Revista Colombiana de Estadística
Special Issue: Colombian and Brazil Graduate Programs
December 2025, volume 48, issue 3, pp. 459–478
http://doi.org/10.15446/rce.v48n3.123657

Labor Income Determinants in Colombia: An Approach based on Generalized Additive Models for Location, Scale and Shape (GAMLSS)

Determinantes del ingreso laboral en Colombia: una aproximación a partir de un modelo aditivo de localización, escala y forma (GAMLSS)

Sofía Gallego Ruiz[1,a], Annamaría Saavedra Ávila[1,b],
Mario E. Arrieta-Prieto[2,c]

[1] Master's Program in Statistics, Faculty of Sciences, Universidad Nacional de Colombia, Bogotá, Colombia

[2] Department of Statistics, Faculty of Sciences, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

Human capital theory posits a central hypothesis: there is a direct relationship between individuals' levels of education and their productivity, which in turn leads to higher earnings. To evaluate this hypothesis, the economic literature commonly estimates the Mincer equation using the classical linear regression model. In this paper, both the traditional approach and a GAMLSS model with the Dagum distribution are estimated for wage earners in private firms and the public sector in Colombia in 2024. The hypothesis that an individual's rate of return increases with years of education, and that it rises with years of experience up to a certain point in the life cycle before declining, is confirmed by both the linear regression and the GAMLSS model. Model selection is based on the Generalized Akaike Information Criterion (GAIC), and the results show that the GAMLSS provides a better fit than the linear regression model.

***Keywords*:** GAMLSS; Labor income; Linear model; Mincer equation.

[a] Master's student. E-mail: kgallego@unal.edu.co
[b] Master's student. E-mail: asaavedraa@unal.edu.co
[c] Ph.D. E-mail: mearrietap@unal.edu.co

Resumen

La teoría del capital humano plantea una hipótesis: existe una relación directa entre los niveles de educación de los individuos y sus niveles de productividad y, por tanto, devengarán ingresos más altos. Para evaluarla, la literatura económica estima la ecuación de Mincer a partir del planteamiento de un modelo de regresión lineal clásico. En el presente documento, se estima el enfoque tradicional y un modelo GAMLSS para la distribución Dagum para los trabajadores de empresas y del gobierno en Colombia en 2024. La hipótesis de que la tasa de ganancia de un individuo se incrementa con los años de educación y que la tasa de ganancia incrementa con los años de experiencia hasta cierto punto en el ciclo de la vida y, posteriormente, desciende, se confirman luego de estimar el modelo de regresión lineal y el GAMLSS. Para seleccionar el modelo, se calcula el Criterio de Información de Akaike Genralizado (AIC) y se concluye que el GAMLSS tiene una mejor bondad de ajuste frente al modelo de regresión lineal.

***Palabras clave*:** Ecuación de Mincer; GAMLSS; Ingreso laboral; Modelo clásico lineal.

1. Introduction

Over the past few decades, empirical studies have attempted to identify the determinants of labor income (Arias & Chávez, 2002; García et al., 2009; Zárate, 2003). Following the seminal work of Mincer (1974) and Becker (1975), it is acknowledged that schooling and labor market experience contribute to human capital formation. To capture the schooling-earning relationship and the experience-earning profiles, the natural logarithm of earnings is modeled as a linear function of years of education and a quadratic function of years of potential experience. The baseline hypothesis is that the return to a year of schooling is positive and that the return to a year of experience declines over the life cycle (Mincer, 1974). Studies of labor income determinants in Colombia follow Mincer's human capital earnings function, including labor market characteristics, socioeconomic differences, and the nonlinear relationship between labor income and education (Arias & Chávez, 2002; García et al., 2009).

In recent years, however, the empirical analysis of income distributions has increasingly incorporated flexible statistical frameworks capable of modeling not only the conditional mean of labor income but also its dispersion, skewness, and tail behavior. A growing body of work highlights the importance of understanding the full conditional distribution of income, rather than focusing solely on average effects, to better characterize inequality, heterogeneity in returns to education, and distributional responses to socio-economic covariates.

Arias & Chávez (2002) estimate the human capital earnings function for the cross-sectional distribution of individual earnings in Colombia for the years 1991, 1999, and 2000, using Maximum Likelihood techniques in the presence of sample selection bias (Heckman, 1979). The authors conclude that returns to high school and undergraduate education increased between 1990 and 1999, but lessened between 1999 and 2000, and that the return to undergraduate education is higher than that to high school education. However, when comparing across groups, women have lower returns to undergraduate education than men, while men have lower returns to high school education than women.

