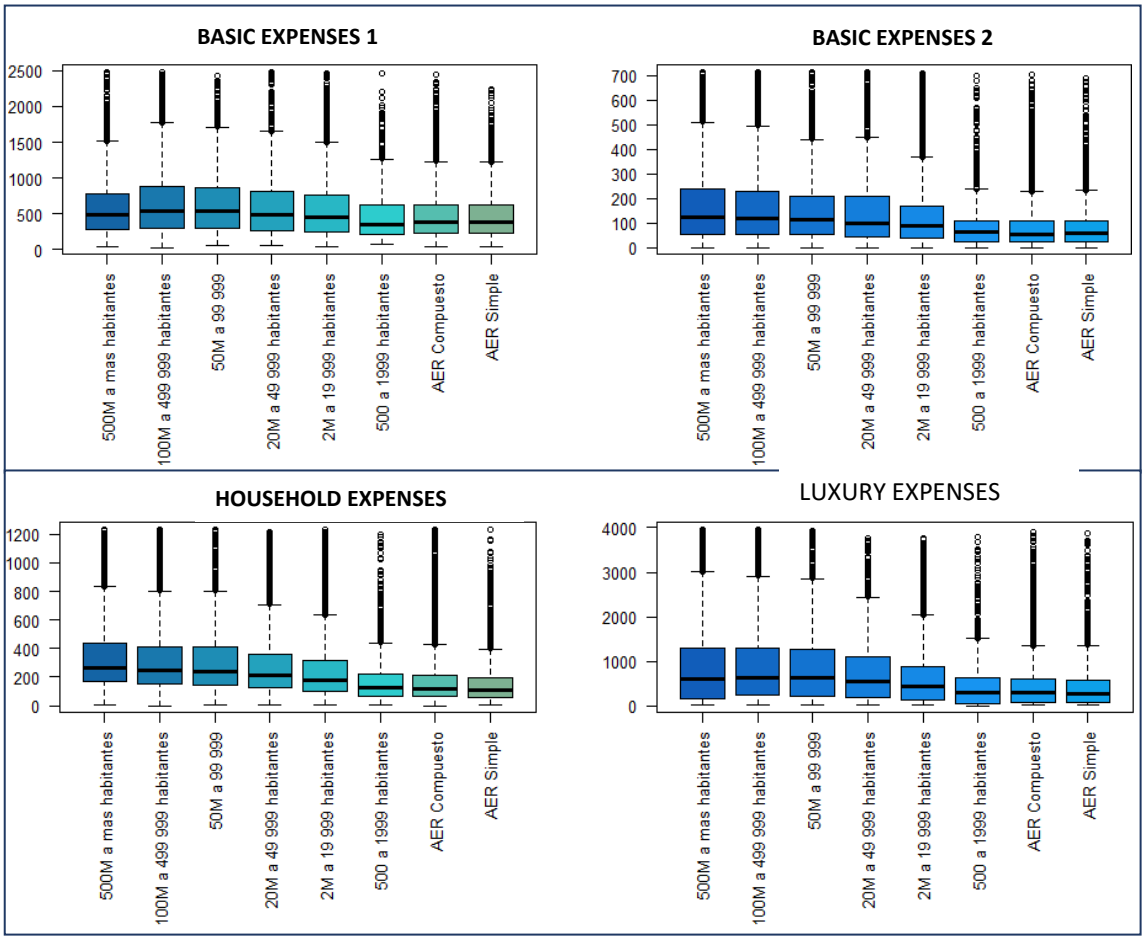
Table 10. Dummy variable for the variable Number of members per household

	Number of members per household				
	1	2	3	4	5
MIEPERHO1	1	0	0	0	0
MIEPERHO2	0	1	0	0	0
MIEPERHO3	0	0	1	0	0
MIEPERHO4	0	0	0	1	0
MIEPERHO5	0	0	0	0	1
MORE THAN 6 MEMBERS	0	0	0	0	0

Source: Based on information from ENAHO 2019
Elaboration: Own

b) Stratum versus the different expense modules:

Graph 10. Expenditure module versus Stratum



Source: Based on information from ENAHO 2019
Elaboration: Own

- The relationship between the Stratum variable and the Basic Expenses 1 module is not linear and the dispersion of the Basic Expenses 1 variable varies according to the different strata, therefore, it is necessary to carry out some transformation of the variable.
- The relationship of the Stratum variable and the Basic Expenses 2 module is not linear and the dispersion of the basic expenses 2 variable varies according to the different strata, therefore, it is necessary to carry out some transformation of the variable.
- The relationship of the Stratum variable and the Household Expenses module is not linear and the dispersion of the household expenses variable varies according to the different strata, therefore, it is necessary to carry out some transformation of the variable.

- The relationship of the Stratum variable and the Luxury Expenses module is not linear and the dispersion of the luxury expenses variable varies according to the different strata, therefore, it is necessary to carry out some transformation of the variable.

It is considered that it is correct to transform the variable Stratum for its respective inclusion in the different models, and include it as a qualitative variable to reflect the elasticity of the modules of expenditures with respect to the stratum.

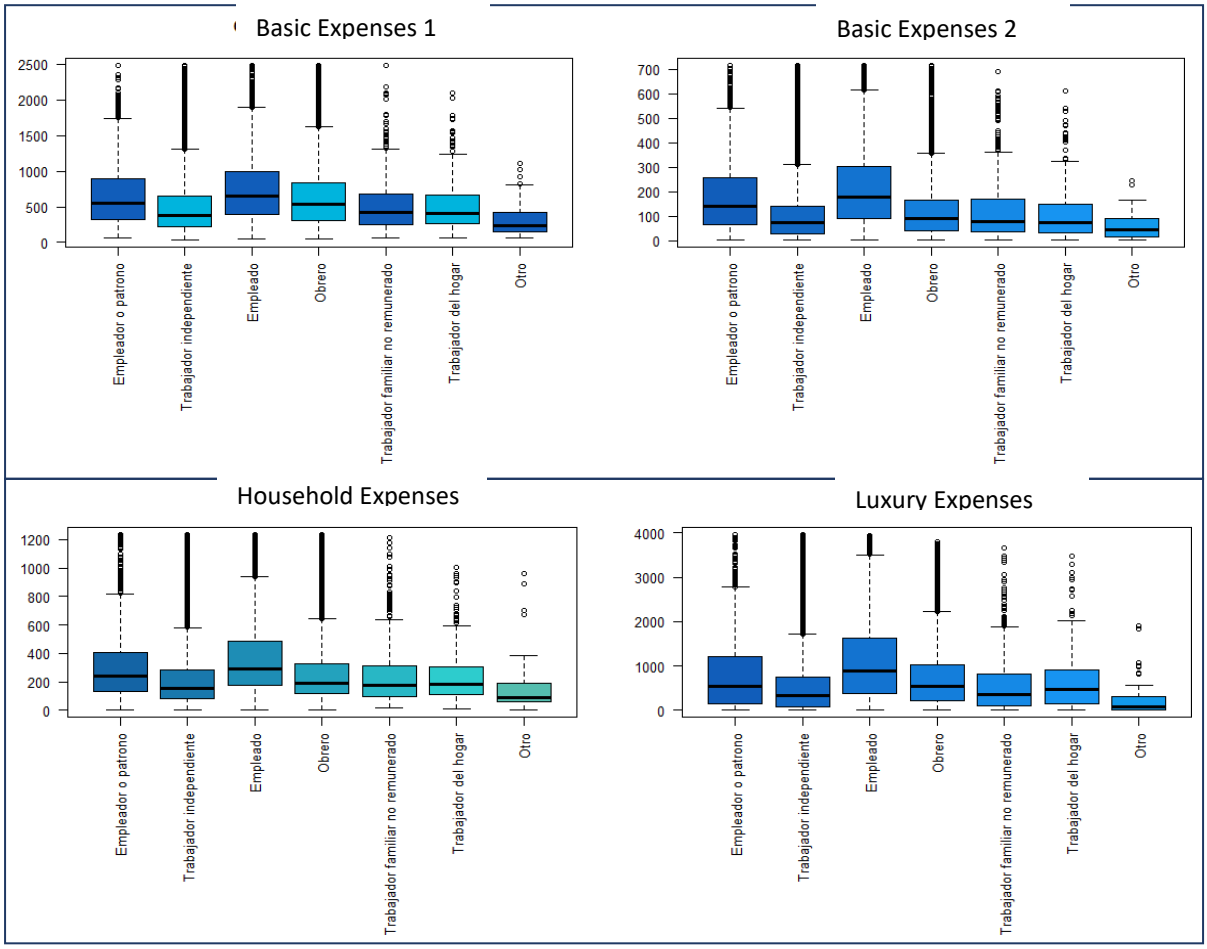
Table 11. *Dummy variable for the variable Stratum*

	Stratum				
	1	2	3	4	5
ESTRATO_1	1	0	0	0	0
ESTRATO_234	0	1	1	1	0
ESTRATO_5	0	0	0	0	1
OTHER STRATUM	0	0	0	0	0

Source: Based on information from ENAHO 2019
Elaboration: Own

c) Type of employment versus different expense modules:

Graph 11. Expenditure module versus Type of employment



Source: Based on information from ENAHO 2019
Elaboration: Own

- The relationship of the variable Type of employment and the Basic Expenses 1 module is not linear and the dispersion of the basic expenses 1 variable varies according to the type of employment, therefore, it is necessary to carry out some transformation of the variable.
- The relationship of the Type of job variable and the Basic Expenses 2 module is not linear and the dispersion of the basic expenses 2 variable varies according to the type of job, therefore, it is necessary to carry out some transformation of the variable.
- The relationship between the Type of employment variable and the Household Expenses module is not linear and the dispersion of the household expenses variable varies according to the type of employment, therefore, it is necessary to carry out some transformation of the variable.

- The relationship of the variable Type of employment and the Luxury Expenses module is not linear and the dispersion of the variable luxury expenses varies according to the type of employment, therefore, some transformation of the variable is necessary.

It is considered that it is correct to transform the variable Type of employment for its respective inclusion in the different models, and include it as a qualitative variable to reflect the elasticity of the modules of expenditures with respect to the Type of employment.

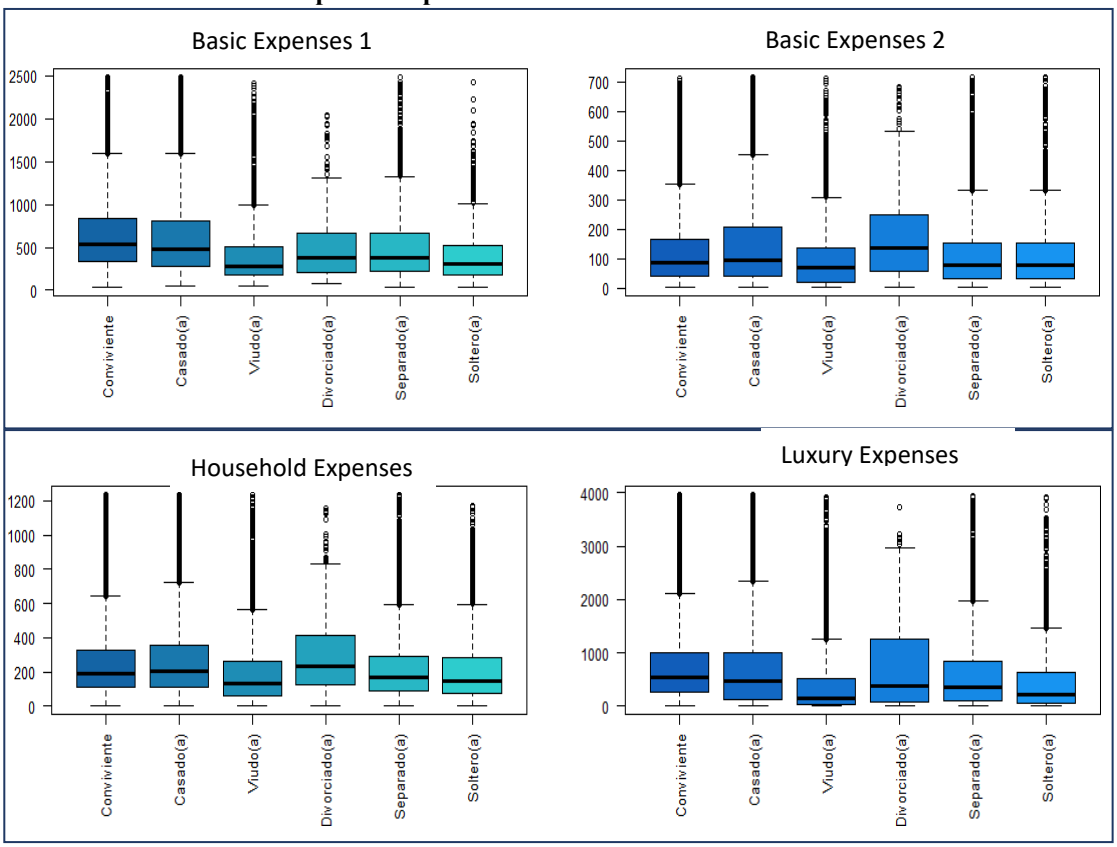
Table 12. *Dummy variable for the variable Type of employment*

	Job type					
	1	2	3	4	5	6
TIPO_EMPLEO_135	1	0	1	0	1	0
TIPO_ EMPLEO_246	0	1	0	1	0	1
TIPO_ EMPLEO_7	0	0	0	0	0	0

Source: Based on information from ENAHO 2019
Elaboration: Own

d) Marital status versus different expense modules:

Graph 12. Expenditure module versus Civil Status



Source: Based on information from ENAHO 2019
Elaboration: Own

- The relationship between the Marital Status variable and the Basic Expenses 1 module is not linear and the dispersion of the Basic Expenses 1 variable varies according to the Marital Status, therefore, it is necessary to carry out some transformation of the variable.
- The relationship between the Marital Status variable and the Basic Expenses 2 module is not linear and the dispersion of the Basic Expenses 2 variable varies according to the Marital Status, therefore, it is necessary to carry out some transformation of the variable.
- The relationship between the Marital Status variable and the Household Expenses module is not linear and the dispersion of the household expenses variable varies according to the Marital Status, therefore, it is necessary to carry out some transformation of the variable.
- The relationship of the Marital Status variable and the Luxury Expenses module is not linear and the dispersion of the luxury expenses variable varies according to the Marital Status, therefore, it is necessary to carry out some transformation of the variable.

It is considered that it is correct to transform the variable Marital status for its respective inclusion in the different models, and to include it as a qualitative variable to reflect the elasticity of the expenditure modules with respect to the Type of employment.

Table 13. *Dummy variable for the variable Marital status*

	Civil Status					
	1	2	3	4	5	6
EST_CIVIL_1	1	0	0	0	0	0
EST_CIVIL_2	0	1	0	0	0	1
EST_CIVIL_3	0	0	1	0	0	0
EST_CIVIL_4	1	0	0	1	0	0
EST_CIVIL_5	0	0	0	0	1	0
EST_CIVIL_6	0	0	0	0	0	1
EST_CIVIL_7	0	0	0	0	0	0

Source: Based on information from ENAHO 2019
Elaboration: Own

e) Age versus different expense modules:

The treatment of the age variable will be used the PieceWise method which will smooth the changes in the sections so that discontinuities do not occur. Likewise, this will allow us to capture in the model different expenditure trends within the same variable, in addition to including as many variables as sections are defined and the sum of the dummy variables must be equal to the original variable.

It is worth indicating that the way in which this criterion of ranges was taken for the variable, the statistical test of heterogeneity will be used, the Wilcoxon-Mann Whitney test.

The objective of this segmentation is to group the people of each household according to some of their attributes, so that the people of each household make up a segment which has characteristics that are homogeneous among them, but as heterogeneous as possible from the other segments with respect to spending. respective of each model.

The test corresponds to contrast the following hypothesis:

H_0: The sections follow the same distribution, they are homogeneous

H_A: The sections do not follow the same distribution, they are heterogeneous

Where the decision criterion will be:

$$\text{If } p < \alpha, \text{ is rejected } H_0$$

$$\text{If } p > \alpha, \text{ Not rejected } H_0$$

Where α is the level of significance, which will be given by the researcher, but according to the Risk Model Development Standard, a significance level α = 0.05 will be used.

Table 14. Piece Wise Treatment for Basic Expenses 1 model

RANK		P-VALUE	P < 0.05
[40-60}	>=60	0.000E+00	TRUE
<40	>=60	0.000E+00	TRUE
[40-60}	<40	5.4431E-15	TRUE

Source: Based on information from ENAHO 2019
Elaboration: Own

Table 15. Piece Wise treatment for the Basic Expenses model 2

RANK		P-VALUE	P < 0.05
[40-60}	>=60	2.246E-64	TRUE
<40	>=60	7.751E-03	TRUE
[40-60}	<40	3.558E-40	TRUE

Source: Based on information from ENAHO 2019
Elaboration: Own

Table 16. Piece Wise Treatment for the Household Expenses model

RANK		P-VALUE	P < 0.05
[40-60}	>=60	3.572E-80	TRUE
<40	>=60	4.537E-40	TRUE
[40-60}	<40	1.249E-03	TRUE

Source: Based on information from ENAHO 2019
Elaboration: Own

Table 17. Piece Wise treatment for the Luxury Expenses model

RANK		P-VALUE	P < 0.05
[40-60}	>60	4,36E-24	TRUE
<40	>60	1,47E-28	TRUE
[40-60}	>60	2,36E-02	TRUE

Source: Based on information from ENAHO 2019
Elaboration: Own

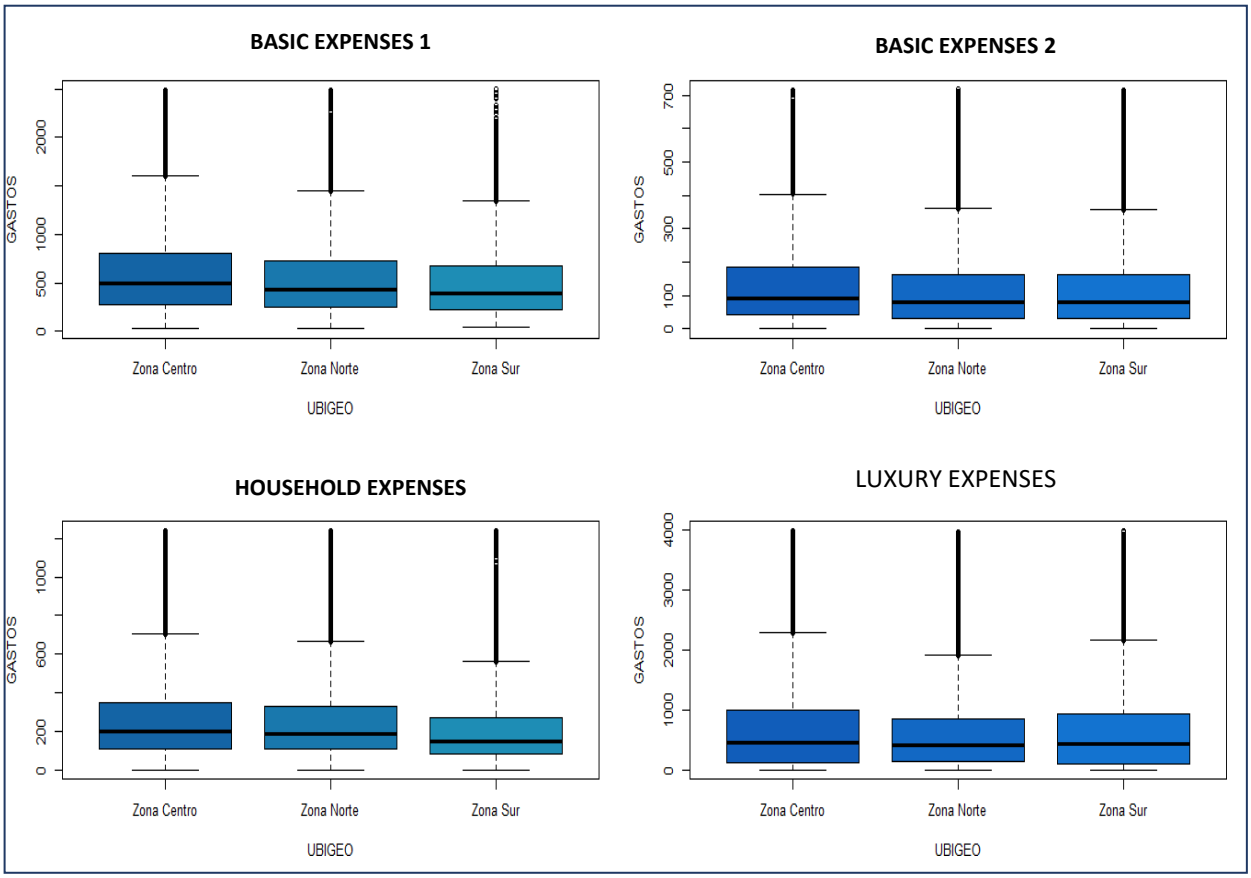
Therefore, it is considered that:

- From Basic Expenses 1: It is observed that for the different ranges it complies with heterogeneity, due to the level of significance it is less than 0.05 where H0 is rejected. Therefore, the ranges that were taken into consideration are valid.
- From Basic Expenses 2: It is observed that for the different ranges it complies with heterogeneity, due to the level of significance it is less than 0.05 where H0 is rejected. Therefore, the ranges that were taken into consideration are valid.
- From Household Expenses: It is observed that for the different ranges it meets heterogeneity, due to the level of significance it is less than 0.05 where H0 is rejected. Therefore, the ranges that were taken into consideration are valid.