García et al. (2009) estimate the Mincer equation under the sample selection bias correction (Heckman, 1979; François Bourguignon & Gurgand, 2007) for salaried and self-employed workers in Colombia, using microdata from the 2007 Great Integrated Household Survey covering seven cities. The authors find that schooling and labor market experience have different effects on labor income, depending on the type of employment. While education has a positive effect on labor income, the semi-elasticity of income to elementary and undergraduate education is lower for salaried workers than for the self-employed. Age, as a proxy for labor market experience, has a positive effect at a decreasing rate for salaried workers, and a negative effect at an increasing rate for the self-employed.

Zárate (2003) examines changes in the returns to education and labor market experience at different points in the wage distribution for 1991–2000, using microdata from the National Household Survey. To account for the mean as a limited measure for explaining wage determinants given skill heterogeneity, Zárate (2003) employs quantile regression. The empirical results indicate that the patterns of changes in rates of return to schooling across 0.10, 0.25, 0.50, 0.75, and 0.90 quantiles are similar. Nevertheless, there are significant differences in magnitude, with returns to schooling being higher in the upper quantiles than in the lower ones. New entrants to the labor market (those with 5 years of experience) obtained higher returns across all five quantiles. Within this group, those with wages in the lower part of the distribution benefited from better returns during the study period. In contrast, the wage returns of experienced workers (those with 15 years of experience) were higher the greater their wage level.

Beyond classical and quantile-based approaches, recent literature in applied statistics and econometrics has emphasized the relevance of distributional regression frameworks, most notably the Generalized Additive Models for Location, Scale and Shape (GAMLSS); for modeling income and welfare-related indicators. These models allow the parameters governing the location, dispersion, and shape of the income distribution to depend on covariates, thereby capturing complex forms of heterogeneity. The flexibility of GAMLSS has made it increasingly popular in empirical work, including the analysis of regional income inequality using Bayesian structured additive distributional regression (Klein et al., 2015), the correction of inequality underestimation in cross-survey imputation (Betti et al., 2024), and the estimation of household economic indicators under small-area frameworks (Mori & Ferrante, 2025).

Furthermore, recent studies extend GAMLSS to mixed discrete-continuous settings (Hohberg et al., 2021), enabling richer multidimensional approaches to poverty and welfare assessments. In parallel, research has continued to refine income-distribution modeling within the GAMLSS framework, as illustrated by Carneosso et al. (2024), who benchmark new unconditional and quantile regression distributions against GAMLSS alternatives.

An additional motivation for adopting a distributional approach is the well-established suitability of Dagum-type distributions for modeling income data, given their ability to capture heavy right tails and provide closed-form expressions for inequality measures. Embedding Dagum or Generalized Beta Type II distributions within GAMLSS combines the economic interpretation of classical income models with the flexibility required to capture distributional heterogeneity across demographic and socioeconomic groups.

In this context, the main objective of this paper is to estimate both a classical linear regression model and a Generalized Additive Model for Location, Scale, and Shape (GAMLSS) for employees in private firms and the public sector in Colombia for the year 2024, using publicly available microdata (Departamento Administrativo Nacional de Estadística, 2024). The hypothesis that the rate of return to a year of schooling is positive and that the return to a year of labor market experience declines over the life cycle is confirmed by both methodologies. For model selection, the GAIC is computed, suggesting that the GAMLSS provides a better fit than the linear regression model.

The remainder is organized as follows. Section 2 describes Mincer's human capital earnings function and a general model to describe the personal distribution of income. Section 3 provides a brief description of the data set, the classical linear regression model, and the GAMLSS. The empirical results are presented in Section 4, Section 5 concludes.

2. Theoretical Framework

2.1. Mincer's Equation

On the basis of both theoretical and empirical arguments, Mincer (1974) proposed the human capital earnings function, modeling the natural logarithm of earnings as the sum of a linear function of years of education and a quadratic function of years of potential labor market experience (Lemieux, 2003)

$$\log y = \log y_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \epsilon, \tag{1}$$

where y denotes hourly income, y_0 is the income level of an individual without schooling nor labor market experience, x_1 is the number of years of education, x_2 is the number of years of potential labor market experience, defined as $\texttt{Age} - x_1 - 6$[1], and β_1 is the rate of return to education.

Since Equation (1) is derived from a human capital investment model and is considered a parsimonious model of earnings determination, the human capital earnings function has guided the empirical study of income determination. However, it is acknowledged that a gap exists between the human capital income function and the dataset available for its estimation (Lemieux, 2003).

[1] To compute potential experience, it is assumed that labor market insertion begins once the scholar cycle is finished (Arias et al., 2003).

2.2. Dagum Distribution

Dagum (1980) introduced the theoretical and empirical foundations of a probability distribution designed to describe the distribution of income. The distribution must be able to capture negative, zero, and positive incomes, without establishing a minimum value, and be expressed with three or four parameters that allow for economic interpretation. Furthermore, the functional form must guarantee the existence of an explicit solution to the Lorenz curve and the Gini concentration coefficient.

Due to their goodness of fit, functional form, and economic interpretation, the Gamma, log-normal, and Pareto distributions have guided the study of income distribution. Nonetheless, it is recognized that they do not adequately capture the left and/or right tails of the distribution, which are critical in the study of income inequality and in the design of public policy (Dagum, 1980).

To overcome these limitations, Dagum (1980) proposed an income distribution for both developed and developing countries that satisfies the empirical and theoretical foundations. Since then, the Dagum family has become one of the leading parametric models for income distribution analysis, particularly because of its flexibility in describing heavy right tails and its ability to provide analytically tractable expressions for inequality measures. Simulation-based comparisons confirm that the Dagum and Singh–Maddala distributions outperform classical families when modeling skewed income data with substantial tail mass (Kakamu, 2016). Moreover, the applicability of the Dagum distribution extends beyond income modeling. For instance, López-Rodríguez et al. (2019) show its suitability for modeling asymmetric and highly variable rainfall patterns, further highlighting its robustness for positively skewed phenomena.

Recent literature has expanded the Dagum framework in several directions. Spasova (2024) evaluates methods for reconstructing income distributions from grouped data and shows that Dagum-based models remain competitive when only quantile or interval information is available. More recently, Alghufily et al. (2025) proposed a multivariate modified Dagum distribution, opening new avenues for the joint modeling of related economic variables and extending the Dagum family to multivariate dependence structures. These developments strengthen the case for employing the Dagum distribution in modern econometric settings and distributional regression frameworks such as GAMLSS.

The random variable X with support equal to the positive real numbers follows a Dagum distribution with scale parameters $a > 0$ and $p > 0$, and shape parameter $b > 0$, that is, $X \mid a, p, b \sim \mathsf{Dagum}(a, p, b)$, if its probability density function is given by

$$f_X(x) = \frac{ap}{x}\left(\frac{\left(\frac{x}{b}\right)^{ap}}{\left(\left(\frac{x}{b}\right)^{a}+1\right)^{p+1}}\right)\cdot \mathcal{I}_{(0,\infty)}(x).$$

After a suitable reparameterization to center the focus on the mean, the dispersion, and the skewness parameters, a generalization of the Dagum distribution is later considered as the candidate for the distribution of the outcome, given its suitability to model income. In the context of GAMLSS, this reparameterization allows each distributional parameter, including inequality-relevant shape parameters, to depend on covariates, providing a flexible framework for capturing heterogeneity in income distributions across demographic and socioeconomic groups.

3. Methodology

3.1. Data

This study uses data from the 2024 Quality of Life Survey (Encuesta de Calidad de Vida in spanish), provided by the National Administrative Department of Statistics (DANE). The survey provides statistical information on the sociodemographic characteristics and the employment and housing conditions of the resident population in Colombia, among other aspects. Since the structure of labor income differs according to the type of employment (García et al., 2009), the target population of this study consists of workers or employees in private firms and government employees who reported a positive labor income.

The dataset contains 27,981 observations, of which 55.4% correspond to male records and 44.6% to female records. The behavior of income for the target population can be observed in Figure 1, disaggregated by biological sex. As shown in Figure 1, income is positively skewed, with a concentration around values close to the statutory monthly minimum wage (SMMLV) in 2024 of COP $1,462,000[2], where a higher peak is observed for men. Furthermore, the mean is higher than the median for both men and women.

In turn, after applying the logarithmic transformation (Figure 2), it becomes clearer that men exhibit a higher density at the main peak compared to women, as well as greater presence in the higher-income range.

[2] The SMMLV for 2024 was set at COP $1,300,000, plus a transportation allowance of COP $162,000, for a total of COP $1,462,000.