- From Luxury Expenses: It is observed that for the different ranges it complies with heterogeneity, due to the level of significance it is less than 0.05 where H0 is rejected. Therefore, the ranges that were taken into consideration are valid.

f) Ubigeo versus different expense modules:

Graph 13. Expenditure module versus Ubigeo



Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 13, for each expenditure module, there is no significant difference between each category, but for convenience and study, the departments of Peru were grouped by geographical area, which are "Central Zone", "North Zone", "South Zone".

For this reason, when observing the previous graph, it was decided to include the Ubigeo variable through dummy variables which will be represented in the following table.

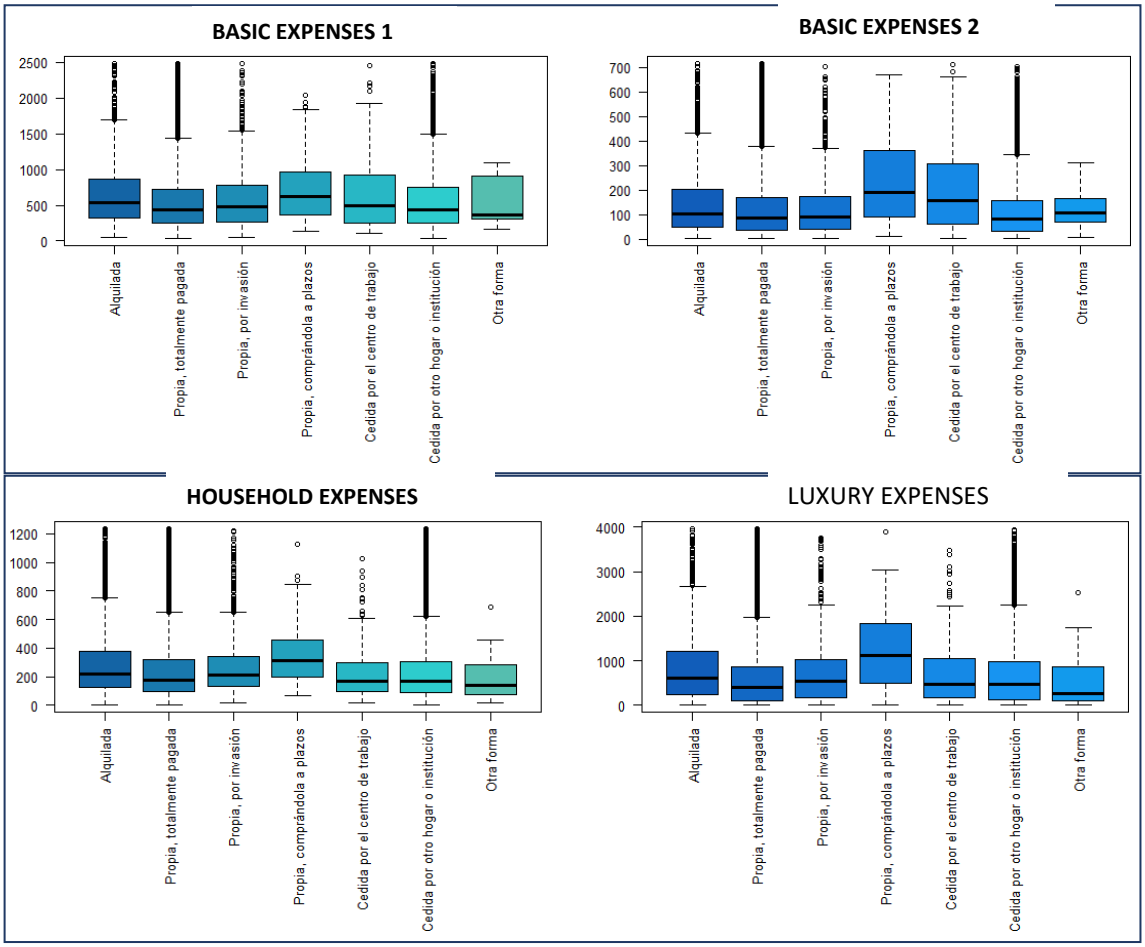
Table 18. Dummy variable for the Ubigeo variable

	UBIGEO	
	1	2
ZONA_CENTRO	1	0
ZONA_NORTE	0	1
ZONA_SUR	0	0

Source: Based on information from ENAHO 2019
Elaboration: Own

g) Housing situation versus different expense modules:

Graph 14. Expenditure module versus Housing situation



Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 14 for each expenditure module, no significant difference is seen between each category, but for convenience and study some categories of housing situation were grouped since they provide us with similar information regarding the dependent variable.

Therefore, when observing the previous graph, it was decided to unify the categories of "Own, fully paid" and "Own, by invasion" in a single dummy variable, in addition to unifying the categories "Assigned by the workplace", "Assigned by another home or institution" and "Another way" due to the fact that in all the expense modules it provides similar behavior, then more detail.

Table 19. *Dummy variable for the variable Housing situation*

	HOUSING SITUATION			
	1	2	3	4
SIT_VIVIENDA_1	1	0	0	0
SIT_VIVIENDA_23	0	1	1	0
SIT_VIVIENDA_4	0	0	0	1
SIT_VIVIENDA_567	0	0	0	0

Source: Based on information from ENAHO 2019
Elaboration: Own

4.6 Discriminant Capacity Analysis

This section shows the results of the predictive capacity tests of the models of each expense module, which will be developed below.

4.6.1 Quantile Regression:

Basic Expenses 1:

Table 14 shows the results of the different estimates proposed for the Basic Expenses 1 model. Among those that were taken into account to show the results are for the quantiles 0.10, 0.25, 0.40, 0.50, 0.60, 0.75, 0.90, and as an additional estimation the multiple linear regression. In almost all the estimates made, as well as for the multiple linear regression and the quantile regression, there are statistically significant parameters. Among those that have a positive influence are all of the variables except for "Estrato_1", "Mieperho_1", "Mieperho_2", "Mieperho_3", "Mieperho_4", "Mieperho_5"," Estado_civil_conviviente"," Estado_civil_casado", "Situacion_vivienda_4", "Ubigeo_Zona_norte"," Edad40", "Edad50" and "Edad60". Now, it can be observed that in the dummy variables "Mieperho" they negatively influence Basic Expenses 1, that is, Basic Expenses 1

is reduced approximately between 0.130 and 1.092 with respect to Basic Expenses 1 of the people who have a 6 to more members in your household, referring to multiple linear regression. It is observed that the relationship of Basic Expenses 1 and the variable "Total Income" tends to increase as we move into segments of the population of people with high monthly Basic Expenses, going from 15% (Q10) to 18.9% (Q75) However, for a more exhaustive analysis, the Q90 could also be reviewed since it increases by 3.4% with respect to Q10, however, this parameter value does not differ much from the different regressions and it is still significant. In the variables "Edad40", "Edad50" and "Edad60" it is observed that they have a negative influence with respect to Basic Expenses 1, however, it can be observed that in the entire range of quantiles and the estimation by the multiple linear regression there is no a significant difference in parameter values. An important fact to observe is that the variable "Estado_marivil_Masado" turns out to be a non-significant variable for the multiple linear regression, however, in the quantiles (Q25, Q40 and Q50) it turns out to be a significant variable. As a conclusion to the estimates of the quantile regression, we can obtain the different behaviors of the distribution with respect to Basic Expenses 1.

Table 20: Results of the quantile regression (Parameters and standard error) of Basic Expenses 1

	OLS	Quantile Regression						
		Quantile 10	Quantile 25	Quantile 40	Quantile 50	Quantile 60	Quantile 75	Quantile 90
LOG_INGRESO_TOTAL (Parameter)	0.177***	0.150***	0.171***	0.183***	0.185***	0.187***	0.189***	0.184***
Std error	0.004	0.007	0.005	0.005	0.004	0.004	0.005	0.006
ESTRATO_1 (Parameter)	-0.073***	-0.100***	-0.115***	-0.127***	-0.118***	-0.099***	-0.046***	0.005
Std error	0.011	0.021	0.015	0.015	0.014	0.014	0.016	0.015
ESTRATO_234 (Parameter)	0.109***	0.089***	0.087***	0.081***	0.090***	0.102***	0.123***	0.144***
Std error	0.008	0.016	0.012	0.011	0.01	0.01	0.011	0.012
ESTRATO_5 (Parameter)	0.054***	0.044**	0.040***	0.031**	0.035***	0.051***	0.055***	0.073***
Std error	0.01	0.017	0.014	0.013	0.012	0.012	0.013	0.015
MIEPERHO_1 (Parameter)	-1.092***	-1.067***	-1.139***	-1.133***	-1.140***	-1.147***	-1.112***	-1.084***
Std error	0.015	0.027	0.02	0.019	0.016	0.017	0.02	0.022
MIEPERHO_2 (Parameter)	-0.773***	-0.793***	-0.844***	-0.821***	-0.813***	-0.805***	-0.763***	-0.706***
Std error	0.012	0.023	0.017	0.016	0.014	0.015	0.016	0.02
MIEPERHO_3 (Parameter)	-0.472***	-0.524***	-0.533***	-0.488***	-0.486***	-0.472***	-0.440***	-0.420***
Std error	0.012	0.024	0.017	0.015	0.013	0.014	0.015	0.02

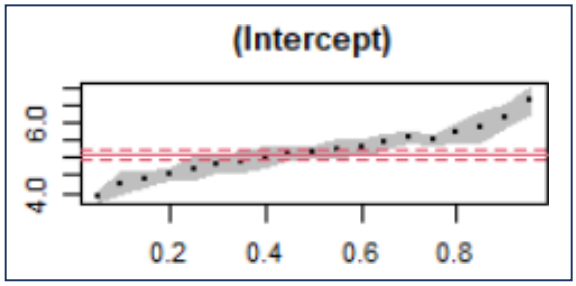
MIEPERHO_4 (Parameter)	-0.254***	-0.272***	-0.277***	-0.258***	-0.240***	-0.241***	-0.226***	-0.251***
Std error	0.012	0.023	0.016	0.015	0.013	0.013	0.014	0.017
MIEPERHO_5 (Parameter)	-0.130***	-0.134***	-0.155***	-0.130***	-0.129***	-0.126***	-0.105***	-0.124***
Std error	0.013	0.024	0.019	0.015	0.014	0.014	0.016	0.019
TIPO_EMPLEO_135 (Parameter)	0.383***	0.538***	0.440**	0.361**	0.388***	0.377***	0.395***	0.118
Std error	0.072	0.167	0.182	0.17	0.074	0.115	0.032	0.161
TIPO_EMPLEO_246 (Parameter)	0.224***	0.361**	0.279	0.198	0.226***	0.219*	0.247***	-0.011
Std error	0.072	0.167	0.182	0.17	0.073	0.115	0.031	0.161
EST_CIVIL_CONVIVIENTE (Parameter)	-0.051***	0.03	-0.021	-0.063***	-0.075***	-0.100***	-0.111***	-0.101***
Std error	0.015	0.031	0.023	0.019	0.019	0.019	0.023	0.02
EST_CIVIL_CASADO(A) (Parameter)	-0.007	0.074**	0.025	-0.025	-0.035*	-0.048**	-0.065***	-0.055***
Std error	0.015	0.029	0.023	0.019	0.019	0.019	0.023	0.02
EST_CIVIL_VIUDO(A) (Parameter)	0.070***	0.149***	0.092***	0.073***	0.052***	0.032	0.011	0.023
Std error	0.017	0.032	0.024	0.021	0.02	0.02	0.024	0.023
EST_CIVIL_DIVORCIADO(A) (Parameter)	0.125***	0.188***	0.132***	0.085	0.076**	0.085**	0.118***	0.116***
Std error	0.035	0.054	0.046	0.063	0.035	0.04	0.035	0.03
EST_CIVIL_SEPARADO(A) (Parameter)	0.046***	0.092***	0.071***	0.039**	0.040**	0.023	0.013	0.016
Std error	0.015	0.029	0.023	0.019	0.019	0.019	0.023	0.019
SIT_VIVIENDA_1 (Parameter)	0.052***	0.044	0.068***	0.057***	0.052***	0.063***	0.027	0.035*
Std error	0.014	0.029	0.021	0.019	0.017	0.018	0.018	0.02
SIT_VIVIENDA_23 (Parameter)	0.026***	0.026	0.030**	0.029**	0.033***	0.033***	0.020*	0.017
Std error	0.009	0.017	0.013	0.012	0.011	0.011	0.012	0.013
SIT_VIVIENDA_4 (Parameter)	0.092	0.09	0.146***	0.07	0.120***	0.036	0.067	0.157
Std error	0.074	0.194	0.029	0.222	0.041	0.027	0.114	0.25
UBIGEO_2_ZONA_CENTRO (Parameter)	0.149***	0.145***	0.156***	0.169***	0.173***	0.171***	0.142***	0.117***
Std error	0.008	0.016	0.012	0.011	0.01	0.01	0.011	0.013
UBIGEO_2_ZONA_NORTE (Parameter)	-0.012	-0.041**	-0.029**	-0.009	-0.001	0.009	0.003	0.007
Std error	0.008	0.016	0.012	0.011	0.01	0.01	0.011	0.014
EDAD40 (Parameter)	-0.0005	0.002	0.001	-0.0003	-0.002	-0.002*	-0.003*	-0.005***
Std error	0.001	0.002	0.002	0.001	0.001	0.001	0.001	0.002
EDAD50 (Parameter)	-0.009***	-0.007***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***
Std error	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

EDAD60 (Parameter)	-0.003***	0.001	-0.002**	-0.004***	-0.004***	-0.004***	-0.004***	-0.004***
Std error	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Constant (Parameter)	5.106***	4.267***	4.687***	5.016***	5.164***	5.317***	5.505***	6.163***
Std error	0.086	0.195	0.195	0.182	0.093	0.129	0.068	0.176
p<0.1; **p<0.05; ***p<0.01 F Statistic= 905.221 (Significant)**								

Source: Based on information from ENAHO 2019
Elaboration: Own

Additionally, the variables plus the intercept are presented below. For this reason, following Koenker and Hallock (2001) we will graph for each quantile regression coefficient for λ = [0.05, 0.10, 0.15…… .0.95] that in each of the figures are represented by the black dotted lines. For each explanatory variable, said estimators can be interpreted as the impact that a unit change of said variables has on the logarithm of Basic Expenses 1 per month, keeping the rest of the explanatory variables constant. In this way, each of the graphs has the quantile scale on the horizontal axis and the monthly Basic Expenses 1 logarithm scale on its vertical axis, which indicates the effect of the explanatory variable on the predicted variable. The contours of the shaded point cloud correspond to the lower and upper values of the confidence bands of the corresponding quantile regression estimator. The horizontal red line corresponds to the value of the estimator of the conditional mean estimated by means of Ordinary Least Squares (OLS). Finally, the horizontal red lines with small stripes correspond to the lower and upper limits of the confidence interval of said OLS estimator. The level of significance used for the confidence intervals is 95%.

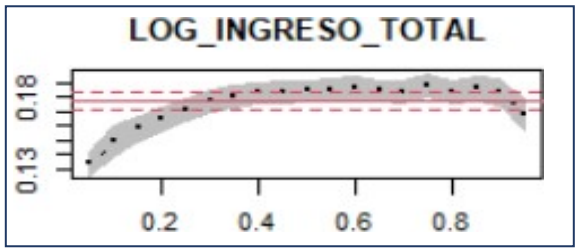
Graph 15. Regression coefficients by quantiles of the intercept



Source: Based on information from ENAHO 2019
Elaboration: Own

Graph 15 shows the evolution of the intercept for the different levels of Basic Expenses 1 of the population of Peru in 2019.

Graph 16. Regression coefficients by quantiles of the Logarithm of total

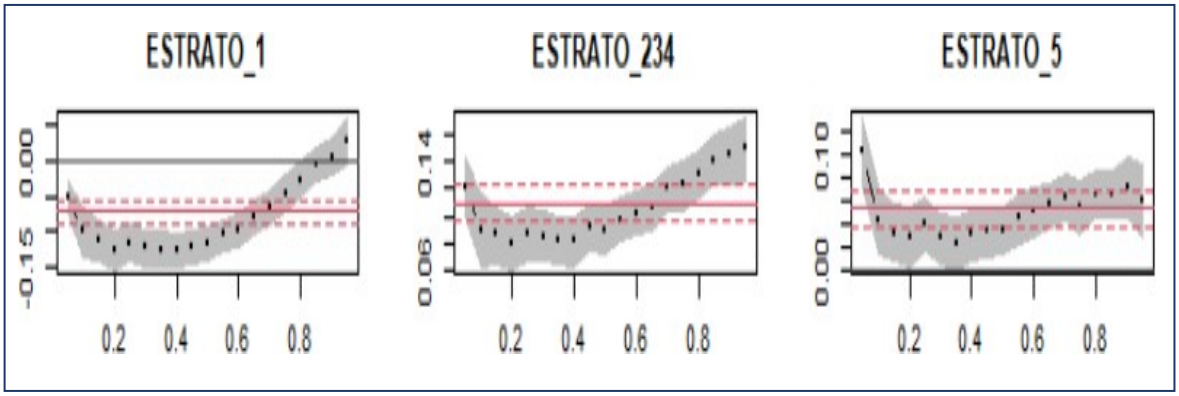


income

Source: Based on information from ENAHO 2019
Elaboration: Own

In Graph 16 it is shown that the logarithm variable of total family income, its effect positively influences Basic Expenses 1 of the population of Peru in the year 2019, in addition, the evolution of the quantile 0.35 is shown, and then the trend tends to stabilize. variable logarithm of total income.

Graph 17. Regression coefficients by quantiles of the Stratum



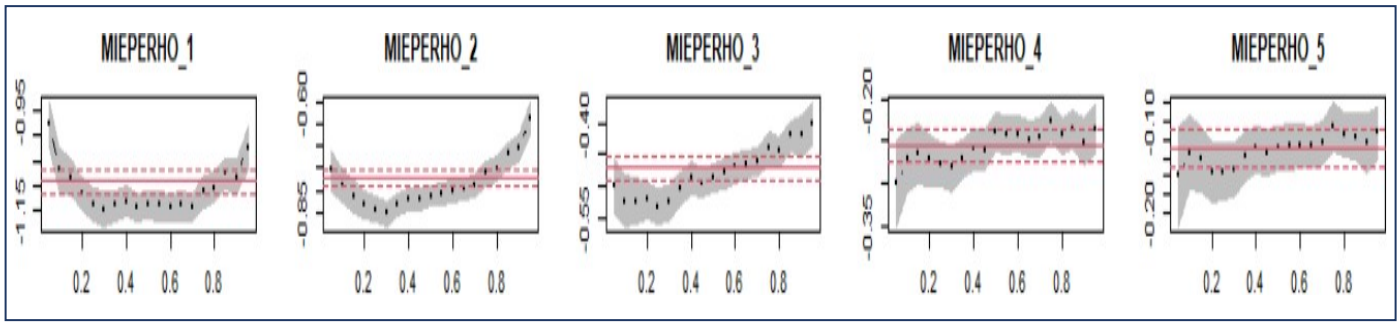
Source: Based on information from ENAHO 2019
Elaboration: Own

Graph 17 shows that:

- The variable "Estrato_1" (People who live in the stratum of 500 thousand or more inhabitants) its effect negatively influences up to the quantile 0.5 and then changes its positive effect in the following quantiles onwards with respect to the Basic Expenses 1 of the population of Peru in 2019.
- The variable "Estrato_234" (People who live in the stratum where between 20 thousand and 500 thousand inhabitants live) its effect positively influences all quantiles, thus it is also observed that up to the 0.75 quantile they spend less money on Basic Expenses 1 than OLS, then they tend to influence more ahead with respect to the Basic Expenses 1 of the population of Peru in the year 2019.

- The variable “Estrato_5” (People who live in the stratum where between 2,000 and 20,000 inhabitants live), its effect positively influences all quantiles, however, there is not much difference between the estimated effects of OLS and quantile regression because the interval the confidence interval shown by the red dotted lines for the OLS is superimposed with the confidence interval of the quantile regression with respect to the Basic Expenditures 1 of the population of Peru in the year 2019.