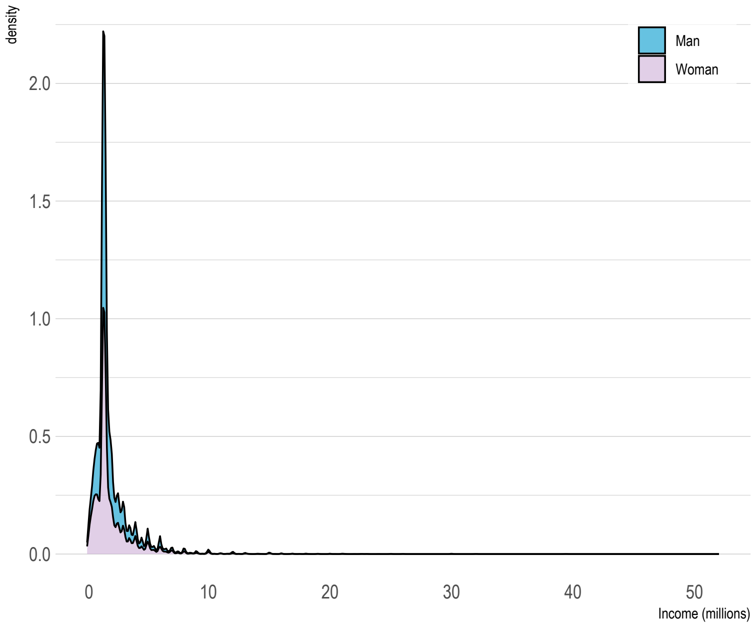

Figure 1: Distribution of the labor income by sex

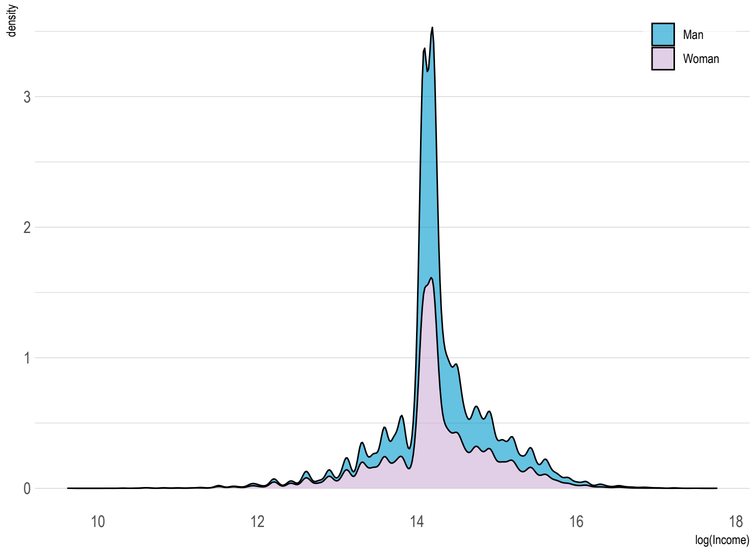

Figure 2: Logarithmic distribution of the labor income by sex

3.2. Linear Model

The linear model is presented below:

$$\begin{cases} \log \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \\ \quad \boldsymbol{\mu} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_1^2 + \beta_3 \mathbf{x}_2 + \sum_{j=4}^{16} \beta_j \mathbf{x}_j, \\ \quad \boldsymbol{\varepsilon} \sim \mathsf{MVN}_n \left(\mathbf{0}, \sigma^2 \mathbf{I}_n \right), \end{cases} \tag{2}$$

where $\mathbf{y}_{n\times 1}$ denotes labor income, $\mathbf{x}_1$ represents age, used as a *proxy* for potential experience, $\mathbf{x}_2$ is a dummy variable that equals 1 if the individual is female and 0 otherwise, $\mathbf{x}_j$, $j = 4, \dots, 16$, are dummy variables that equal 1 if the highest educational level attained by the individual is the j-th level and 0 otherwise, and $\boldsymbol{\varepsilon}$ is the error term. A detailed presentation of the different categories in educational level is presented in Table 1. Since a feature of the Colombian labor market is the bias in female wage remuneration (Arias & Chávez, 2002), gender is included as an explanatory variable. In addition, dummy variables are incorporated to capture differences across educational levels, which allows identifying potential differences in the returns to primary, secondary, and higher education. It is important to note that schooling in the dataset is categorical, which represents a notable difference from other studies that incorporate schooling as a quantitative variable measured in years of education.

3.3. GAMLSS Model

Rigby & Stasinopoulos (2005) assume that the response variable for each individual, y_i, $i = 1, \dots, n$, has a conditional probability function $f(y_i \mid \mu_i, \sigma_i, \nu_i, \tau_i)$[3], with μ_i as the location parameter, σ_i as the scale parameter, and ν_i and τ_i, if present, as the shape parameters (skewness and kurtosis). This approach allows modeling not only the mean but also other parameters of the response distribution as parametric and/or nonparametric functions of explanatory variables and/or random effects, and is specified as

[3] The only restriction on the distributional form in R is that the function $\log f(y_i \mid \mu_i, \sigma_i, \nu_i, \tau_i)$ and its first partial derivatives with respect to each parameter can be computed explicitly or numerically.