Graph 18. *Mierpeho quantile regression coefficients*

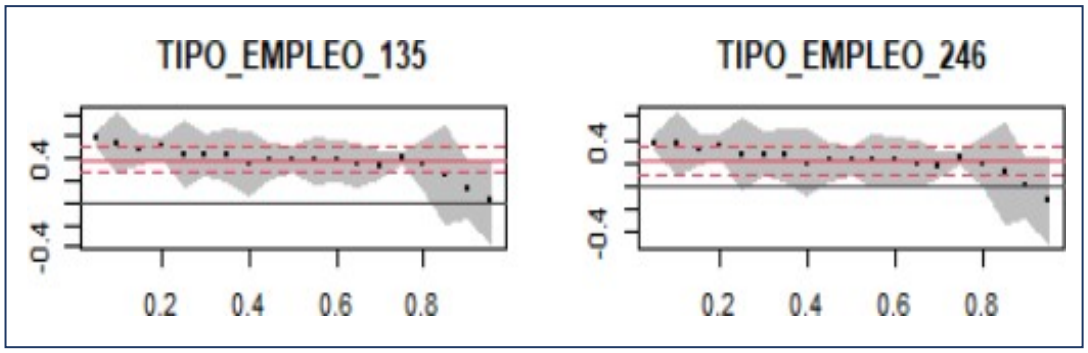


Source: Based on information from ENAHO 2019
Elaboration: Own

Graph 18 shows that:

- The variable "Mierperho_1" (1 household members in their home) its effect negatively influences all quantiles with respect to Basic Expenses 1 of the population of Peru in 2019.
- The variable "Mierperho_2" (2 household members in their home), its effect negatively influences all quantiles with respect to Basic Expenses 1 of the population of Peru in 2019.
- The variable "Mierperho_3" (3 household members in their home), its effect negatively influences all quantiles with respect to Basic Expenses 1 of the population of Peru in the year 2019.
- The variable “Mierperho_4” (4 household members in their home) and “Mierperho_5” (5 household members in their home), its effect negatively influences all quantiles, however, there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression with respect to Basic Expenditures 1 of the population of Peru in the year 2019.

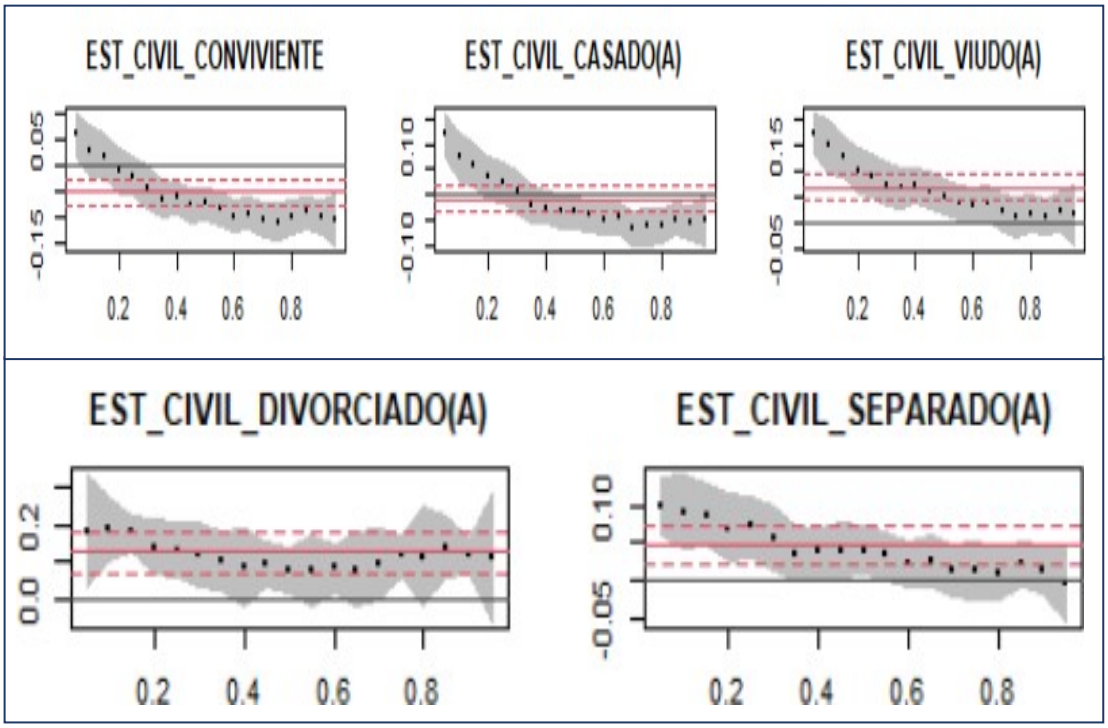
Graph 19. *Regression coefficients by quantiles of the Type of employment*



Source: Based on information from ENAHO 2019
Elaboration: Own

The variable "Tipo_Empleo_135" (People whose characteristics are "Employer or employer", "Employee" and "Unpaid family worker") and "Tipo_Empleo_246" (People whose characteristics are "Independent worker", "Worker" and " Domestic worker ") its effect positively influences all quantiles, however, there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the interval confidence of the quantile regression with respect to Basic Expenses 1 of the population of Peru in the year 2019.

Graph 20. Regression coefficients by quantiles of the Civil Status

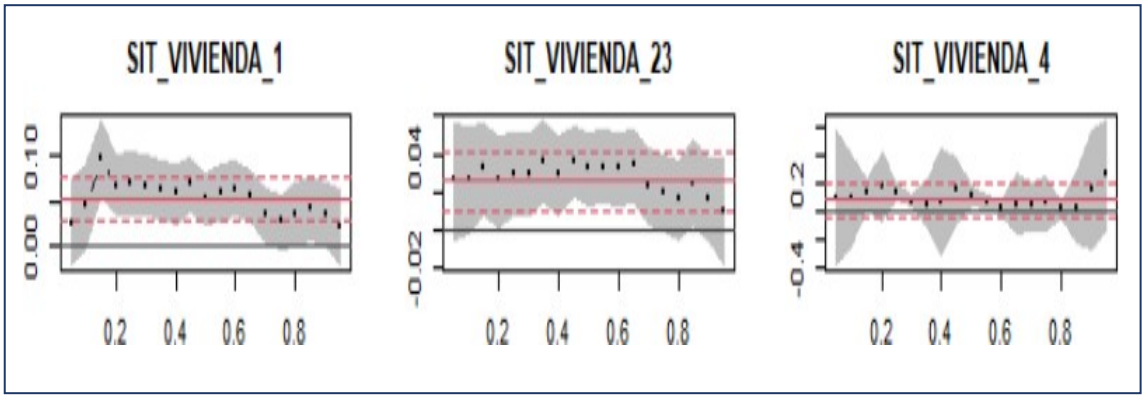


Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 20 it is observed that:

- The variable "Estado_Civil_Conviviente" its effect negatively influences all quantiles with respect to Basic Expenses 1 of the population of Peru in the year 2019.
- The variable "Estado_Civil_Casado" its effect negatively influences all quantiles with respect to Basic Expenses 1 of the population of Peru in 2019.
- The variable "Estado_Civil_Viudo" its effect positively influences all quantiles with respect to Basic Expenses 1 of the population of Peru in 2019.
- The variable "Estado_Civil_Divorciado" and "Estado_Civil_Separado" its effect negatively influences all quantiles, however, there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS is overlaps with the confidence interval of the quantile regression with respect to Basic Expenses 1 of the population of Peru in the year 2019.

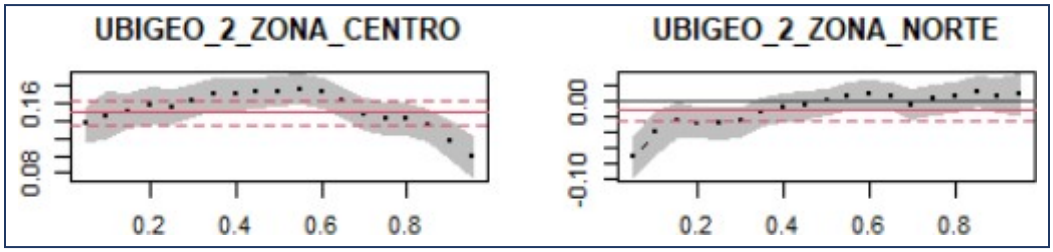
Gráfico 21. Coeficientes de regresiones por cuantiles de la Situación de la vivienda



Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 21 it can be seen that for the dummy variables “Sit_Vivienda_1” (The home you occupy is “Rented”), “Sit_Vivienda_23” (The home you occupy are “Own, fully paid” and “Own, by invasion”) and “Sit_Vivienda_4 "(The home you occupy is" Own, buying it in installments), there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression with respect to Basic Expenses 1 of the population of Peru in the year 2019.

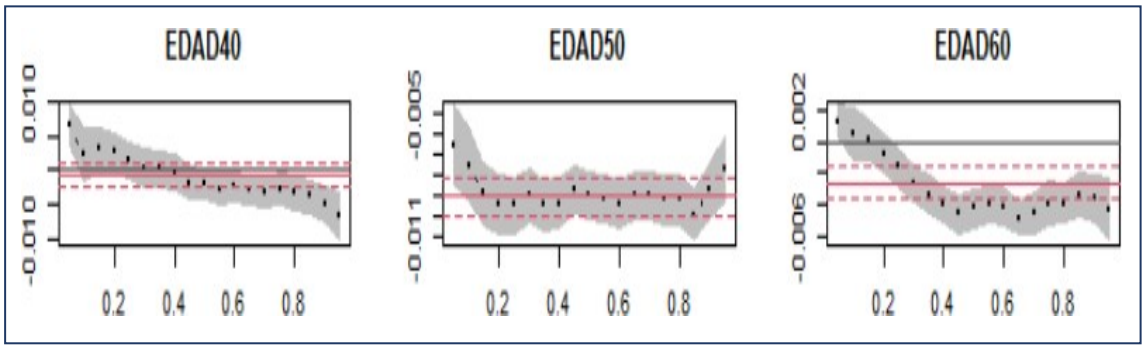
Graph 22. *Regression coefficients by quantiles of Ubigeo*



Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 22 it can be seen that for the dummy variables “Ubigeo_2_Zona_Centro” (People who live in the Central Zone) and “Ubigeo_2_Zona_Norte” (People who live in the North Zone), there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS is superimposed with the confidence interval of the quantile regression, except in the distribution tails, that is, starting with the quantile and ending with the 0.1 and 0.9 quantile respectively with respect to Expenses. Basic 1 of the population of Peru in the year 2019.

Graph 23. *Regression coefficients by quantiles of Age*



Source: Based on information from ENAHO 2019
Elaboration: Own

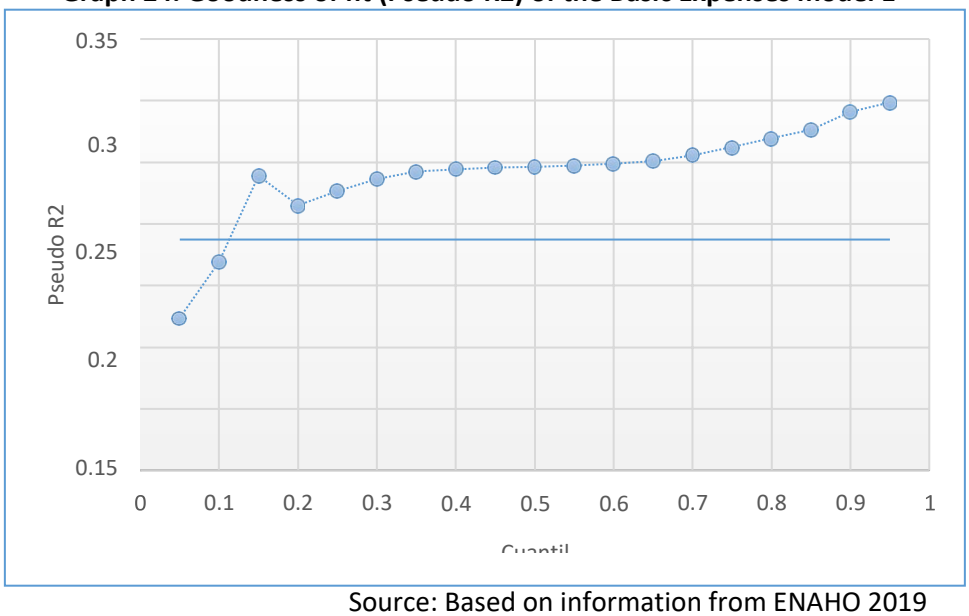
From Graph 23 it can be seen that:

- In the variable “Edad_40” (Age of the head of the household under 40 years) it is observed that it has a negative influence with respect to Basic Expenses 1 of the population of Peru in 2019, in addition that as the quantiles increase, it is observed that When the head of the household is less than 40 years old, he spends less on Basic Expenses 1.
- In the variable "Edad_50" (Age of the head of the household between 40 and 59 years) it is observed that it has a negative influence with respect to Basic Expenses 1 of the population of Peru in 2019, in addition there is not much difference between the estimated effects OLS and quantile regression because the confidence interval shown by the red dotted lines for the

OLS overlaps with the confidence interval of the quantile regression, except in the tails of the distribution, that is, starting with the quantile and ending with the quantile 0.1 and 0.9 respectively

- In the variable "Edad_60" (Age of the head of the household greater than 60 years) it is observed that it has a negative influence with respect to Basic Expenses 1 of the population of Peru in 2019, in addition to that as the quantiles increase, it is observed that when the head of the household is over 60 years of age, Basic Expenses 1 stabilize.

Graph 24. Goodness of fit (Pseudo R2) of the Basic Expenses model 1



Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 24 it can be observed that as the quantiles on the response variable Basic Expenses 1 increase, we notice that the goodness of fit (Pseudo R2) tends to increase, it is more it can be observed that in the quantile 0.95 is the highest Pseudo R2 with a value of 29.84%, due to this it is evidenced that the quantile regression is more flexible to explain the extremes of the response variable.

Basic Expenses 2:

Table 15 shows the results of the different estimates proposed for the Basic Expenses model 2. Among those that were taken into account to show the results are for the quantiles 0.10, 0.25, 0.40, 0.50, 0.60, 0.75, 0.90, and as an additional estimation the multiple linear regression. In almost all the estimates made, as well as for the multiple linear regression and the quantile regression, there are statistically significant parameters. Among those that have a positive influence are all variables except "Mieperho_1", "Mieperho_2", "Mieperho_3", "Mieperho_4", "Mieperho_5", "Estado_civil_conviviente", "Estado_civil_viudo", "Estado_civil_separado" and "Edad60". Now, it can be observed that in the dummy variables "Mieperho_1", "Mieperho_2", "Mieperho_3" negatively influence Basic Expenses 2, that is, Basic Expenses 2 are reduced approximately between 0.01 and 0.436 with respect to Basic Expenses 2 of people with 6 or more household members in their household, in reference to multiple linear regression. It is observed that the relationship of Basic Expenses 2 and the variable "Total Income" tends to decrease as we are situated in segments of the population of people with high monthly Basic Expenses 2, going from 53.6% (Q10) to 33.4% (Q90). In the variables "Edad40" and "Edad50" it is observed that they have a positive influence with respect to Basic Expenses 2, however, it can be observed that in the entire range of quantiles and the estimation by the multiple linear regression there is no significant difference in the parameter values. An important fact to observe is that the dummy variables "Mieperho_3", "Mieperho_4", "Mieperho_5" are non-significant variables both for the multiple linear regression and in the quantile regression in all its trajectory (Q10 to Q90), the same case is You can see for the variable “Estado_Civil_Casado”, however, in Q10 it is observed that the parameter is significant, which results in an analysis in that population. As a conclusion to the estimates of the quantile regression, we can obtain the different behaviors of the distribution with respect to Basic Expenses 2.

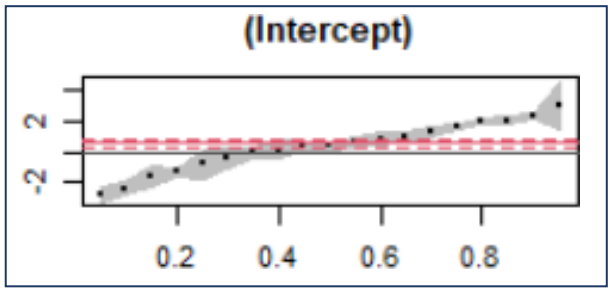
***Table 21**. Results of the quantile regression (Parameters and standard error) of Basic Expenses 2*

	OLS	Quantile Regression						
		Quantile 10	Quantile 25	Quantile 40	Quantile 50	Quantile 60	Quantile 75	Quantile 90
LOG_INGRESO_TOTAL (Parameter)	0.415***	0.536***	0.498***	0.471***	0.465***	0.445***	0.378***	0.334***
Std error	0.007	0.016	0.011	0.009	0.008	0.008	0.008	0.008
ESTRATO_1 (Parameter)	0.279***	0.150***	0.221***	0.258***	0.259***	0.239***	0.292***	0.248***
Std error	0.021	0.047	0.031	0.026	0.023	0.021	0.02	0.023
ESTRATO_234 (Parameter)	0.310***	0.344***	0.281***	0.288***	0.254***	0.239***	0.287***	0.239***
Std error	0.016	0.036	0.023	0.019	0.017	0.017	0.016	0.019
ESTRATO_5 (Parameter)	0.183***	0.174***	0.163***	0.186***	0.160***	0.131***	0.166***	0.132***
Std error	0.019	0.048	0.027	0.022	0.02	0.02	0.019	0.023
MIEPERHO_1 (Parameter)	-0.436***	-0.742***	-0.538***	-0.416***	-0.286***	-0.248***	-0.296***	-0.193***
Std error	0.028	0.072	0.041	0.037	0.034	0.031	0.027	0.032
MIEPERHO_2 (Parameter)	-0.194***	-0.325***	-0.264***	-0.204***	-0.154***	-0.134***	-0.127***	-0.068**
Std error	0.024	0.053	0.03	0.026	0.025	0.026	0.023	0.028
MIEPERHO_3 (Parameter)	-0.01	-0.042	-0.028	-0.008	0.007	0.018	-0.006	0.037
Std error	0.023	0.049	0.029	0.023	0.022	0.024	0.02	0.024
MIEPERHO_4 (Parameter)	0.017	0.05	0.025	0.040*	0.039*	0.039*	-0.008	0.012
Std error	0.022	0.048	0.026	0.023	0.021	0.022	0.021	0.023
MIEPERHO_5 (Parameter)	0.026	0.076	0.037	0.03	0.023	0.024	-0.009	0.024
Std error	0.024	0.05	0.031	0.023	0.022	0.024	0.022	0.026
TIPO_EMPLEO_135 (Parameter)	0.725***	1.479***	0.955*	0.617*	0.659***	0.602***	0.609***	0.773***
Std error	0.138	0.208	0.526	0.319	0.127	0.233	0.031	0.033
TIPO_EMPLEO_246 (Parameter)	0.384***	1.025***	0.558	0.254	0.323**	0.306	0.338***	0.572***
Std error	0.138	0.206	0.526	0.319	0.126	0.233	0.028	0.03
EST_CIVIL_CONVIVIENTE (Parameter)	-0.045	0.077	-0.063	-0.068*	-0.089***	-0.079**	-0.053*	-0.088**
Std error	0.029	0.072	0.05	0.039	0.03	0.032	0.029	0.035
EST_CIVIL_CASADO(A) (Parameter)	0.050*	0.207***	0.051	0.034	0.009	0.023	0.034	-0.002
Std error	0.029	0.071	0.05	0.039	0.03	0.031	0.029	0.034
EST_CIVIL_VIUDO(A) (Parameter)	-0.062*	-0.013	-0.027	-0.085*	-0.108***	-0.057	-0.031	-0.095**
Std error	0.032	0.087	0.057	0.044	0.038	0.036	0.032	0.038
EST_CIVIL_DIVORCIADO(A) (Parameter)	0.290***	0.504***	0.255***	0.195***	0.137*	0.200***	0.261***	0.243***

Std error	0.067	0.134	0.051	0.056	0.079	0.071	0.053	0.09
EST_CIVIL_SEPARADO(A) (Parameter)	-0.047	0.069	-0.099*	-0.102**	-0.112***	-0.085***	-0.054*	-0.062*
Std error	0.029	0.071	0.051	0.04	0.032	0.032	0.029	0.034
SIT_VIVIENDA_1 (Parameter)	0.046*	0.104**	0.052	0.026	0.004	0.002	0.017	0.002
Std error	0.027	0.053	0.034	0.03	0.03	0.028	0.027	0.027
SIT_VIVIENDA_23 (Parameter)	0.100***	0.160***	0.140***	0.114***	0.100***	0.074***	0.056***	0.039**
Std error	0.018	0.041	0.027	0.022	0.021	0.018	0.017	0.019
SIT_VIVIENDA_4 (Parameter)	0.261*	0.294	0.204	-0.015	-0.037	0.118	0.184***	0.473***
Std error	0.141	0.26	0.158	0.03	0.159	0.204	0.069	0.103
UBIGEO_2_ZONA_CENTRO (Parameter)	0.141***	0.263***	0.197***	0.148***	0.114***	0.099***	0.071***	0.003
Std error	0.016	0.035	0.022	0.018	0.018	0.016	0.016	0.019
UBIGEO_2_ZONA_NORTE (Parameter)	0.043***	0.035	0.064***	0.066***	0.052***	0.056***	0.043***	-0.01
Std error	0.016	0.039	0.024	0.02	0.017	0.017	0.016	0.019
EDAD40 (Parameter)	0.006***	0.009**	0.009***	0.006***	0.007***	0.006***	0.005**	0.006***
Std error	0.002	0.004	0.003	0.002	0.002	0.002	0.002	0.002
EDAD50 (Parameter)	0.005***	0.005*	0.008***	0.007***	0.006***	0.006***	0.006***	0.004***
Std error	0.001	0.002	0.001	0.001	0.001	0.001	0.001	0.001
EDAD60 (Parameter)	-0.015***	-0.033***	-0.022***	-0.011***	-0.006***	-0.006***	-0.007***	-0.003**
Std error	0.001	0.003	0.002	0.002	0.001	0.001	0.001	0.001
Constant (Parameter)	0.594***	-2.443***	-0.85	0.177	0.428***	0.855***	1.680***	2.271***
Std error	0.164	0.285	0.545	0.336	0.156	0.252	0.098	0.11
p<0.1; **p<0.05; ***p<0.01 F Statistic= 416.898 (Significant)**								

Source: Based on information from ENAHO
Preparation: Own

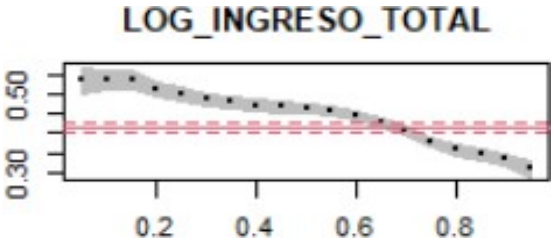
Graph 25. *Regression coefficients by quantiles of the intercept*



Source: Based on information from ENAHO 2019
Elaboration: Own

Graph 25 shows the evolution of the intercept for the different levels of Basic Expenses 2 of the population of Peru in 2019.