$$
\begin{aligned}
\mathbf{Y} &\overset{\text{ind}}{\sim} \mathsf{D}(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\nu},\boldsymbol{\tau}),\\
g_1(\boldsymbol{\mu}) &= \eta_1 = h_1(\mathbf{X}_1,\boldsymbol{\beta}_1) + \sum_{j=1}^{J_1} \mathbf{Z}_{j,1}\boldsymbol{\gamma}_{j,1},\\
g_2(\boldsymbol{\sigma}) &= \eta_2 = h_2(\mathbf{X}_2,\boldsymbol{\beta}_2) + \sum_{j=1}^{J_2} \mathbf{Z}_{j,2}\boldsymbol{\gamma}_{j,2},\\
g_3(\boldsymbol{\nu}) &= \eta_3 = h_3(\mathbf{X}_3,\boldsymbol{\beta}_3) + \sum_{j=1}^{J_3} \mathbf{Z}_{j,3}\boldsymbol{\gamma}_{j,3},\\
g_4(\boldsymbol{\tau}) &= \eta_4 = h_4(\mathbf{X}_4,\boldsymbol{\beta}_4) + \sum_{j=1}^{J_4} \mathbf{Z}_{j,4}\boldsymbol{\gamma}_{j,4},
\end{aligned}
$$

where D denotes the distribution of the response variable $\mathbf{Y}$, $g_k(\cdot)$ is a known monotonic link function that relates the distribution parameters to the explanatory variables; $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$, and $\boldsymbol{\tau}$ are vectors of dimension $n \times 1$; $\boldsymbol{\beta}_k^\top = (\beta_{1,k},\ldots,\beta_{J'_k,k})^\top$ is a parameter vector of dimension $J'_k \times 1$; $\mathbf{X}_k$ is a fixed and known design matrix of dimension $n \times J'_k$; $\mathbf{Z}_{j,k}$ is a design matrix of dimension $n \times q_{j,k}$; $\boldsymbol{\gamma}_{j,k}$ is a random vector of dimension $q_{j,k} \times 1$, such that $\boldsymbol{\gamma}_{j,k} \overset{\text{ind}}{\sim} \mathsf{MVN}_{q_{j,k}}\left(\mathbf{0}, [\mathbf{G}_{j,k}(\boldsymbol{\lambda}_{k,j})]^{-1}\right)$, where $\mathbf{G}_{j,k}(\boldsymbol{\lambda}_{k,j})$ is the (generalized) inverse of $\mathbf{G}_{j,k} = \mathbf{G}_{j,k}(\boldsymbol{\lambda}_{j,k})$, which may depend on a vector of hyperparameters, $\boldsymbol{\lambda}_{j,k}$; and h_i are linear and/or nonlinear functions of $\mathbf{X}_k$, $k = 1,\ldots,4$ (Rigby & Stasinopoulos, 2005).

Accordingly, the GAMLSS model for the case study is specified as follows:

$$
\begin{aligned}
\mathbf{Y} &\overset{\text{ind}}{\sim} \mathsf{GB}(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\nu}),\\
g_1(\boldsymbol{\mu}) &= \eta_1 = \beta_{0,1} + \beta_{1,1}\mathbf{x}_1 + \beta_{2,1}\mathbf{x}_1^2 + \beta_{3,1}\mathbf{x}_2 + \sum_{j=4}^{11} \beta_{j,1}\mathbf{x}_j,\\
g_2(\boldsymbol{\sigma}) &= \eta_2 = \beta_{0,2} + \beta_{1,2}\mathbf{x}_1 + \beta_{2,2}\mathbf{x}_1^2 + \beta_{3,2}\mathbf{x}_2 + \sum_{j=4}^{11} \beta_{j,2}\mathbf{x}_j,\\
g_3(\boldsymbol{\nu}) &= \eta_3 = \beta_{0,3} + \beta_{1,3}\mathbf{x}_1 + \beta_{2,3}\mathbf{x}_1^2 + \beta_{3,3}\mathbf{x}_2 + \sum_{j=4}^{11} \beta_{j,3}\mathbf{x}_j,
\end{aligned}
\tag{3}
$$

where GB denotes the Generalized Beta Type II distribution and incorporates the Dagum distribution as a special case. The link function for the parameters is the logarithmic function ($g_1(\cdot) = g_2(\cdot) = g_3(\cdot) = \log(\cdot)$). The parameters $\boldsymbol{\beta}_k$ are estimated by maximizing the log-likelihood function (Rigby & Stasinopoulos, 2005) given by

$$
l_p = \sum_{i=1}^{n} \log f(y_i \mid \mu_i, \sigma_i, \nu_i).
$$

4. Results

4.1. Linear Model

Table A1 reports the Ordinary Least Squares (OLS) estimates for the linear regression model. For interpretation purposes, Table 1 is presented, considering that the reference category for the variable `Sex` is female, and for `Schooling` it is 4 – Lower Secondary (6th–9th grade).

The results indicate that the effect of education on income is positive and increasing for both sexes; however, from upper secondary education onwards, men obtain higher incremental returns than women. The most pronounced difference is observed for men at university levels: in particular, an average increase of 88.0% in income is predicted for a man with incomplete tertiary education, holding other factors constant, compared to a woman with lower secondary education. By contrast, for a woman with the same schooling level, compared to a woman with lower secondary education, the model predicts an average increase of 51.7%, ceteris paribus. Regarding `Age`, since the squared term of the variable is included, the interpretation is affected: the effect on income is not constant, as it depends on the value of `Age`. Nevertheless, the general pattern indicates that as age increases, the positive impact diminishes due to the negative quadratic term.

With respect to the assumptions of the linear model, there are observations with high leverage, as well as outliers and influential cases. However, when performing a comparative exercise by removing the influential observations from the model estimation, their exclusion does not significantly alter the coefficients or the overall fit of the model. Likewise, the RESET test indicates that the model does not lack an unexplained quadratic or cubic pattern.

Table 1: Estimated effects by sex and educational level on the logarithm of wages.

Schooling Level	Category	Female (coef)	Male (total coef)	Change %
None	(1)	−0.364	−0.145	Female: −30.5%, Male: −13.5%
Preschool	(2)	−0.344	−0.125	Female: −29.1%, Male: −11.8%
Primary	(3)	−0.141	0.078	Female: −13.2%, Male: +8.1%
Upper Secondary	(5)	0.150	0.369	Female: +16.2%, Male: +44.6%
Technical, no degree	(6)	0.296	0.515	Female: +34.4%, Male: +67.3%
Technical, with degree	(7)	0.378	0.597	Female: +45.9%, Male: +81.7%
Technological, no degree	(8)	0.363	0.582	Female: +43.7%, Male: +78.9%
Technological, with degree	(9)	0.510	0.729	Female: +66.5%, Male: +107.3%
University, no degree	(10)	0.417	0.636	Female: +51.7%, Male: +89.0%
University, with degree	(11)	0.879	1.098	Female: +140.7%, Male: +199.9%
Postgraduate, no degree	(12)	1.295	1.514	Female: +263.3%, Male: +373.0%
Postgraduate, with degree	(13)	1.383	1.602	Female: +297.5%, Male: +395.9%

Note: Effects are in relation to women with Lower Secondary (reference category). Percentages are calculated as $(\exp(\text{coef}) - 1) \times 100$.

Regarding the assumption of multicollinearity, the variables `Sex` and `Schooling` do not exhibit this issue; however, when including `Age` and `Age`^{2}, multicollinearity is present. On the other hand, heteroskedasticity is detected, mainly due to the rejection of the null hypothesis in the Breusch–Pagan test. As a correction, robust inference was applied to construct a new variance–covariance matrix using an adjustment matrix of type HC3. The results of the model with corrected standard errors and confidence intervals can be found in Table A1.

The model does not satisfy the assumption of independence, since the Runs Test provides evidence that the signs of the residuals are not random with respect to each of the variables (`Age`, `Sex`, and `Schooling`). Finally, based on the Q–Q plot, the worm plot, and the Jarque–Bera test, it is concluded that the assumption of normality is not met and that the residuals exhibit heavy tails. Figure A1 summarizes the assumptions of *unexplained patterns*, *multicollinearity*, *heterogeneity*, and *normality* of the residuals.

To assess the predictive performance of the model, a 10-fold cross-validation was implemented. Figure 3 presents the Mean Squared Error (MSE) by fold, which shows reduced variability in prediction errors, suggesting that the model is stable across different data partitions (fold 5 has the smallest MSE, while fold 6 exhibits the largest MSE, around 0.283). The estimated average MSE was approximately 0.267 (gray line), indicating a good overall fit in terms of prediction on observations not used in training. In addition, using the Generalized Akaike Information Criterion (GAIC) with $\kappa = 2$, a value of 58,927 was obtained.

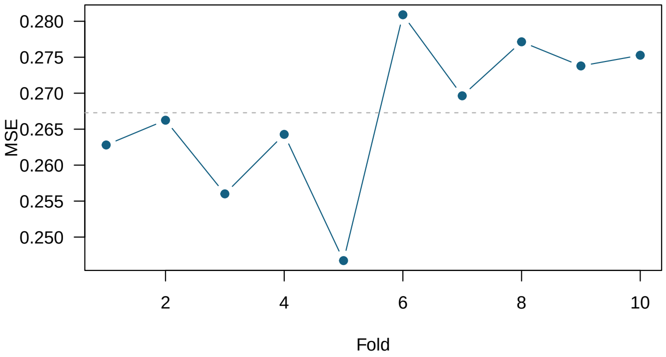

FIGURE 3: 10-fold cross-validation for linear model in Equation (2).

4.2. GAMLSS Model

Tables A2, A3, and A4 present the estimation results for the parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\nu}$ of the GAMLSS model specified in Equation (3). In addition, Figures A2, A3, and A4 illustrate the relationship between each predictor and the response variable. For $\boldsymbol{\mu}$, the interpretations are very similar to those of the classical linear model: there is an increasing relationship between `Age` and `Income`, although with a decreasing rate of change (as also evidenced by the negative effect of the quadratic term of `Age`). Likewise, controlling for other variables, men exhibit on average slightly higher incomes than women. Moreover, as the level of schooling increases, the contribution to expected income also rises.