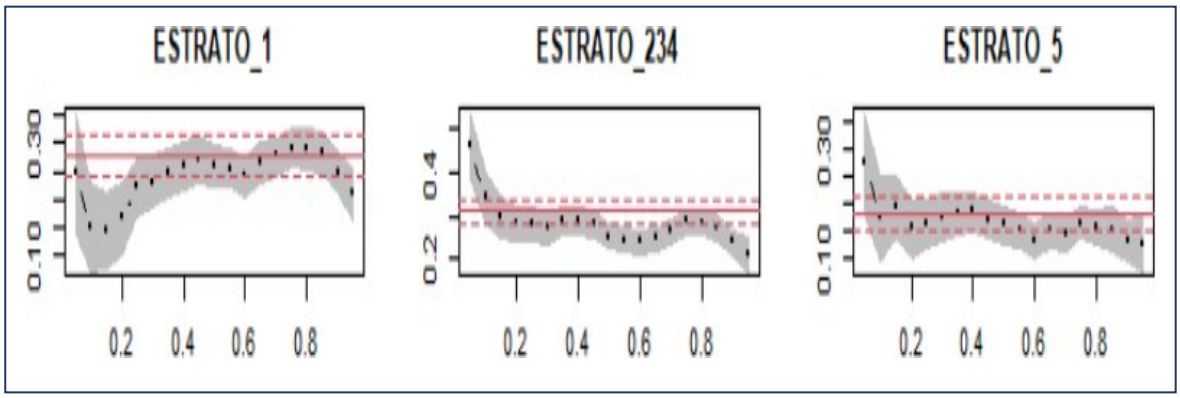
Graph 26. *Regression coefficients by quantiles of the Logarithm of total income*



Source: Based on information from ENAHO 2019
Elaboration: Own

In Graph 26 it is shown that the logarithm variable of total family income, its effect positively influences the monthly Basic Expenses 2 of the population of Peru in 2019, it also shows that the segments of the monthly Basic Expenses 2 high, their influence effect decreases because the population with high incomes does not necessarily have to spend more money on Basic Expenses 2.

Graph 27. Regression coefficients by quantiles of the Stratum



Source: Based on information from ENAHO 2019
Elaboration: Own

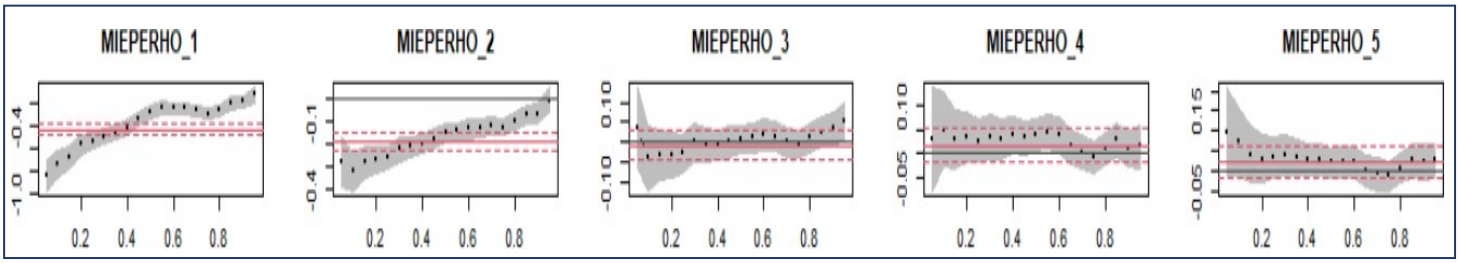
Graph 27 shows that:

- The variable "Estrato_1" (People who live in the stratum of 500 thousand or more inhabitants) its effect positively influences the Basic Expenses 2 of the population of Peru in 2019, thus it is also observed that up to the quantile 0.35 they spend less money in Basic Expenses 2 than OLS, however from the 0.40 quantile there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red

dotted lines for the OLS overlaps with the interval of confidence of the quantile regression with respect to the Basic Expenditures 2 of the population of Peru in the year 2019.

- The variable "Estrato_234" (People who live in the stratum where between 20,000 and 500,000 inhabitants live), its effect positively influences all quantiles, however, starting between the 0.15 and 0.85 quantile, there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression with respect to Basic Expenditures 2 of the population of Peru in 2019, a data not lower is that the quantile regression is seen to estimate in a better way in the tails.
- The variable "Estrato_5" (People who live in the stratum where between 2,000 and 20,000 inhabitants live) its effect positively influences all quantiles, however, starting between the quantile 0.10 and 0.85, there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression with respect to Basic Expenditures 2 of the population of Peru in 2019, a data not The lower is that the quantile regression is seen to estimate in a better way in the tails.

Graph 28. *Regression coefficients by quantiles of Mierpeho*



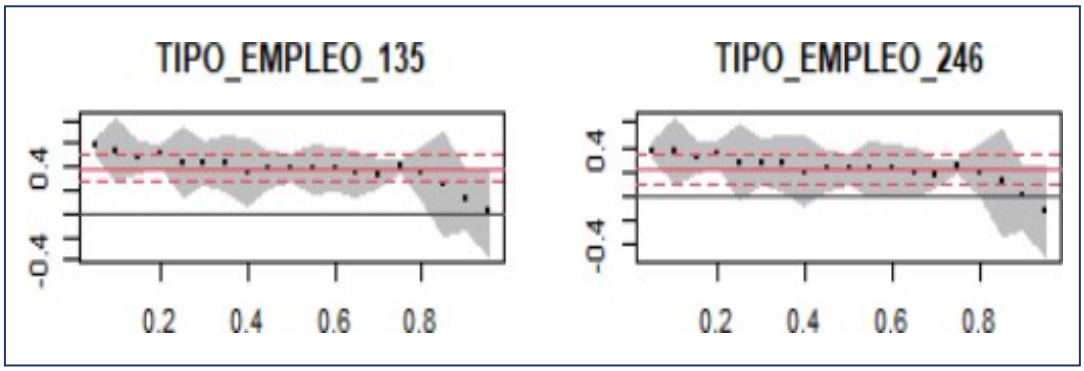
Source: Based on information from ENAHO 2019
Elaboration: Own

Graph 28 shows that:

- The variable "Mierperho_1" (1 household members in their home) and the variable "Mierperho_2" (2 household members in their home) its effect negatively influences all quantiles with respect to Basic Expenses 2 of the Peruvian population in In 2019, it is also observed that up to the 0.40 quantile they spend less money on Basic Expenses 2 than the OLS.
- The variable "Mierperho_3" (3 household members in their home), the variable "Mierperho_4" (4 household members in their home) and the variable "Mierperho_5" (5

household members in their home), there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression with respect to Basic Expenditures 2 of the population of Peru in the year 2019 in the entire path of the quantiles..

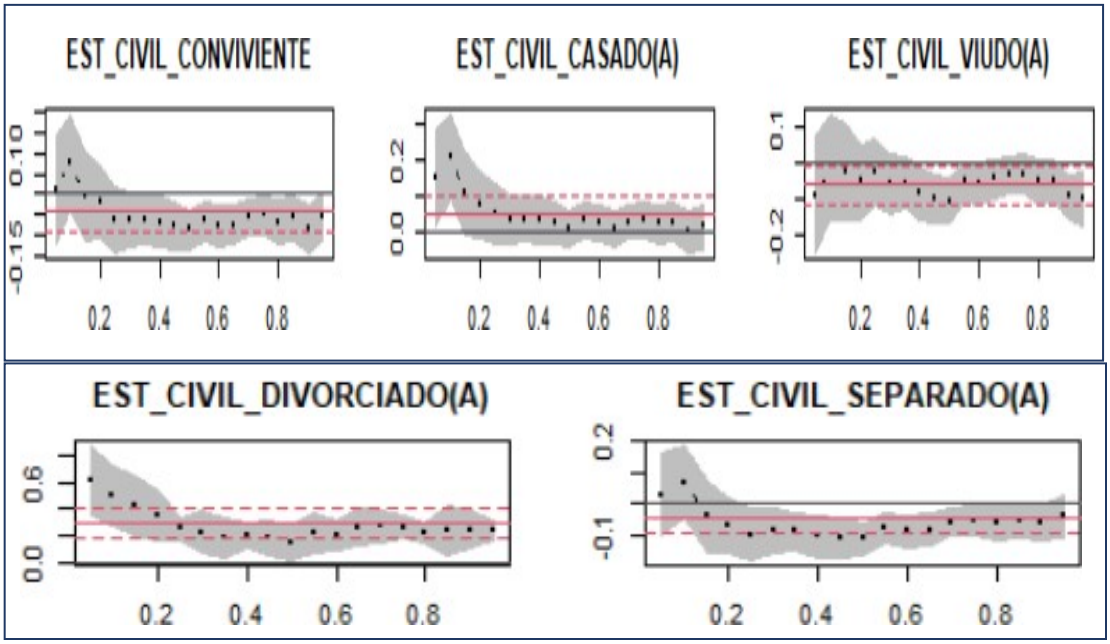
Graph 29. *Regression coefficients by quantiles of the Type of employment*



Source: Based on information from ENAHO 2019
Elaboration: Own

The variable “Tipo_Empleo_135” (People whose characteristics are “Employer or employer”, “Employee” and “Unpaid family worker”) and “Tipo_Empleo_246” (People whose characteristics are “Independent worker”, “Worker” and “ Domestic worker ”) its effect positively influences all quantiles, however, there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the interval of confidence of the quantile regression with respect to the Basic Expenditures 2 of the population of Peru in the year 2019.

*G**raph 30.** Regression coefficients by quantiles of the Civil Status*

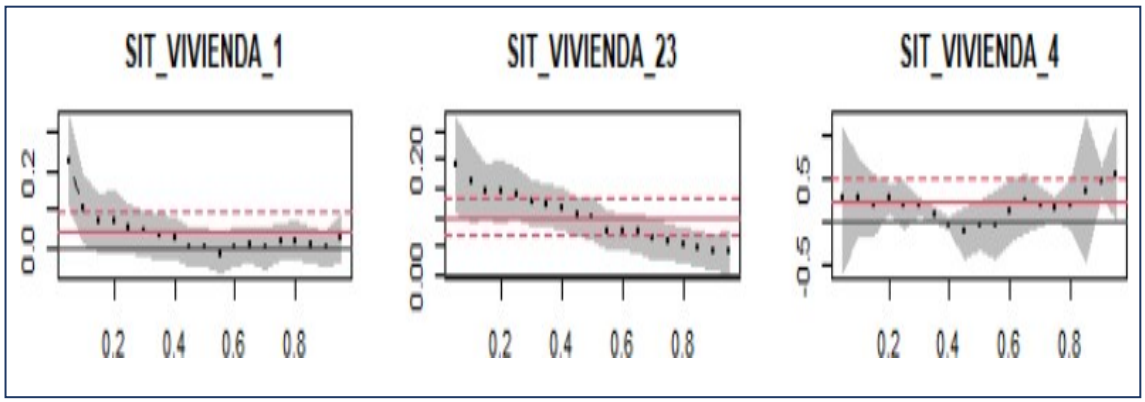


Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 30 it is observed that:

- The variable "Estado_Civil_Conviviente" and the variable "Estado_Civil_Viudo" their effect negatively influences all quantiles with respect to Basic Expenses 2 of the population of Peru in the year 2019.
- The variable “Estado_Civil_Casado”, its effect positively influences all quantiles with respect to Basic Expenses 2 of the population of Peru in 2019.
- The variable "Estado_Civil_Divorciado" and "Estado_Civil_Separado" their effect positively and negatively influences all quantiles respectively, however, there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for The OLS overlaps with the confidence interval of the quantile regression with respect to Basic Expenditures 2 of the population of Peru in the year 2019.

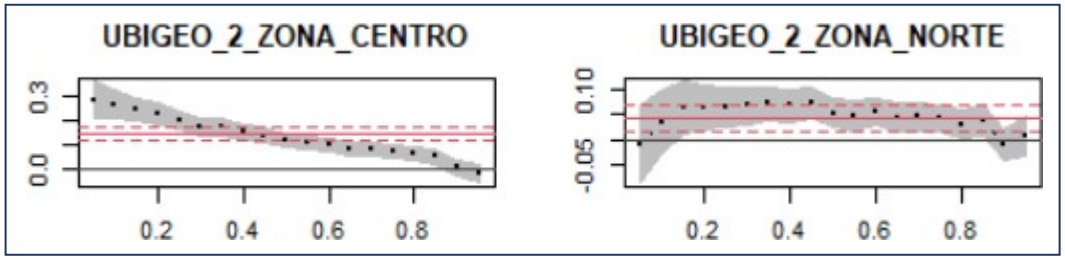
Graph 31. *Regression coefficients by quantiles of the Housing situation*



Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 31 it can be seen that for the dummy variables “Sit_Vivienda_1” (The dwelling you occupy is “Rented”), “Sit_Vivienda_23” (The dwelling you occupy are “Own, fully paid” and “Own, by invasion”) and “Sit_Vivienda_4 "(The home you occupy is" Own, buying it in installments), there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression with respect to Basic Expenses 2 of the population of Peru in the year 2019, except in the tails of the distribution.

Graph 32. *Quantile regression coefficients of Ubigeo*

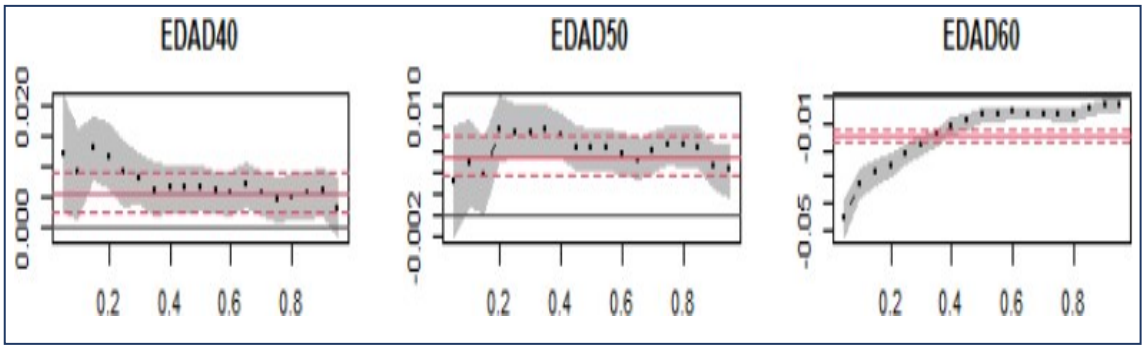


Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 33 it is observed that for the variables “Ubigeo_2_Zona_Centro” (People who live in the central Zone) and “Ubigeo_2_Zona_Norte” (People who live in the North Zone) its effect negatively influences all quantiles with respect to Basic Expenditures 2 of the population of Peru in 2019, it can also be observed that for people who are located in the Central Zone of Peru, Basic Expenses 2 have to decrease, in addition to the variable "Ubigeo_2_Zona_Norte" (People who live in the North Zone) no There is a lot of difference between the estimated effects of OLS and quantile

regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression in all quantiles with respect to the Basic Expenses 2 of the population of Peru in the year 2019.

Graph 34. Regression coefficients by quantiles of Age

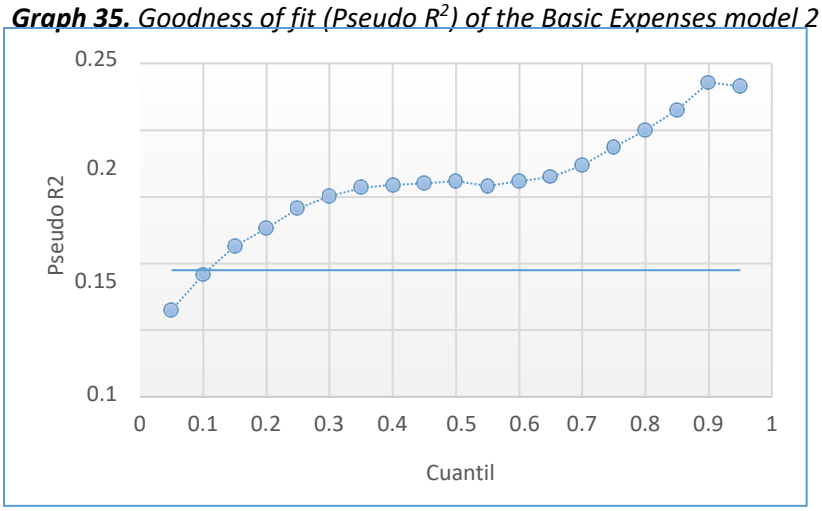


Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 34 it is observed that:

- In the variable "Edad_40" (Age of the head of the household under 40 years) and the variable "Edad_50" (Age of the head of the household between 40 and 59 years) it is observed that it has a positive influence with respect to the Basic Expenses 2 of the population of Peru in 2019, in addition to that as the quantiles increase, it is observed that when the head of the household is less than 40 years old, he spends less on Basic Expenses 2, in addition, there is not much difference between the estimated effects of OLS and regression quantile because the confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression, except for the tails of the variable “Edad_40” up to the quantile 0.20.
- In the variable "Edad_60" (Age of the head of the household greater than 60 years) it is observed that it negatively influences all quantiles with respect to Basic Expenses 2 of the population of Peru in 2019, it is also observed that people over 60 years of age spend less on Basic Expenses 2.

Graph 35. *Goodness of fit (Pseudo R^2) of the Basic Expenses model 2*



Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 35 it can be observed that as the quantiles on the response variable Basic Expenses 2 increase, we notice that the goodness of fit (Pseudo R2) tends to increase, it is more it can be observed that in the quantile 0.90 and 0.95 there are higher Pseudo R2 with a value of 23.56% and 23.27% respectively, due to this it is evidenced that the quantile regression is more flexible to explain the extremes of the response variable.

Household Expenses:

Table 16 shows the results of the different estimates made for the Household Expenses model. Among those that were taken into account to show the results are for the quantiles 0.10, 0.25, 0.40, 0.50, 0.60, 0.75, 0.90, and as an additional estimate the multiple linear regression. In almost all the estimates made, as well as for the multiple linear regression and the quantile regression, there are statistically significant parameters. Among those that have a positive influence are all variables except "Mieperho_1", "Mieperho_2", "Mieperho_3", "Mieperho_4", "Mieperho_5", "Estado_Civil_Viudo", "Edad40" and "Edad60". Now, it can be observed that in the dummy variables "Mieperho_1", "Mieperho_2", "Mieperho_3", "Mieperho_4", "Mieperho_5" negatively influence Household Expenses, that is, Household Expenses are reduced approximately between 0.050 and 0.685 with respect to tHousehold Expenditures of people who have 6 or more household members in their household, in reference to the multiple linear regression. It is observed that the relationship of Household Expenses and the variable "Total Income" tends to decrease as we place ourselves in segments of the population of people with high monthly Household Expenses, going from 21.8%

(Q10) to 27% (Q75), Now, for a more exhaustive analysis, the Q90 could also be reviewed since it increases by 3.75% with respect to the Q10, however, this parameter value does not differ much from the different regressions and it is still significant. In the variables "Edad50" it is observed that they have a positive influence with respect to Household Expenses, however, it can be observed that in the entire range of quantiles and the estimation by the multiple linear regression there is no significant difference in the values of the parameters. An important piece of information to observe is that the dummy variables "Estado_Civil_Conviviente", "Estado_Civil_Viudo", "Estado_Civil_Separado" are non-significant variables both for the multiple linear regression and in the quantile regression throughout their journey (Q10 to Q90). As a conclusion to the estimates of the quantile regression, we can obtain the different behaviors of the distribution with respect to Household Expenditures.