When examining the effect of the explanatory variables on the dispersion parameter $\boldsymbol{\sigma}$, `Age` has a negative effect on dispersion, while the quadratic term has a positive effect. Note that the age extremes are associated with greater uncertainty in income. Regarding the levels of `Schooling`, a decreasing pattern in income dispersion is observed as the educational level increases. This may be explained by the fact that individuals with lower schooling levels exhibit greater dispersion—that is, more heterogeneous incomes—whereas university and postgraduate levels tend to show more homogeneous incomes. Finally, although the line corresponding to women lies slightly above that of men (Figure A3, row 2, column 1), suggesting greater income dispersion among women, the confidence bands overlap. This indicates that there is no conclusive evidence of a significant difference in income dispersion between men and women.

Finally, for the skewness parameter $\boldsymbol{\nu}$, it is noteworthy that income skewness tends to increase with `Age`; among older individuals, extreme values on the right-hand side of the distribution are expected. With respect to `Sex`, the red estimation line for men lies above that for women, indicating that men's income distribution exhibits greater positive skewness. In the case of `Schooling`, a clear pattern emerges: lower schooling levels (categories 1 to 4) display less skewness—that is, more symmetric incomes (in other words, less skewed toward high incomes). Starting from medium levels (categories 5 and above), positive skewness increases; however, it stabilizes at higher levels (categories 10 and above).

4.3. Model Comparison

The global deviance for a GAMLSS is defined as

$$\mathsf{D}_{\text{GAMLSS}} = -2\log \hat{L}_c,$$

where $\hat{L}_c$ is the maximized likelihood of the fitted model. To compare two models, the generalized Akaike information criterion can be computed as

$$\mathsf{GAIC}(\kappa) = -2\log \hat{L}_c + (\kappa \cdot \mathsf{df}),$$

where df denotes the effective degrees of freedom (the effective number of parameters) of the model and κ is the penalty per degree of freedom used. The model that yields the smallest value of $\mathsf{GAIC}(\kappa)$, for a given κ, is selected. Table 2 presents the GAIC for $\kappa = 2$ and $\kappa = \log n$. Since the GAIC is smaller for the GAMLSS model, it is concluded that the latter provides a better fit than the classical linear model.

Table 2: Goodness of fit of the proposed models.

Model	$\mathsf{GAIC}(\kappa = 2)$	$\mathsf{GAIC}(\kappa = \log n)$	Effective degreesof freedom
Classical linear model	58,927	59,059	16
GAMLSS	32,574	32,970	48

In the validation set, we compared a GAMLSS model with GB2 errors against a linear model with Normal errors. The GB2 specification attained a lower total and per-observation deviance than the Normal model (Table 3,) indicating a superior probabilistic fit, particularly in capturing skewness and tail behavior. It also achieved a lower MAE (669 276.2 vs. 710 735.3). The classic model delivered a slightly smaller $RMSE$ (a difference of approximately 0.8%), suggesting a few larger squared errors under GB2 without overturning its overall advantage. In sum, when the evaluation focuses on full predictive distributions or on L_1-type loss, the GAMLSS–GB2 model is preferable; the linear–Normal benchmark remains competitive only if squared–error loss is the exclusive operational criterion.

Table 3: Validation results on the training set.

Model	Deviance (total)	Deviance / obs.	RMSE	MAE
GAMLSS–GB2	331,181	29.6625	1,420,278	669,276
Linear–NO	347,876	31.1578	1,408,340	710,735

5. Conclusions, Limitations, and Future Work

This study analyzed the determinants of labor income in Colombia using both a classical linear regression and a Generalized Additive Model for Location, Scale, and Shape (GAMLSS) with a Dagum-type distribution. While both approaches confirm the human capital hypothesis: higher education and experience are associated with greater income, with diminishing returns over the life cycle, the GAMLSS framework offers a more comprehensive understanding of the income-generating process by modeling not only the mean but also the dispersion and asymmetry of income as functions of individual characteristics.

The estimates for the location parameter ($\hat{\mu}$) reaffirm that education has a strong positive association with expected income, while age exhibits the conventional concave pattern of earnings across the life cycle. The estimated scale equation ($\hat{\sigma}$) reveals that income dispersion decreases with higher education, indicating more homogeneous returns among highly educated individuals and greater variability among those with lower schooling levels. Finally, the estimated skewness equation ($\hat{\nu}$) highlights that income distributions become more asymmetric with age and education, reflecting a longer right tail particularly among men and more educated groups.

These findings demonstrate the flexibility and explanatory power of the GAMLSS model in capturing heterogeneity across multiple dimensions of the income distribution. Compared with the classical model, the GAMLSS achieves a substantially better fit and provides richer insights into inequality dynamics, as confirmed by the GAIC. Future work could extend this approach by incorporating additional labor market, demographic, and regional variables, or by exploring temporal extensions to study changes in income distribution over time.