Table 22. *Results of the quantile regression (Parameters and standard error) of Household Expenditures*

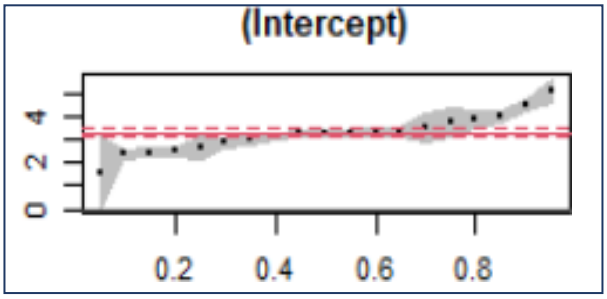
	OLS	**Quantile regression**						
		Quantile 10	**Quantile 25**	**Quantile 40**	**Quantile 50**	**Quantile 60**	**Quantile 75**	**Quantile 90**
LOG_INGRESO_TOTAL (Parameter)	0.234***	0.218***	0.231***	0.240***	0.244***	0.255***	0.270***	0.255***
Std error	0.005	0.008	0.006	0.006	0.005	0.006	0.007	0.01
ESTRATO_1 (Parameter)	0.557***	0.717***	0.593***	0.531***	0.511***	0.457***	0.456***	0.447***
Std error	0.015	0.022	0.017	0.018	0.017	0.02	0.023	0.027
ESTRATO_234 (Parameter)	0.493***	0.658***	0.533***	0.476***	0.457***	0.416***	0.386***	0.361***
Std error	0.011	0.019	0.014	0.013	0.013	0.014	0.016	0.021
ESTRATO_5 (Parameter)	0.305***	0.415***	0.328***	0.287***	0.283***	0.259***	0.240***	0.248***
Std error	0.014	0.021	0.017	0.016	0.016	0.018	0.019	0.026
MIEPERHO_1 (Parameter)	-0.685***	-1.043***	-0.807***	-0.725***	-0.682***	-0.613***	-0.510***	-0.319***
Std error	0.02	0.039	0.027	0.027	0.024	0.026	0.029	0.036
MIEPERHO_2 (Parameter)	-0.287***	-0.415***	-0.318***	-0.314***	-0.310***	-0.294***	-0.246***	-0.125***
Std error	0.017	0.025	0.02	0.019	0.018	0.021	0.022	0.031
MIEPERHO_3 (Parameter)	-0.135***	-0.180***	-0.148***	-0.169***	-0.172***	-0.158***	-0.118***	-0.027
Std error	0.016	0.023	0.018	0.018	0.016	0.019	0.022	0.026
MIEPERHO_4 (Parameter)	-0.084***	-0.101***	-0.086***	-0.112***	-0.113***	-0.108***	-0.078***	-0.018
Std error	0.016	0.02	0.016	0.017	0.016	0.019	0.02	0.025
MIEPERHO_5 (Parameter)	-0.050***	-0.041*	-0.048***	-0.069***	-0.065***	-0.065***	-0.070***	-0.025

Std error	0.017	0.021	0.018	0.02	0.017	0.02	0.023	0.031
TIPO_EMPLEO_135 (Parameter)	0.396***	0.300***	0.41	0.369***	0.479***	0.592***	0.375	0.223***
Std error	0.099	0.041	0.26	0.057	0.041	0.076	0.357	0.058
TIPO_EMPLEO_246 (Parameter)	0.209**	0.105***	0.229	0.191***	0.299***	0.407***	0.21	0.08
Std error	0.098	0.038	0.259	0.056	0.039	0.075	0.357	0.057
EST_CIVIL_CONVIVIENTE (Parameter)	0.005	-0.011	0.009	-0.041	-0.026	-0.026	-0.009	0.031
Std error	0.021	0.033	0.028	0.027	0.026	0.026	0.028	0.038
EST_CIVIL_CASADO(A) (Parameter)	0.061***	0.077**	0.081***	0.023	0.029	0.021	0.049*	0.069*
Std error	0.021	0.032	0.028	0.027	0.026	0.027	0.028	0.037
EST_CIVIL_VIUDO(A) (Parameter)	-0.014	-0.014	0.011	-0.037	-0.014	-0.026	-0.007	-0.03
Std error	0.023	0.04	0.032	0.03	0.029	0.03	0.034	0.041
EST_CIVIL_DIVORCIADO(A) (Parameter)	0.215***	0.223*	0.215***	0.188***	0.218***	0.214***	0.233***	0.164***
Std error	0.048	0.122	0.059	0.061	0.053	0.055	0.052	0.042
EST_CIVIL_SEPARADO(A) (Parámetro)	0.012	0.025	0.032	-0.028	-0.004	-0.002	-0.02	0.043
Std error	0.021	0.035	0.028	0.027	0.026	0.027	0.028	0.038
SIT_VIVIENDA_1 (Parameter)	0.041**	0.026	0.053**	0.033	0.022	0.021	0.038	0.019
Std error	0.019	0.032	0.021	0.02	0.019	0.023	0.027	0.032
SIT_VIVIENDA_23 (Parameter)	0.119***	0.116***	0.103***	0.105***	0.122***	0.110***	0.102***	0.097***
Std error	0.013	0.017	0.015	0.015	0.014	0.016	0.018	0.024
SIT_VIVIENDA_4 (Parameter)	0.133	0.133***	0.246***	0.115***	0.178	0.233*	0.054**	-0.01
Std error	0.1	0.037	0.048	0.036	0.118	0.136	0.022	0.087
UBIGEO_2_ZONA_CENTRO (Parameter)	0.130***	0.114***	0.129***	0.156***	0.151***	0.145***	0.146***	0.078***
Std error	0.011	0.017	0.013	0.013	0.013	0.015	0.016	0.021
UBIGEO_2_ZONA_NORTE (Parameter)	0.181***	0.167***	0.199***	0.213***	0.204***	0.185***	0.196***	0.127***
Std error	0.011	0.018	0.013	0.013	0.012	0.015	0.016	0.022
EDAD40 (Parameter)	-0.007***	-0.0002	-0.002	-0.006***	-0.009***	-0.010***	-0.012***	-0.013***
Std error	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.003
EDAD50 (Parameter)	0.002**	0.003***	0.003***	0.002***	0.002***	0.003***	0.001	-0.001
Std error	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
EDAD60 (Parameter)	-0.010***	-0.018***	-0.011***	-0.007***	-0.006***	-0.005***	-0.003**	-0.001
Std error	0.001	0.002	0.001	0.001	0.001	0.001	0.001	0.002
Constant (Parameter)	3.298***	2.367***	2.626***	3.112***	3.226***	3.286***	3.772***	4.481***
Std error	0.117	0.11	0.272	0.095	0.083	0.108	0.368	0.131

p<0.1; **p<0.05; ***p<0.01** **F Statistic= 586.127 (Significant)**

Source: Based on information from ENAHO
Preparation: Own

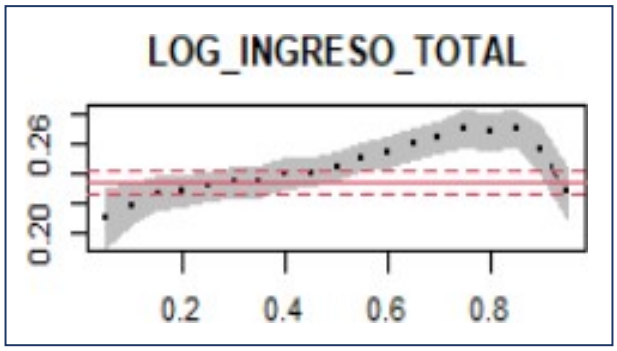
***Graph 36**. Regression coefficients by quantiles of the intercept*



Source: Based on information from ENAHO
Preparation: Own

Graph 36 shows the evolution of the intercept for the different levels of Household Expenditures of the population of Peru in 2019.

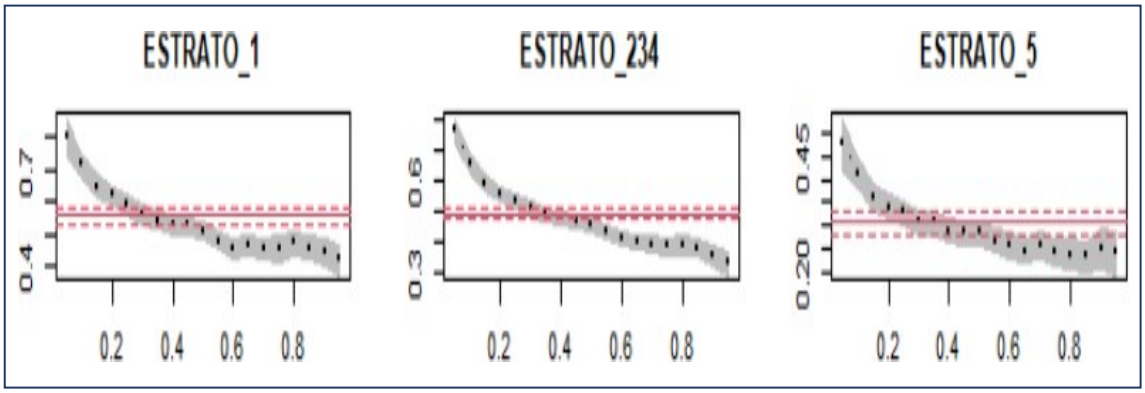
Graph 37. Regression coefficients by quantiles of the Logarithm of total



Source: Based on information from ENAHO
Preparation: Own

In Graph 37 it is shown that the logarithm variable of total family income, its effect positively influences the monthly Household Expenditures of the population of Peru in the year 2019, it also shows that the segments of the high monthly Household Expenditures its effect of influence is increasing to the high income of people up to the quantile 0.85.

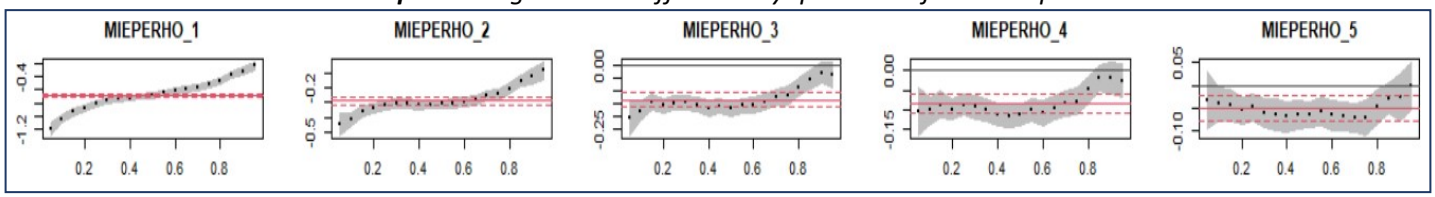
Graph 38. *Regression coefficients by quantiles of the Stratum*



Source: Based on information from ENAHO 2019
Elaboration: Own

Graph 38 shows that the variable "Estrato_1" (People who live in the stratum of 500 thousand or more inhabitants), the variable "Estrato_234" (People who live in the stratum where between 20 thousand and 500 thousand inhabitants live), The variable "Estrato_5" (People who live in the stratum where between 2,000 and 20,000 inhabitants live) its effect has a positive influence with respect to Household Expenditures in all quantiles of the population of Peru in 2019, as well as that the segments of high monthly Household Expenses are lower for the three variables, thus it is also shown that up to the 0.25 quantile Household Expenses for the quantile regression are higher compared to the OLS estimate.

Graph 39. *Regression coefficients by quantiles of the Mierpeho*



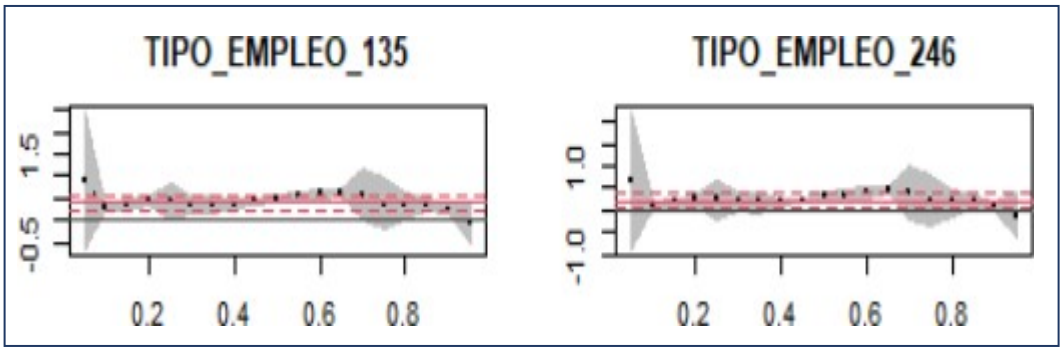
Source: Based on information from ENAHO 2019
Elaboration: Own

Graph 39 shows that:

- The variable "Mierperho_1" (1 household members in their home) and the variable "Mierperho_2" (2 household members in their home) their effect negatively influences all quantiles with respect to Household Expenditures of the Peruvian population in the year 2019, thus it is also observed that Household Expenses tend to decrease when you have 1 or 2 members per household.

- The variable “Mierperho_3” (3 household members in their home), the variable “Mierperho_4” (4 household members in their home) and the variable “Mierperho_5” (5 household members in their home) there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression with respect to Household Expenditures of the population of Peru in the year 2019 in all the path of the quantiles, except for the distribution tails as shown in the variables "Mierperho_3" and "Mieperho_4".

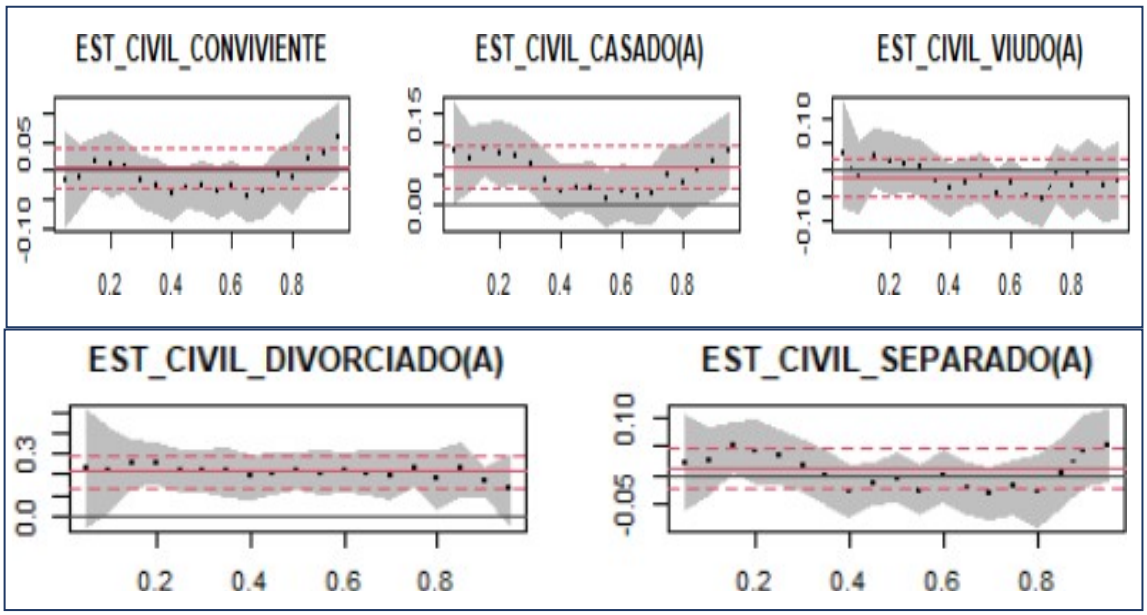
Graph 40. *Regression coefficients by quantiles of the Type of employment*



Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 40 it is observed that the variable "Type_Empleo_135" (People who have the characteristic of being "Employer or employer", "Employee" and "Unpaid family worker") and "Type_Empleo_246" (People who have the characteristic of being "Independent worker "," Worker "and" Household worker ") their effect positively influences all quantiles, however, there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for The OLS overlaps with the confidence interval of the quantile regression with respect to Household Expenditures of the Peruvian population in 2019, except in both in the quantile 0.05

Graph 41. *Regression coefficients by quantiles of Civil Status*

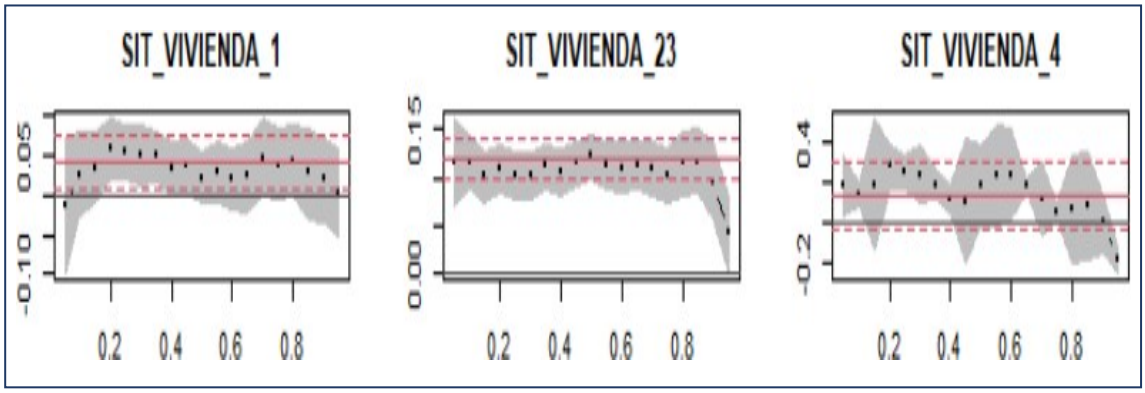


Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 41 it is observed that:

- The variable "Estado_Civil_Conviviente", the variable "Estado_Civil_Viudo" and the variable "Estado_Civil_Separado" their effect positively influence up to the
- quantile 0.35, then onwards they influence negatively with respect to Household Expenditures of the population of Peru in the year 2019, as well as Note there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression with respect to Household Expenditures of the population of the Peru in 2019.
- The variable “Estado_Civil_Casado” and the variable “Estado_Civil_Divorciado” their effect positively influences all quantiles with respect to Household Expenditures, however, like the previous paragraph, there is no significant difference between the estimated effects of OLS and quantile regression in all quantiles.

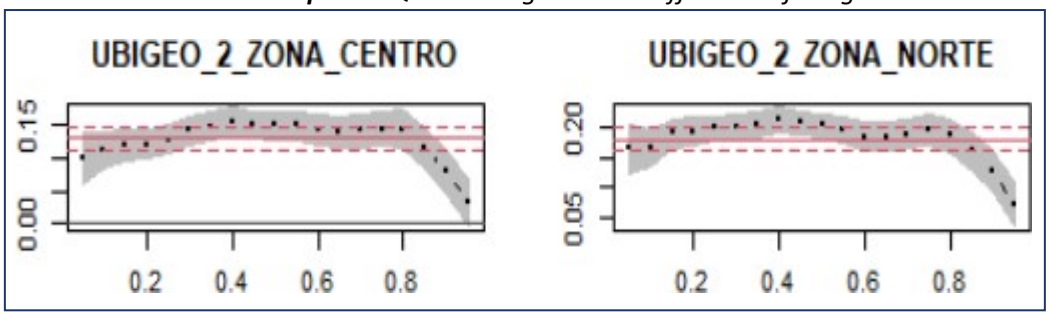
Graph 42. *Regression coefficients by quantiles of the Housing*



Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 42 it can be observed that for the dummy variables “Sit_Vivienda_1” (The dwelling you occupy is “Rented”), “Sit_Vivienda_23” (The dwelling you occupy are “Own, fully paid for” and “Own, by invasion”) and “Sit_Vivienda_4 "(The home you occupy is" Own, buying it in installments) its effect positively influences all quantiles, thus it is also observed that there is no difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the lines Red dots for the OLS overlap with the confidence interval of the quantile regression with respect to Household Expenditures of the population of Peru in the year 2019, except in the tails (Q95) of the distribution of the variable "Sit_Vivienda_23" and " Sit_Vivienda_4 ”.

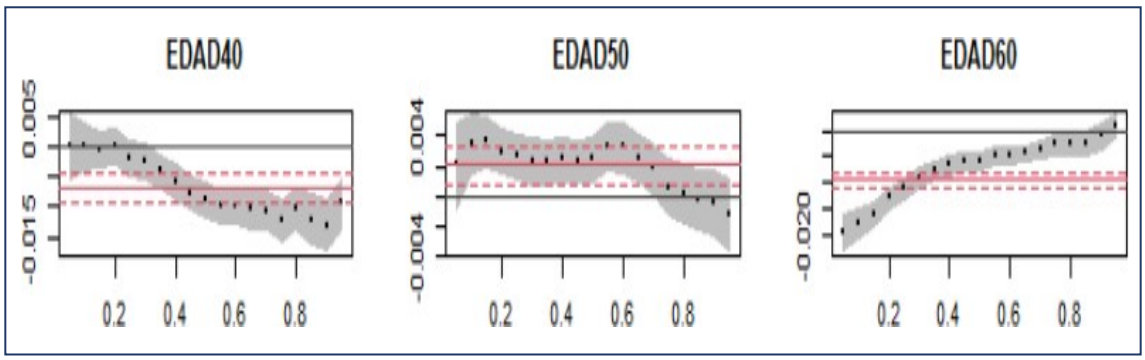
Graph 43. *Quantile regression coefficients of Ubigeo*



Source: Based on information from ENAHO 2019
Elaboration: Own

Del From Graph 43 it is observed that for the variables “Ubigeo_2_Zona_Centro” (People who live in the central Zone) and “Ubigeo_2_Zona_Norte” (People who live in the North Zone) its effect positively influences all the quantiles with respect to the Expenses Monthly household expenses of the population of Peru in 2019, it can also be observed that there is no significant difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression in all the quantiles with respect to the Household Expenditures of the population of Peru in the year 2019, except in the last quantiles of the distribution.

Graph 44. *Regression coefficients by quantiles of Age*

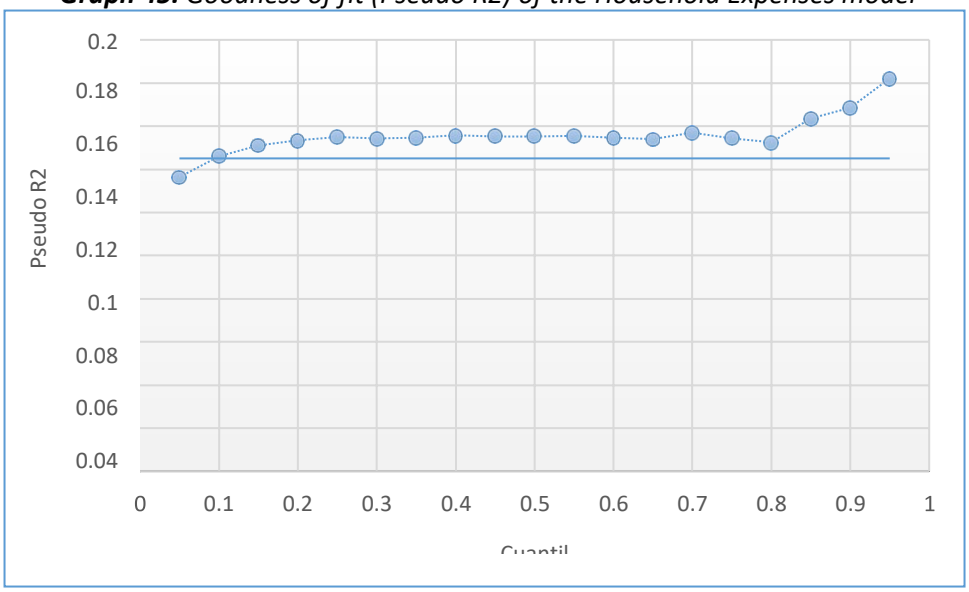


Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 44 it is observed that:

- In the variable “Edad_40” (Age of the head of the household less than 40 years) and the variable “Edad_60” (Age of the head of the household greater than 60 years) it is observed that it has a negative influence with respect to the Household Expenditures of the population of the Peru in 2019, in addition to that as the quantiles increase, it is observed that when the head of the household is less than 40 years old, he spends less on Household Expenses and vice versa, it happens when the head of the household is over 60 years old.
- In the variable "Edad_50" (Age of the head of the household older than 40 and 60 years) it is observed that it positively influences all the quantiles with respect to the Household Expenditures of the population of Peru in the year 2019, it is also observed that there is no significant difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression in all quantiles with respect to Household Expenditures of the population of Peru in 2019, except in the last quantiles of the distribution.

Graph 45. *Goodness of fit (Pseudo R2) of the Household Expenses model*



Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 45 it can be observed that as the quantiles on the response variable Household Expenses increase, we note that the goodness of fit (Pseudo R2) tends to increase, it is more it can be observed that in the quantile 0.95 where there is greater Pseudo R2 with a value of 18.21%, due to this it is evident how the quantile regression is more flexible to explain the extremes of the response variable.

Luxury Expenses:

Table 17 shows the results of the different estimates made for the Luxury Expenses model. Among those that were taken into account to show the results are for the quantiles 0.10, 0.25, 0.40, 0.50, 0.60, 0.75, 0.90, and as an additional estimate the multiple linear regression. In almost all the estimates made, as well as for the multiple linear regression and the quantile regression, there are statistically significant parameters. Among those that have a positive influence are all variables except for "Mieperho_1", "Mieperho_2", "Mieperho_3", "Mieperho_4", "Mieperho_5", "Estado_Civil_Conviviente", "Estado_Civil_Casado", "Estado_Civil_Separadoorte2", "Estado_Civil_Separadoorte2", "Age_Civil_Separadoorte2" "And" Age60 ". Now, it can be seen that in the dummy variables "Mieperho_1", "Mieperho_2", "Mieperho_3", "Mieperho_4", "Mieperho_5", they negatively influence Luxury Expenses, that is, Luxury Expenses are reduced

approximately between 0.154 and 2.3 Regarding the Luxury Expenses of people who have 6 or more members in their household in their household, in reference to the multiple linear regression. It is observed that the relationship of Luxury Expenses and the variable “Total Income” tends to decrease as we place ourselves in segments of the population of people with high monthly Luxury Expenses, going from 27.3% (Q10) to 26.6% (Q90). An important fact to observe is that the dummy variable “Sit_Vivienda_23” is a non-significant variable both for the multiple linear regression and in the quantile regression throughout its journey (Q10 to Q90), however, the variable “Sit_Vivienda_1” is only significant in Q75 and the variable "Sit_Vivienda_4" is only significant in Q25. As a conclusion to the estimates of the quantile regression, we can obtain the different behaviors of the distribution with respect to Luxury Expenses.

Table 23. *Results of the quantile regression (Parameters and standard error) of Luxury Expenses*

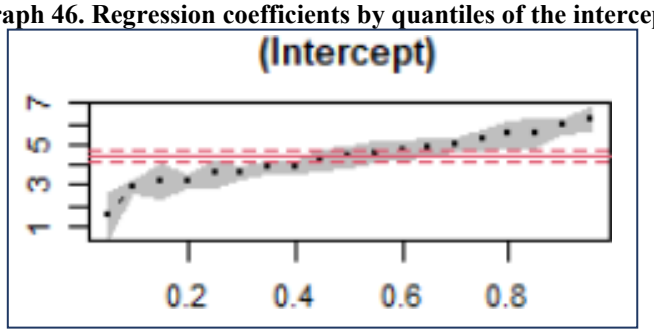
	OLS	**Quantile Regression**						
		Quantile 10	**Quantile 25**	**Quantile 40**	**Quantile 50**	**Quantile 60**	**Quantile 75**	**Quantile 90**
LOG_INGRESO_TOTAL (Parameter)	0.278***	0.273***	0.272***	0.274***	0.268***	0.268***	0.268***	0.266***
Std error	0.007	0.015	0.01	0.009	0.008	0.008	0.008	0.008
ESTRATO_1 (Parameter)	0.353***	0.232***	0.293***	0.333***	0.370***	0.407***	0.466***	0.443***
Std error	0.023	0.046	0.034	0.028	0.027	0.024	0.023	0.023
ESTRATO_234 (Parameter)	0.506***	0.535***	0.475***	0.484***	0.475***	0.491***	0.498***	0.434***
Std error	0.017	0.032	0.022	0.019	0.017	0.017	0.017	0.019
ESTRATO_5 (Parameter)	0.249***	0.302***	0.246***	0.231***	0.232***	0.242***	0.267***	0.213***
Std error	0.02	0.037	0.025	0.023	0.02	0.021	0.019	0.023
MIEPERHO_1 (Parameter)	-2.300***	-2.846***	-2.691***	-2.516***	-2.410***	-2.285***	-1.950***	-1.364***
Std error	0.03	0.052	0.042	0.038	0.037	0.037	0.042	0.037
MIEPERHO_2 (Parameter)	-1.587***	-2.165***	-1.981***	-1.731***	-1.621***	-1.443***	-1.195***	-0.886***
Std error	0.025	0.045	0.034	0.031	0.028	0.029	0.027	0.028
MIEPERHO_3 (Parameter)	-0.734***	-1.175***	-0.900***	-0.693***	-0.646***	-0.575***	-0.492***	-0.368***
Std error	0.024	0.044	0.03	0.023	0.023	0.021	0.021	0.023
MIEPERHO_4 (Parameter)	-0.313***	-0.424***	-0.350***	-0.270***	-0.263***	-0.253***	-0.234***	-0.174***
Std error	0.023	0.038	0.023	0.021	0.019	0.017	0.018	0.02

MIEPERHO_5 (Parameter)	-0.154***	-0.180***	-0.181***	-0.128***	-0.137***	-0.131***	-0.106***	-0.090***
Std error	0.026	0.041	0.024	0.021	0.019	0.021	0.019	0.021
TIPO_EMPLEO_135 (Parameter)	0.731***	0.573***	0.646*	0.837***	0.808***	0.676***	0.604**	0.268
Std error	0.146	0.095	0.357	0.126	0.299	0.222	0.259	0.208
TIPO_EMPLEO_246 (Parameter)	0.447***	0.287***	0.347	0.565***	0.526*	0.407*	0.369	0.058
Std error	0.145	0.089	0.357	0.125	0.299	0.221	0.259	0.208
EST_CIVIL_CONVIVIENTE (Parameter)	-0.347***	-0.141**	-0.139***	-0.293***	-0.376***	-0.383***	-0.452***	-0.510***
Std error	0.031	0.065	0.048	0.044	0.048	0.044	0.044	0.036
EST_CIVIL_CASADO(A) (Parameter)	-0.260***	-0.078	-0.076	-0.203***	-0.284***	-0.296***	-0.362***	-0.426***
Std error	0.031	0.067	0.048	0.045	0.049	0.044	0.044	0.037
EST_CIVIL_VIUDO(A) (Parameter)	0.087**	0.102	0.176***	0.081	0.012	0.022	-0.029	-0.085**
Std error	0.034	0.071	0.057	0.051	0.053	0.05	0.05	0.041
EST_CIVIL_DIVORCIADO(A) (Parameter)	0.241***	0.285***	0.415***	0.166	0.236**	0.224***	0.192**	-0.003
Std error	0.071	0.108	0.149	0.134	0.118	0.082	0.093	0.063
EST_CIVIL_SEPARADO(A) (Parameter)	-0.001	0.09	0.122**	0.016	-0.044	-0.046	-0.123***	-0.233***
Std error	0.031	0.065	0.049	0.045	0.049	0.045	0.045	0.038
SIT_VIVIENDA_1 (Parameter)	0.051*	0.057	0.044	0.059*	0.047	0.034	0.044*	0.012
Std error	0.029	0.058	0.036	0.033	0.03	0.029	0.024	0.031
SIT_VIVIENDA_23 (Parameter)	0.026	0.026	0.012	0.029	0.01	0.009	0.015	0.004
Std error	0.019	0.034	0.025	0.022	0.02	0.019	0.018	0.02
SIT_VIVIENDA_4 (Parameter)	0.169	0.418	0.338***	0.175*	0.122	0.1	0.076	0.104
Std error	0.149	0.255	0.128	0.092	0.227	0.157	0.101	0.136
UBIGEO_2_ZONA_CENTRO (Parameter)	0.02	0.193***	0.082***	0.031	-0.0003	-0.034*	-0.097***	-0.111***
Std error	0.017	0.036	0.025	0.021	0.018	0.018	0.017	0.018
UBIGEO_2_ZONA_NORTE (Parameter)	-0.082***	0.05	-0.034	-0.088***	-0.121***	-0.150***	-0.182***	-0.175***
Std error	0.017	0.035	0.024	0.02	0.018	0.018	0.018	0.018
EDAD40 (Parameter)	0.002	0.015***	0.012***	0.008***	0.005***	0.004**	-0.00005	0.002
Std error	0.002	0.003	0.003	0.002	0.002	0.002	0.002	0.002
EDAD50 (Parameter)	-0.024***	-0.041***	-0.030***	-0.023***	-0.020***	-0.018***	-0.013***	-0.009***
Std error	0.001	0.002	0.002	0.001	0.001	0.001	0.001	0.001
EDAD60 (Parameter)	-0.026***	-0.031***	-0.033***	-0.032***	-0.030***	-0.028***	-0.023***	-0.012***
Std error	0.001	0.003	0.002	0.002	0.002	0.002	0.002	0.002
Constant (Parameter)	4.390***	2.939***	3.588***	3.917***	4.368***	4.707***	5.223***	5.880***

Std error	0.173	0.19	0.38	0.163	0.316	0.243	0.277	0.229
p<0.1; **p<0.05; ***p<0.01 F Statistic= 1120.261 (Significant)**								

Source: Based on information from ENAHO
Preparation: Own

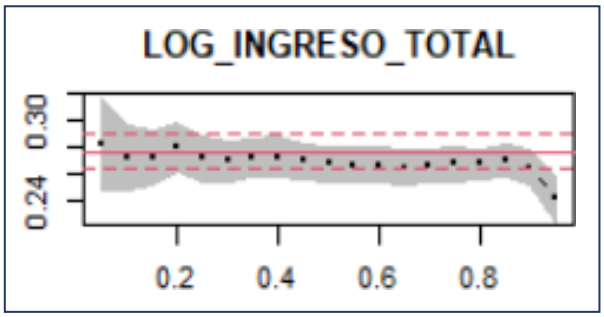
Graph 46. Regression coefficients by quantiles of the intercept



Source: Based on information from ENAHO 2019
Elaboration: Own

Graph 46 shows the evolution of the intercept for the different levels of monthly Luxury Expenses of the population of Peru in 2019.

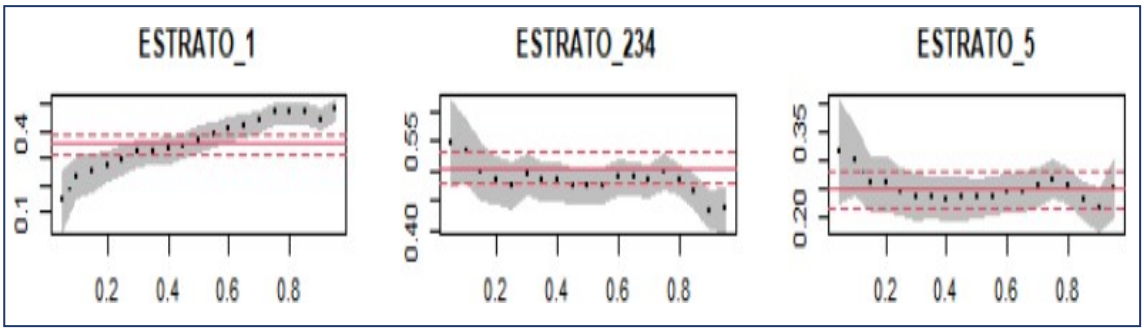
Graph 47. *Regression coefficients by quantiles of the Logarithm of total income total*



Source: Based on information from ENAHO 2019
Elaboration: Own

Graph 47 shows that the logarithm variable of total family income, its effect positively influences the monthly Luxury Expenses of the population of Peru in 2019, in addition there are no significant differences between the estimated effects of OLS and quantile regression because The confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression with respect to the monthly Luxury Expenditures of the population of Peru in the year 2019 in the entire range of the quantiles, except in the distribution queues.

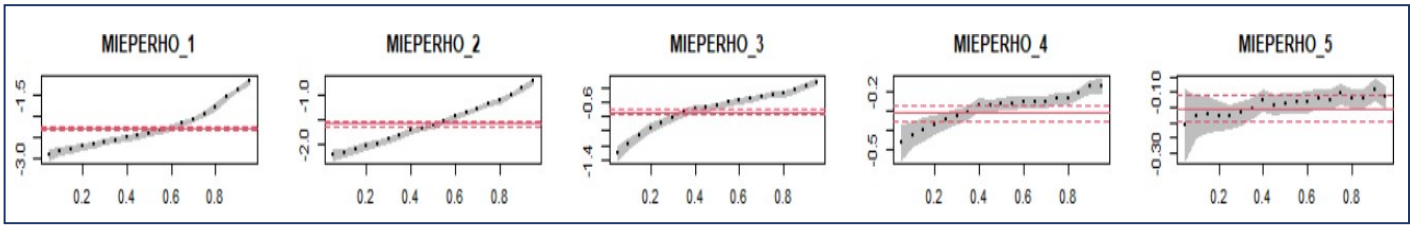
Graph 48. *Regression coefficients by quantiles of the Stratum*



Source: Based on information from ENAHO 2019
Elaboration: Own

Graph 48 shows that the variable “Estrato_1” (People who live in the stratum of 500 thousand or more inhabitants), the variable “Estrato_234” (People who live in the stratum where between 20 thousand and 500 thousand inhabitants live), the variable “Estrato_5” (People who live in the stratum where between 2 thousand and 20 thousand inhabitants live), its effect positively influences the Luxury Expenses in all the quantiles of the population of Peru in the year 2019, as well as that people who are in stratum 1 have a higher expenditure in Luxury Expenses, in addition there is no significant difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for the OLS overlaps with the confidence interval of the quantile regression with respect to the monthly Luxury Expenditures of the population of Peru in the year 2019 in the entire range of the quantiles of the variables " Estrato_234" and "Estrato_5", except in the distribution tails.

Graph 49. *Regression coefficients by quantiles of Mierpeho*



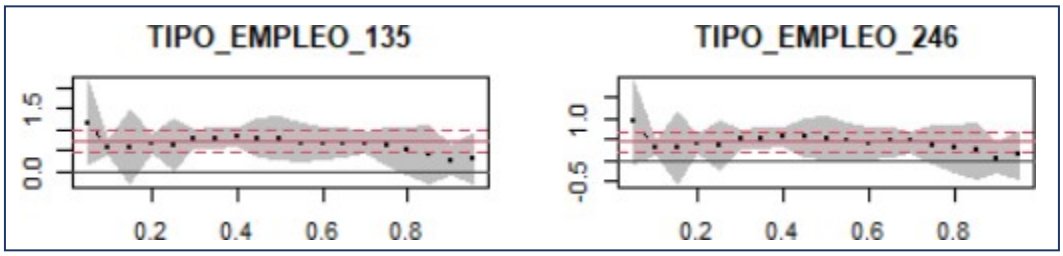
Source: Based on information from ENAHO 2019
Elaboration: Own

Graph 49 shows that:

The variable “Mierperho_1” (1 household members in their home), the variable “Mierperho_2” (2 household members in their home), the variable “Mierperho_3” (3 household members in their

home), the variable “Mierperho_4” (4 household members in their home) and the variable "Mierperho_5" (5 household members in their home), its effect negatively influences all quantiles with respect to the monthly Luxury Expenditures of the Peruvian population in the year 2019, thus It is also observed that Luxury expenses tend to increase in relation to more members per household in families.

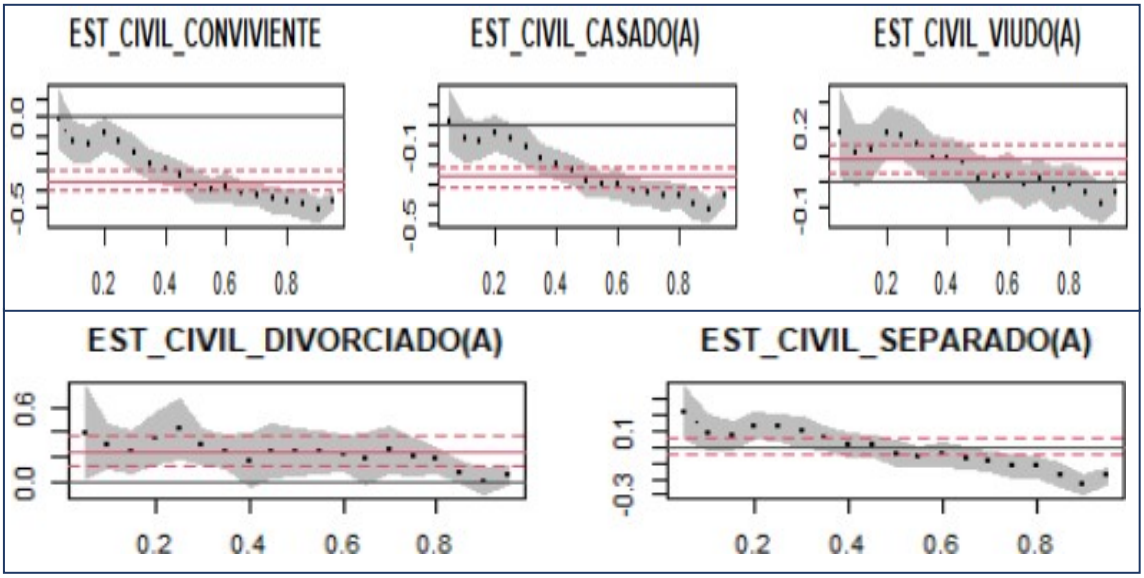
Graph 50. *Regression coefficients by quantiles of the Type of employment*



Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 50 it is observed that the variable "Type_Empleo_135" (People whose characteristics are "Employer or employer", "Employee" and "Unpaid family worker") and "Type_Empleo_246" (People who have the characteristic of being "Independent worker "," Worker "and" Household worker ") their effect positively influences all quantiles, however, there is not much difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the red dotted lines for The OLS overlaps with the confidence interval of the quantile regression with respect to the monthly Luxury Expenditures of the population of Peru in the year 2019, except in both in the quantile 0.05.

Graph 51. *Regression coefficients by quantiles of Civil Status*

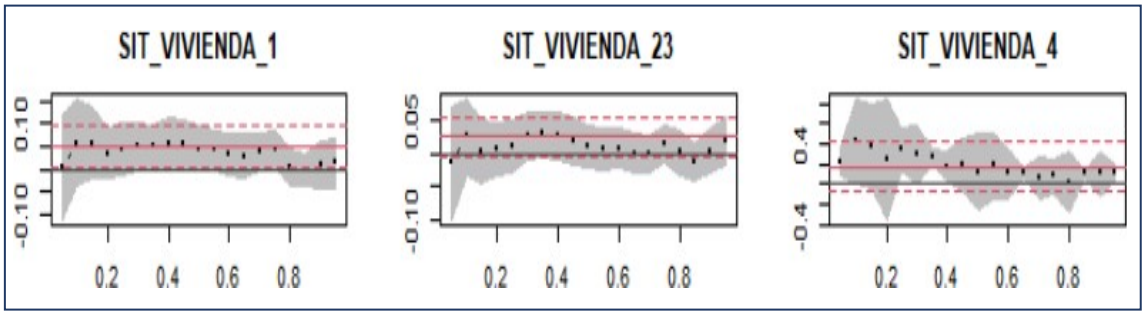


Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 51 it is observed that:

- The variable “Estado_Civil_Conviviente” and the variable “Estado_Civil_Civil_Casado” their effect negatively influences all the quantiles with respect to the monthly Luxury Expenses of the population of Peru in the year 2019, thus also as the quantiles increase, the monthly Luxury Expenses tend to decrease in the heads of household with cohabiting and married marital status.
- The variable "Estado_Civil_Viudo" and the variable "Estado_Civil_Divorciado" their effect positively influences all quantiles with respect to Household Expenditures, however, there is no significant difference between the estimated effects of OLS and quantile regression because the confidence interval shown The red dotted lines for the OLS overlap with the confidence interval of the quantile regression with respect to the monthly Luxury Expenditures of the population of Peru in the year 2019, except in the quantile 0.90 and 0.95.
- The variable “Estado_Civil_Separado”, its effect has a positive influence up to the 0.5 quantile and then onwards it has a negative influence with respect to the monthly Luxury Expenses of the Peruvian population in the year 2019.

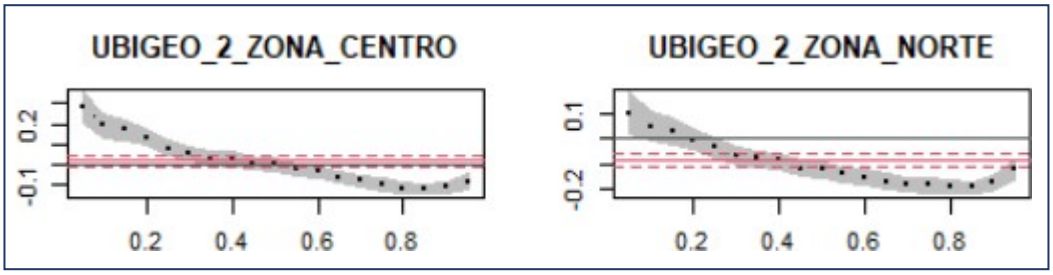
Graph 52. *Regression coefficients by quantiles of the Housing situation*



Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 52 it can be seen that for the dummy variables “Sit_Vivienda_1” (The dwelling you occupy is “Rented”), “Sit_Vivienda_23” (The dwelling you occupy are “Own, fully paid for” and “Own, by invasion”) and “Sit_Vivienda_4 "(The home you occupy is" Own, buying it in installments) its effect positively influences all quantiles, thus it is also observed that there is no significant difference between the estimated effects of OLS and quantile regression because the confidence interval shown by the Red dotted lines for the OLS overlap with the confidence interval of the quantile regression with respect to the Luxury Expenditures of the population of Peru in the year 2019.

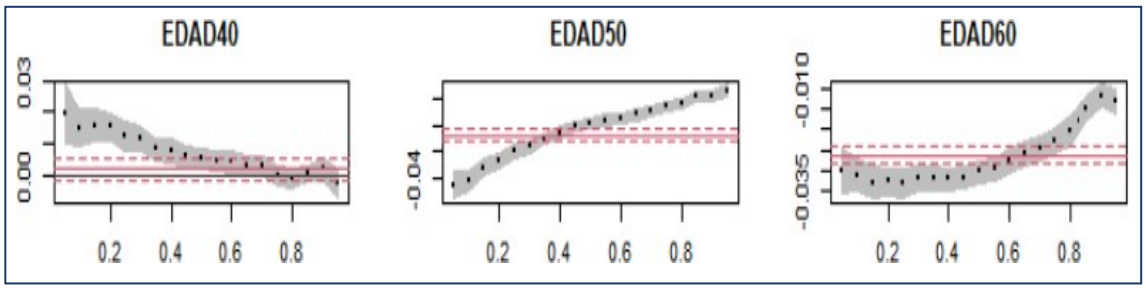
Graph 53. *Quantile regression coefficients of Ubigeo*



Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 53 it can be observed that for the variable “Ubigeo_2_Zona_Centro” (People who live in the central Zone) that has a positive influence up to the quantile 0.4 and then onwards has a negative influence with respect to the monthly Luxury Expenses of the population of Peru in the year 2019 and “Ubigeo_2_Zona_Norte” (People who live in the North Zone) its effect up to the quantile 0.2 and then onwards has a negative influence with respect to the monthly Luxury Expenses of the population of Peru in the year 2019.

Graph 54. Regression coefficients by quantiles of Age

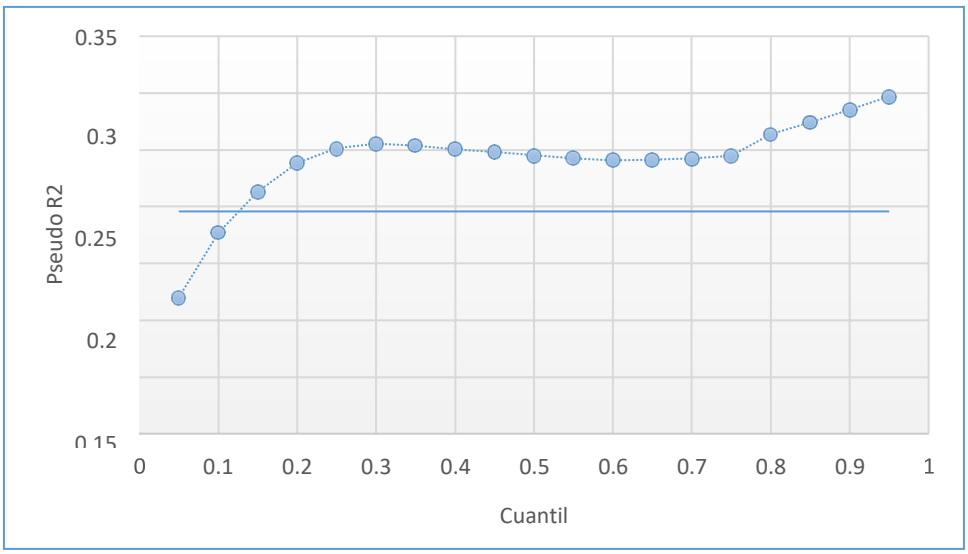


Source: Based on information from ENAHO 2019
Elaboration: Own

From Graph 54 it is observed that:

- In the variable “Edad_50” (Age of the oldest head of the household between 40 and 60 years) and the variable “Edad_60” (Age of the oldest head of the household equal to 60 years) it is observed that it negatively influences all the quantiles with respect to the Monthly Luxury Expenses of the Peruvian population in 2019, in addition to increasing the quantiles, it is observed that when the head of the household is over 40 years old, he spends less on Luxury Expenses.
- In the variable "Age_40" (Age of the head of the household under 40 years) it is observed that it positively influences all the quantiles with respect to the monthly Luxury Expenses of the population of Peru in the year 2019, its influence effect
- decreases this Due to the fact that the population that has high Luxury Expenses does not necessarily tend to spend the heads of household under 40 years of age.

Graph 55. *Goodness of fit (Pseudo R^2) of the Luxury Expenses model*



- Source: Based on information from ENAHO 2019
- Elaboration: Own

From Graph 55 it can be observed that as the quantiles on the Luxury Expenses response variable increase, we notice that the goodness of fit (Pseudo R2) tends to increase, it is more can be observed that in the 0.95 quantile where there is a greater Pseudo R2 with a value of 29.69%, due to this it is evidenced that the quantile regression is more flexible to explain the extremes of the response variable.

4.6.2 Multiple linear regression

Next, the parameters and predictive capacity of each spending model will be shown.

Graph 56. *Parameters of the Basic expenses model 1*

```
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.1055211  0.0857396  59.547  < 2e-16 ***
LOG_INGRESO_TOTAL        0.1773896  0.0036982  47.966  < 2e-16 ***
ESTRATO_1               -0.0730862  0.0112076  -6.521 7.09e-11 ***
ESTRATO_234              0.1094432  0.0083921  13.041  < 2e-16 ***
ESTRATO_5                0.0537662  0.0101458   5.299 1.17e-07 ***
MIEPERHO_1              -1.0915528  0.0148182 -73.663  < 2e-16 ***
MIEPERHO_2              -0.7727229  0.0123850 -62.392  < 2e-16 ***
MIEPERHO_3              -0.4716911  0.0118763 -39.717  < 2e-16 ***
MIEPERHO_4              -0.2536910  0.0115488 -21.967  < 2e-16 ***
MIEPERHO_5              -0.1304971  0.0126501 -10.316  < 2e-16 ***
TIPO_EMPLEO_135          0.3825882  0.0722133   5.298 1.18e-07 ***
TIPO_EMPLEO_246          0.2240380  0.0718740   3.117 0.001828 **
EST_CIVIL_CONVIVIENTE   -0.0509611  0.0153016  -3.330 0.000868 ***
EST_CIVIL_CASADO(A)     -0.0074207  0.0153270  -0.484 0.628276
EST_CIVIL_VIUDO(A)       0.0701589  0.0167425   4.190 2.79e-05 ***
EST_CIVIL_DIVORCIADO(A)  0.1250571  0.0351939   3.553 0.000381 ***
EST_CIVIL_SEPARADO(A)    0.0462368  0.0151530   3.051 0.002280 **
SIT_VIVIENDA_1           0.0517402  0.0142597   3.628 0.000286 ***
SIT_VIVIENDA_23          0.0258354  0.0093884   2.752 0.005930 **
SIT_VIVIENDA_4           0.0919990  0.0735452   1.251 0.210975
UBIGEO_2_ZONA_CENTRO     0.1487661  0.0084100  17.689  < 2e-16 ***
UBIGEO_2_ZONA_NORTE     -0.0124074  0.0084043  -1.476 0.139873
EDAD40                  -0.0004875  0.0010436  -0.467 0.640397
EDAD50                  -0.0090556  0.0005827 -15.541  < 2e-16 ***
EDAD60                  -0.0026351  0.0006086  -4.330 1.50e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5505 on 29340 degrees of freedom
Multiple R-squared:  0.4254,    Adjusted R-squared:  0.425
F-statistic: 905.2 on 24 and 29340 DF,  p-value: < 2.2e-16
```

Source: Based on information from ENAHO 2019
Elaboration: Own

- It can be seen that the variable “Marital status - married” involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the basic expenses 1 model using multiple linear regression.
- It can be seen that the variable "Housing situation - Own, buying it in installments" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the basic expenses 1 model using multiple linear regression.
- It can be seen that the variable "Ubigeo - Zona Norte" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the basic expenses 1 model using multiple linear regression.
- It can be seen that the variable “Edad_40” involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the basic expenses 1 model using multiple linear regression.

Graph 57. *Parameters of the Basic expenses model 2*

```
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.593631   0.164378   3.611 0.000305 ***
LOG_INGRESO_TOTAL        0.415261   0.007090  58.568  < 2e-16 ***
ESTRATO_1                0.279057   0.021487  12.987  < 2e-16 ***
ESTRATO_234              0.310324   0.016089  19.288  < 2e-16 ***
ESTRATO_5                0.182923   0.019451   9.404  < 2e-16 ***
MIEPERHO_1              -0.435633   0.028409 -15.334  < 2e-16 ***
MIEPERHO_2              -0.194438   0.023744  -8.189 2.74e-16 ***
MIEPERHO_3              -0.009662   0.022769  -0.424 0.671326
MIEPERHO_4               0.017105   0.022141   0.773 0.439808
MIEPERHO_5               0.026469   0.024253   1.091 0.275117
TIPO_EMPLEO_135          0.725210   0.138446   5.238 1.63e-07 ***
TIPO_EMPLEO_246          0.384416   0.137795   2.790 0.005278 **
EST_CIVIL_CONVIVIENTE   -0.045266   0.029336  -1.543 0.122833
EST_CIVIL_CASADO(A)      0.050210   0.029385   1.709 0.087512 .
EST_CIVIL_VIUDO(A)      -0.061999   0.032098  -1.932 0.053429 .
EST_CIVIL_DIVORCIADO(A)  0.290260   0.067473   4.302 1.70e-05 ***
EST_CIVIL_SEPARADO(A)   -0.046917   0.029051  -1.615 0.106328
SIT_VIVIENDA_1           0.046249   0.027338   1.692 0.090712 .
SIT_VIVIENDA_23          0.100104   0.017999   5.562 2.70e-08 ***
SIT_VIVIENDA_4           0.260688   0.140999   1.849 0.064488 .
UBIGEO_2_ZONA_CENTRO     0.140507   0.016123   8.714  < 2e-16 ***
UBIGEO_2_ZONA_NORTE      0.043309   0.016113   2.688 0.007195 **
EDAD40                   0.005765   0.002001   2.881 0.003963 **
EDAD50                   0.005388   0.001117   4.823 1.42e-06 ***
EDAD60                  -0.014521   0.001167 -12.445  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.055 on 29340 degrees of freedom
Multiple R-squared:  0.2543,    Adjusted R-squared:  0.2537
F-statistic: 416.9 on 24 and 29340 DF,  p-value: < 2.2e-16
```

Source: Based on information from ENAHO 2019
Elaboration: Own

- It can be seen that the variable "Mieperho - 3 members per household" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the basic expenses model 2 using multiple linear regression.
- It can be seen that the variable "Mieperho - 4 members per household" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the basic expenses model 2 using multiple linear regression.
- It can be seen that the variable "Mieperho - 5 members per household" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the basic expenses model 2 using multiple linear regression.

- It can be seen that the variable "Marital status - Coexistence" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the basic expenses 2 model using multiple linear regression.
- It can be seen that the variable “Marital status - Married” involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the basic expenses model 2 using multiple linear regression.
- It can be seen that the variable "Marital status - Widower" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the basic expenses model 2 using multiple linear regression.
- It can be seen that the variable "Marital status - Separated" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the basic expenses model 2 using multiple linear regression.
- It can be seen that the variable "Housing situation - Rented" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the basic expenses model 2 using multiple linear regression.
- It can be seen that the variable “Housing situation - Own, buying it in installments” involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the basic expenses model 2 using the multiple linear regression.

Graph 58. Parameters of the Household Expenses model

```
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.2983565  0.1171077  28.165  < 2e-16 ***
LOG_INGRESO_TOTAL        0.2344910  0.0050513  46.422  < 2e-16 ***
ESTRATO_1                0.5570514  0.0153079  36.390  < 2e-16 ***
ESTRATO_234              0.4934996  0.0114623  43.054  < 2e-16 ***
ESTRATO_5                0.3051065  0.0138577  22.017  < 2e-16 ***
MIEPERHO_1              -0.6847132  0.0202395 -33.831  < 2e-16 ***
MIEPERHO_2              -0.2874913  0.0169161 -16.995  < 2e-16 ***
MIEPERHO_3              -0.1351212  0.0162212  -8.330  < 2e-16 ***
MIEPERHO_4              -0.0835417  0.0157739  -5.296 1.19e-07 ***
MIEPERHO_5              -0.0500696  0.0172782  -2.898  0.00376 **
TIPO_EMPLEO_135          0.3958014  0.0986327   4.013 6.01e-05 ***
TIPO_EMPLEO_246          0.2089802  0.0981693   2.129  0.03328 *
EST_CIVIL_CONVIVIENTE    0.0046283  0.0208998   0.221  0.82474
EST_CIVIL_CASADO(A)      0.0608264  0.0209345   2.906  0.00367 **
EST_CIVIL_VIUDO(A)      -0.0141732  0.0228678  -0.620  0.53540
EST_CIVIL_DIVORCIADO(A)  0.2146527  0.0480697   4.465 8.02e-06 ***
EST_CIVIL_SEPARADO(A)    0.0116033  0.0206968   0.561  0.57505
SIT_VIVIENDA_1           0.0414386  0.0194767   2.128  0.03338 *
SIT_VIVIENDA_23          0.1193741  0.0128232   9.309  < 2e-16 ***
SIT_VIVIENDA_4           0.1330585  0.1004519   1.325  0.18531
UBIGEO_2_ZONA_CENTRO     0.1299892  0.0114868  11.316  < 2e-16 ***
UBIGEO_2_ZONA_NORTE      0.1811624  0.0114791  15.782  < 2e-16 ***
EDAD40                  -0.0068560  0.0014254  -4.810 1.52e-06 ***
EDAD50                   0.0020129  0.0007959   2.529  0.01144 *
EDAD60                  -0.0095914  0.0008313 -11.538  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7519 on 29340 degrees of freedom
Multiple R-squared:  0.3241,    Adjusted R-squared:  0.3235
F-statistic: 586.1 on 24 and 29340 DF,  p-value: < 2.2e-16
```

Source: Based on information from ENAHO 2019
Elaboration: Own

- It can be seen that the variable “Marital status - Cohabitant” involved in the model obtained a statistically non-significant parameter with a 5% level of significance for the household expenses model using multiple linear regression.
- It can be seen that the variable "Marital status - Widower" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the household expenses model using multiple linear regression.
- It can be seen that the variable “Marital status - Separated” involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the household expenses model using multiple linear regression.

- It can be seen that the variable "Housing situation - Own, buying it in installments" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the household expenses model using multiple linear regression.

Graph 59. Parameters of the Luxury Expenses model

```
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.3899053  0.1732307  25.341  < 2e-16 ***
LOG_INGRESO_TOTAL        0.2775766  0.0074720  37.149  < 2e-16 ***
ESTRATO_1                0.3528853  0.0226441  15.584  < 2e-16 ***
ESTRATO_234              0.5063232  0.0169556  29.862  < 2e-16 ***
ESTRATO_5                0.2488369  0.0204989  12.139  < 2e-16 ***
MIEPERHO_1              -2.3000726  0.0299391 -76.825  < 2e-16 ***
MIEPERHO_2              -1.5872486  0.0250230 -63.432  < 2e-16 ***
MIEPERHO_3              -0.7339856  0.0239951 -30.589  < 2e-16 ***
MIEPERHO_4              -0.3135000  0.0233335 -13.436  < 2e-16 ***
MIEPERHO_5              -0.1536255  0.0255587  -6.011 1.87e-09 ***
TIPO_EMPLEO_135          0.7313635  0.1459018   5.013 5.40e-07 ***
TIPO_EMPLEO_246          0.4465544  0.1452163   3.075 0.002106 **
EST_CIVIL_CONVIVIENTE   -0.3470518  0.0309159 -11.226  < 2e-16 ***
EST_CIVIL_CASADO(A)     -0.2602186  0.0309672  -8.403  < 2e-16 ***
EST_CIVIL_VIUDO(A)       0.0870196  0.0338271   2.572 0.010102 *
EST_CIVIL_DIVORCIADO(A)  0.2409513  0.0711068   3.389 0.000703 ***
EST_CIVIL_SEPARADO(A)   -0.0006847  0.0306156  -0.022 0.982157
SIT_VIVIENDA_1           0.0514985  0.0288108   1.787 0.073871 .
SIT_VIVIENDA_23          0.0257085  0.0189686   1.355 0.175326
SIT_VIVIENDA_4           0.1689536  0.1485928   1.137 0.255538
UBIGEO_2_ZONA_CENTRO     0.0201195  0.0169918   1.184 0.236395
UBIGEO_2_ZONA_NORTE     -0.0822490  0.0169803  -4.844 1.28e-06 ***
EDAD40                   0.0021145  0.0021086   1.003 0.315946
EDAD50                  -0.0237291  0.0011773 -20.156  < 2e-16 ***
EDAD60                  -0.0261670  0.0012296 -21.280  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.112 on 29340 degrees of freedom
Multiple R-squared:  0.4782,    Adjusted R-squared:  0.4778
F-statistic:  1120 on 24 and 29340 DF,  p-value: < 2.2e-16
```

Source: Based on information from ENAHO 2019
Elaboration: Own

- It can be seen that the variable "Marital status - Divorced" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the luxury expenses model using multiple linear regression.
- It can be observed that the variable "Marital status - Separated" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the luxury expenses model using multiple linear regression.

- It can be seen that the variable "Housing situation - Rented" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the luxury expenses model using multiple linear regression.
- It can be seen that the variable “Housing situation - Own” involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the luxury expenses model using multiple linear regression.
- It can be seen that the variable "Housing situation - Own, buying it in installments" involved in the model obtained a statistically non-significant parameter with a 5% level of significance, for the luxury expenses model using multiple linear regression.

In addition, the performance indicators for the different expenditure models are presented, summarized in the following tables.

Table 24. *Predictive capacity of each expenditure module model*

	Indicators	
Models	R^2	R^2 - Adjusted
Basic expenses 1	42,54%	42,50%
Basic expenses 2	25,43%	25,37%
Household expense	32,41%	32,35%
Luxury expenses	47,82%	47,78%

Source: Based on information from ENAHO 2019
Elaboration: Own

It can be concluded that the performance of the R^2 – Adjusted models, which varies between 25.43% and 47.82% for the different expense modules

4.6.3 Overestimates and underestimates for quantile regression and multiple linear regression

Now, we will review the underestimations and overestimations of the quantile regression and multiple linear regression models. Now, for both estimates to be comparative, the quantile regression will be estimated at the quantile 0.40 and 0.50 because they are values close to the estimate of the mean of multiple linear regression.

Table 25. *Overestimation and underestimation of multiple linear regression*

BASIC EXPENSE 1		BASIC EXPENSE 2		HOUSEHOLD EXPENSE		LUXURY EXPENSE	
(-1,-0.5]	2843	(-1,-0.5]	7541	(-1,-0.5]	5050	(-1,-0.5]	7343
(-0.5,-0.4]	2271	(-0.5,-0.4]	2200	(-0.5,-0.4]	2001	(-0.5,-0.4]	2026
(-0.4,-0.3]	2651	(-0.4,-0.3]	2015	(-0.4,-0.3]	1875	(-0.4,-0.3]	1926
(-0.3,-0.2]	2625	(-0.3,-0.2]	1707	(-0.3,-0.2]	1930	(-0.3,-0.2]	1608
(-0.2,-0.1]	2511	(-0.2,-0.1]	1461	(-0.2,-0.1]	1782	(-0.2,-0.1]	1507
(-0.1,0]	2262	(-0.1,0]	1317	(-0.1,0]	1720	(-0.1,0]	1331
(0,0.1]	2059	(0,0.1]	1118	(0,0.1]	1698	(0,0.1]	1136
(0.1,0.2]	1697	(0.1,0.2]	989	(0.1,0.2]	1453	(0.1,0.2]	1061
(0.2,0.3]	1482	(0.2,0.3]	839	(0.2,0.3]	1290	(0.2,0.3]	913
(0.3,0.4]	1272	(0.3,0.4]	703	(0.3,0.4]	1193	(0.3,0.4]	769
(0.4,0.5]	1121	(0.4,0.5]	719	(0.4,0.5]	1044	(0.4,0.5]	642
(0.5,1]	3447	(0.5,1]	2355	(0.5,1]	3718	(0.5,1]	2427
(1,30]	3124	(1,30]	6162	(1,30]	4590	(1,30]	6511
(30,86]	0	(30,86]	229	(30,86]	21	(30,86]	165
(+ -)	29.04%	(+ -)	16.64%	(+ -)	22.66%	(+ -)	17.15%

Source: Based on information from ENAHO 2019
Elaboration: Own

Now, it is observed that the overestimation and underestimation of the multiple linear regression are in + - 20% of the estimations are somewhat in the Basic Expense 1 where the data fits better.

Table 26. *Overestimation and underestimation of the quantile regression (Q40)*

BASIC EXPENSE 1		BASIC EXPENSE 2		HOUSEHOLD EXPENSE		LUXURY EXPENSE	
(-1,-0.5]	2306	(-1,-0.5]	4203	(-1,-0.5]	3025	(-1,-0.5]	9122
(-0.5,-0.4]	2284	(-0.5,-0.4]	2074	(-0.5,-0.4]	1985	(-0.5,-0.4]	946
(-0.4,-0.3]	2315	(-0.4,-0.3]	2030	(-0.4,-0.3]	2508	(-0.4,-0.3]	849
(-0.3,-0.2]	1892	(-0.3,-0.2]	2084	(-0.3,-0.2]	2536	(-0.3,-0.2]	850
(-0.2,-0.1]	1715	(-0.2,-0.1]	1948	(-0.2,-0.1]	2398	(-0.2,-0.1]	850
(-0.1,0]	1704	(-0.1,0]	1797	(-0.1,0]	2025	(-0.1,0]	833
(0,0.1]	1693	(0,0.1]	1667	(0,0.1]	1826	(0,0.1]	819

(0.1,0.2]	1726	(0.1,0.2]	1481	(0.1,0.2]	1653	(0.1,0.2]	829
(0.2,0.3]	1672	(0.2,0.3]	1334	(0.2,0.3]	1572	(0.2,0.3]	829
(0.3,0.4]	1744	(0.3,0.4]	1115	(0.3,0.4]	1416	(0.3,0.4]	782
(0.4,0.5]	1639	(0.4,0.5]	1038	(0.4,0.5]	1296	(0.4,0.5]	784
(0.5,1]	6024	(0.5,1]	3849	(0.5,1]	4786	(0.5,1]	3529
(1,30]	2651	(1,30]	4745	(1,30]	2339	(1,30]	8343
(30,86]	0	(30,86]	0	(30,86]	0	(30,86]	0
(+ -)	23.29%	(+ -)	23.47%	(+ -)	26.91%	(+ -)	11.34%

Source: Based on information from ENAHO 2019
Elaboration: Own

Table 27. Overestimation and underestimation of quantile regression (Q50)

BASIC EXPENSE 1		BASIC EXPENSE 2		HOUSEHOLD EXPENSE		LUXURY EXPENSE	
(-1,-0.5]	3302	(-1,-0.5]	4612	(-1,-0.5]	3398	(-1,-0.5]	9625
(-0.5,-0.4]	2452	(-0.5,-0.4]	2065	(-0.5,-0.4]	2244	(-0.5,-0.4]	1103
(-0.4,-0.3]	2193	(-0.4,-0.3]	2226	(-0.4,-0.3]	2708	(-0.4,-0.3]	1114
(-0.3,-0.2]	1879	(-0.3,-0.2]	2279	(-0.3,-0.2]	2700	(-0.3,-0.2]	1100
(-0.2,-0.1]	1815	(-0.2,-0.1]	2285	(-0.2,-0.1]	2455	(-0.2,-0.1]	1043
(-0.1,0]	1802	(-0.1,0]	1991	(-0.1,0]	2055	(-0.1,0]	1090
(0,0.1]	1845	(0,0.1]	1847	(0,0.1]	1869	(0,0.1]	1097
(0.1,0.2]	1788	(0.1,0.2]	1499	(0.1,0.2]	1746	(0.1,0.2]	973
(0.2,0.3]	1846	(0.2,0.3]	1328	(0.2,0.3]	1584	(0.2,0.3]	975
(0.3,0.4]	1777	(0.3,0.4]	1150	(0.3,0.4]	1448	(0.3,0.4]	922
(0.4,0.5]	1588	(0.4,0.5]	985	(0.4,0.5]	1251	(0.4,0.5]	909
(0.5,1]	5403	(0.5,1]	3686	(0.5,1]	4380	(0.5,1]	3575
(1,30]	1675	(1,30]	3412	(1,30]	1527	(1,30]	5839
(30,86]	0	(30,86]	0	(30,86]	0	(30,86]	0
(+ -)	24.69%	(+ -)	25.96%	(+ -)	27.67%	(+ -)	14.31%

Source: Based on information from ENAHO 2019
Elaboration: Own

Now, already estimating the underestimations and overestimations of both quantiles (Q40 and Q50), it is observed that the multiple linear regression fits better to the data in the Basic Expenses 1 and Luxury Expenses models since a greater amount of approximation is shown in the + - 20%, however, the quantile regression, the approximation at + - 20%, is higher in the Basic Expenses 2 and Household Expenses models for both quantiles (Q40 and Q50). Therefore, it is concluded that both models can be conveniently taken as prediction models the study defined by a researcher.

.

Conclusions

An empirical investigation has been carried out on the relationships between the predictive variables, which are the stratum to which the dwelling belongs, number of members per household, type of occupation of the head of the household, age of the head of the household, ubiquitous, housing situation, Marital status of the head of the household, income and the different expenditure modules, such as Basic Expenses1, Basic Expenses2, Household Expenses and Luxury Expenses for a sample of the population of Peru, based on data from the ENAHO survey (National household survey). Given the asymmetry of the predicted variables, the methodology has consisted of using quantile regression, which is more appropriate in cases of non-normality. Furthermore, this technique offers the possibility of quantifying the effect of the regressors at different points in the domain of the expenditure modules. The conclusions regarding the objectives set out in the report are as follows:

1) The results indicate that where the data best fits through quantile regression is in the quantile 0.90 to 0.95, observing the Pseudo R2 for the different expenditure modules mentioned, which is consistent with the theoretical framework developed in this research work.
2) The model obtained shows that the influence of the explanatory variables that are significant on the total monthly expenses for each module are the variable "Income", "Stratum" and "Number of members per household", as well as some variables vary considerably from one quantile to another, so that some of them are significant only in some quantiles, such as the variable “Stratum”, “Type of employment” and “Marital status”.
3) Thus, it is also observed that the variable “Housing situation” in the total monthly expenses in education, health, home and luxury, there is no significant difference between the effects estimated by the multiple linear regression and the effects estimated by the quantile regression

4) The underestimations and overestimates of the quantile regression of the approximation in the + - 20% fit better in the Basic Expenses 2 and Household Expenses models for both quantiles (Q40 and Q50), and the multiple linear regression better fits the estimates in the Basic Expenses 1 and Luxury Expenses models. Therefore, it is concluded that both models can be conveniently taken as prediction models the study defined by a researcher.

Therefore, the quantile regression offers a comprehensive strategy in all the behavior of the monthly expenses of education, home, health and luxury, which allows the researcher to pose the question of relationship between the response variable and the covariates in any quantile of the function conditional distribution.

Bibliography

1. A.Cameron and P.Triveldi,(2005), Macroeconometrics, Methods and applications, Cambridge University Press.
2. R.Koenker, (2005), Quantile Regression, Econometric Society Monographs.
3. C.M.Kuan, (2004), An Introduction to Quantile Regression, Institute of Economics, Academia Sinica, Taiwan.
4. B.Reyes (2011), Evolution of the population of the municipalities of Extremadura: Parametric and semi-parametric applications, Autonomous University of Madrid, Madrid.
5. R.Huiman (2016), Analysis of the quantile regression for the distribution of the total monthly income of the economically active population in Metropolitan Lima, Universidad Nacional Mayor de San Marcos, Lima.
6. M.Medina y O.Vicens (2011), Determinants of household electricity demand in Spain: an approximation through quantile regression, Department of Applied Economics, Spain.
7. Buchinsky (1994), Changes in the US wage structure between 1963 and 1987: an application with quantile regression, USA.
8. SELOWSKY.M.: "The effect of unemployment and growth on the profitability of educational investment: an application to Colombia" Planning and Development, 1,2, Bogotá, National Planning Department.
9. Nuñez, J; Ramírez, J.C.; Cuesta, Laura. (2005): "Determinants of poverty in Colombia ". University of the Andes. CEDE Documents, 2006.

Appendix

Graph 60. *Project development schedule*

Points	Definition	nov-20			dic-20					ene-21					feb-21			
		S2	S3	S4	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5	S1	S2	S3	S4
		09/11 al 13/11	16/11 al 20/11	23/11 al 30/11	01/12 al 05/12	08/12 al 12/12	15/12 al 19/12	22/12 al 26/12	29/12 al 31/12	01/01 al 03/01	04/01 al 08/01	11/01 al 15/01	18/01 al 22/01	25/01 al 29/01	01/02 al 05/02	08/02 al 12/02	15/02 al 19/02	22/02 al 26/02
1. introduction	Presentation of the Problem																	
	Formulation of the research problem and its objectives																	
	Justification, scope and limits of research																	
2. Theoretical Foundations	Investigative background																	
	Conceptual framework																	
3. Materials, methods and procedures	Research design, population and sample																	
4. Presentation and analysis of results	Database																	
	Descriptive analysis of variables																	
	Univariate Analysis																	
	Bivariate Analysis																	
	Discriminant capacity analysis																	
5. Conclusions																		
6. Appendix																		

Source: Based on information from ENAHO 2019
Elaboration: Own

CRISP-DM METHODOLOGY

CRISP-DM (Cross Industry Standard Process for Data Mining) provides a standardized description of the life cycle of a standard data analysis project, analogously to software engineering with development life cycle models of software. The CRISP-DM model covers the phases of a project, their respective tasks, and the relationships between these tasks. At this level of description, it is not possible to identify all relationships; Relationships could exist between any task depending on the objectives, the context, and the user's interest in the data.

The CRISP-DM methodology views the data analysis process as a professional project, thus establishing a much richer context that influences the modeling. This context takes into account the existence of a client that is not part of the development team, as well as the fact that the project not only does not finish once the ideal model is found (since it requires deployment and maintenance afterwards) Rather, it is related to other projects, and needs to be fully documented for other development teams to use and build on the knowledge.

The data mining project life cycle consists of six phases shown in the following figure.

Next, we will briefly describe each of the phases

Phase I. Business Understanding. Defining customer needs (understanding the business)

This initial phase focuses on understanding the project objectives. This knowledge of the data is then converted into the definition of a data mining problem and a preliminary plan designed to achieve the objectives.

Phase II. Data Understanding. Study and understanding of data

The data understanding phase begins with the initial data collection and continues with activities that allow you to become familiar with the data, identify quality problems, discover preliminary knowledge about the data, and / or discover interesting subsets to form hypotheses about hidden information.

Phase III. Data Preparation. Data analysis and feature selection

The data preparation phase covers all the activities necessary to build the final data set (the data that will be used in the modeling tools) from the initial raw data. Tasks include selecting tables, records, and attributes, as well as data transformation and cleansing for the modeling tools.

Phase IV. Modeling. Modeling

In this phase, the modeling techniques that are relevant to the problem (the more the better) are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same type of data mining problem. Some techniques have specific requirements on the shape of the data. Therefore, almost always any project ends up going back to the data preparation phase.

Phase V. Evaluation. Evaluation (obtaining results)

At this stage in the project, one or more models have been built that seem to achieve sufficient quality from a data analysis perspective. Before proceeding to the final deployment of the model, it is important to thoroughly evaluate it and review the steps taken to create it, compare the model obtained with the business objectives. A key objective is to determine if there are any important business issues that have not been sufficiently considered. At the end of this phase, a decision should be obtained on the application of the results of the data analysis process.

Phase VI. Deployment. Deployment (put into production)

Generally, creating the model is not the end of the project. Even if the objective of the model is to increase the knowledge of the data, the knowledge obtained will have to be organized and presented so that the client can use it. Depending on the requirements, the development phase can be as simple as the generation of a report or as complex as the periodic and perhaps automated performance of a data analysis process in the organization.

INTERNATIONAL STANDARDS

GUIDANCE ON STATISTICAL TECHNIQUES FOR THE NTP ISO-9001: 2001 STANDARD

Regression analysis:

1. What is it?

Regression analysis relates the behavior of a characteristic of interest (usually called the "response variable") with potentially causal factors (usually called "explanatory variables"). Such a relationship is specified by a model that can come from the field of science, economics, engineering, etc., or it can be derived empirically. The goal is to help understand the potential cause of variation in response, and to explain how each factor contributes to the variation. This is achieved by statistically relating the variation in the response variable with the variation in the explanatory variables, and obtaining the best fit, minimizing the deviations between the prediction and the actual response.

2. What is it used for?

Regression analysis allows the user to do the following:

- Test hypotheses regarding the influence of potential explanatory variables on the response and use this information to describe the estimated change in the response for a given change in the explanatory variable.
- Predict (at a declared level of confidence) the range of values within which the response is expected to lie, given the specific values for the explanatory variables;
- Estimate the direction and degree of association between a response variable and an explanatory variable (although such association does not imply causality). Such information could be used, for example, to determine the effect of changing a factor such as temperature on process performance, while other factors are held constant.

Profits:

Regression analysis can provide insight into the relationship between various factors and the response of interest, and such understanding can help guide decisions related to the process being studied and eventually improve it.

The insight produced by regression analysis comes from its ability to concisely describe the behavior of response data, compare different but related data subsets, and analyze potential cause-and-effect relationships.

When the relationships are well modeled, regression analysis can give an estimate of the relative magnitudes of the effect of the explanatory variables, as well as identify the relative importance of these variables on the outcome. This information is potentially of great value in controlling or improving the results of the process.

Regression analysis can also provide estimates of the magnitude and source of influence on the response caused by factors not measured or omitted from the analysis. This information can be used to improve the measurement system or the process.

Regression analysis can be used to predict the values of the response variable, for certain values of one or more explanatory variables; it can also be used to forecast the effect of changes in explanatory variables on an existing or predicted response. It can be useful to conduct such analyzes before investing time or money on a problem when the effectiveness of an action is unknown.

Limitations and Cautions.

When modeling a process, skill is required in specifying a suitable regression model (eg, linear, exponential, multivariate), and in using diagnostics to improve the model. The presence of omitted variables, measurement errors, and other sources of unexplained variations in the response can complicate modeling. The specific assumptions behind the regression model in question, and the characteristics of the available data, determine which estimation technique is appropriate in a regression analysis problem.

A problem that is sometimes encountered in the development of a regression model is the presence of data whose validity is questionable. The validity of such data should be investigated when possible, since the inclusion or omission of data from the analysis could influence the estimates of the model parameters, and thus the response.

Simplification of the model, minimizing the number of explanatory variables, is important when making the model. The inclusion of unnecessary variables can mask the influence of explanatory variables and reduce the precision of the prediction model.

However, the omission of an important explanatory variable can seriously limit the model and reduce the usefulness of the results.

Sampling:

1. What is it?

   Sampling is a systematic statistical method to obtain information about some characteristic of a population by studying a representative fraction (that is, sample) of the population. There are several sampling techniques that can be employed (such as simple random sampling, stratified sampling, systematic sampling, sequential sampling, skip-lot sampling, etc.), and the selection of techniques is determined by the purpose of the sampling and the conditions under which it will be carried out.

2. What is it for?

   The sampling can be loosely divided into two broad, non-exclusive areas: “acceptance sampling” and “survey sampling”.

Acceptance sampling deals with the decision making regarding whether to accept or not accept a "lot" (ie, a group of items) based on the result of a selected sample from that lot. A wide range of acceptance sampling plans are available to meet specific requirements or applications.

Survey sampling is used in enumerative or analytical studies to estimate the values of one or more characteristics in a population, or to estimate how those characteristics are distributed among the population. Survey sampling is often associated with polls where information is gathered from public opinions on an issue, such as in customer surveys. It can equally apply to data collection for other purposes, such as audits.

A specialized form of survey sampling is exploratory sampling, which is used in enumerative studies to obtain information on one or more characteristics of a population or a subset of a population. So is production sampling, which can be done to carry out, for example, a process capability analysis.

Another application is the sampling of bulk materials (eg minerals, liquids and gases) for which sampling plans have been developed.

Profits

A properly developed sampling plan offers time, cost and labor savings compared to a total population census or 100% inspection of a lot. Where product inspection involves destructive testing, sampling is the only practical way to obtain relevant information.

Sampling offers an inexpensive and timely way to obtain preliminary information regarding the value or distribution of a characteristic of interest in a population.

Limitations and cautions

In constructing a sampling plan, attention should be paid to decisions regarding sample size, sampling frequency, sample selection, the basis for subgroups, and various other aspects of sampling methodology.

Sampling requires that the sample be selected in an unbiased manner (that is, the sample is representative of the population from which it was drawn). Failure to do this will result in a poor estimate of the characteristics of the population. In the case of acceptance sampling, unrepresentative samples may result in the unnecessary rejection of lots of acceptable quality, or the improper acceptance of lots of unacceptable quality.

Even with unbiased samples, information derived from samples is subject to a certain degree of error. The magnitude of this error can be reduced by taking a larger sample size, but it cannot be eliminated. Depending on the specific question and the context of the sampling, the sample size required to achieve the desired level of confidence and precision may be too large to be of practical value.