[Received: September 2025 — Accepted: November 2025]

References

Alghufily, N., Sultan, K. S. & Radwan, H. M. M. (2025), 'Multivariate modified Dagum distribution and its applications', *Mathematics* **13**(10), 1620.

Arias, Y. & Chávez, A. (2002), Cálculo de la tasa interna de retorno de la educación en colombia, Technical Report Documento de Trabajo No. 2, Universidad Externado de Colombia, Bogotá.

Becker, G. S. (1975), 'Investment in human capital: Effects on earnings'.

Betti, G., Molini, V. & Mori, L. (2024), 'An attempt to correct the underestimation of inequality measures in cross-survey imputation through generalized additive models for location, scale and shape', *Socio-Economic Planning Sciences* **91**, 101784.

Carneosso, C. C., De Andrade, T. A. N. & Bisognin, C. (2024), 'New unconditional and quantile regression model Erf-Weibull: An alternative to gamma, gumbel and exponentiated exponential distributions', *Revista Colombiana de Estadística* **47**(2), 301–327.

Dagum, C. (1980), 'The generation and distribution of income, the Lorenz curve and the Gini ratio', *Économie Appliquée* **33**(2), 327–367.

Departamento Administrativo Nacional de Estadística (2024), 'National quality of life survey (ENCV 2024)'.

François Bourguignon, M. F. & Gurgand, M. (2007), 'Selection bias corrections based on the multinomial logit model: Monte carlo comparisons', *Journal of Economic Surveys* **21**(1), 174–205.

García, A., Guataquí, J. & Rodríguez, M. (2009), Estimaciones de los determinantes de los ingresos laborales en colombia, Technical Report Documentos de Trabajo No. 70, Universidad del Rosario, Bogotá.

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica* **47**(1), 153–161. http://www.jstor.org/stable/1912352

Hohberg, M., Donat, F., Marra, G. & Kneib, T. (2021), 'Beyond unidimensional poverty analysis using distributional copula models for mixed ordered–continuous outcomes', *Journal of the Royal Statistical Society: Series C* **70**(5), 1365–1390.

Kakamu, K. (2016), 'Simulation studies comparing Dagum and Singh–Maddala income distributions', *Computational Economics* **48**(4), 593–605.

Klein, N., Kneib, T., Lang, S. & Sohn, A. (2015), Bayesian structured additive distributional regression with an application to regional income inequality in germany. Working paper.

Lemieux, T. (2003), The Mincer equation thirty years after schooling, experience, and earnings, *in* S. Grossbard, ed., 'Jacob Mincer: A pioneer of modern labor economics', Springer, Boston, MA.

López-Rodríguez, F., García-Sanz-Calcedo, J., Moral-García, F. J. & García-Conde, A. J. (2019), 'Statistical study of rainfall control: The Dagum distribution and applicability to the southwest of spain', *Water* **11**(3), 453.

Mincer, J. A. (1974), The human capital earnings function, *in* 'Schooling, experience, and earnings', National Bureau of Economic Research, pp. 83–96.

Mori, L. & Ferrante, M. R. (2025), 'Small area estimation of household economic indicators under unit-level generalized additive models for location, scale and shape', *Journal of Survey Statistics and Methodology* **13**(1), 160–196.

Rigby, R. A. & Stasinopoulos, D. M. (2005), 'Generalized additive models for location, scale and shape', *Journal of the Royal Statistical Society: Series C* **54**, 507–554.

Spasova, T. (2024), 'Estimating income distributions from grouped data: A minimum quantile distance approach', *Computational Economics* **64**(4), 2079–2096.

Zárate, H. (2003), Cambios en la estructura salarial: Una historia desde la regresión cuantílica, Technical Report Borradores de Economía No. 245, Banco de la República, Bogotá.

Appendix A. Estimates of Linear Model by OLS and Linear Model Validation

Table A1 presents the estimated coefficients by ordinary least squares. Besides, Figure A1 summarizes the assessment of the classic assumptions (unexplained patterns, multicollinearity, heteroscedasticity, and normality of residuals).

Table A1: Estimates of Equation 2 by OLS.

Explanatory variable	Estimate	Standard error	Lower limit 95%	Upper limit 95%
Intercept	13.028***	0.034	12.962	13.094
Sex (Male = 1)	0.219***	0.007	0.205	0.231
Age	0.033***	0.002	0.030	0.037
Age^2	−0.000***	0.000	−0.000	−0.000
None	−0.364***	0.035	−0.432	−0.297
Preschool	−0.344*	0.163	−0.664	−0.025
Elementary	−0.141***	0.014	−0.168	−0.113
Middle and High School	0.150***	0.011	0.129	0.171
Technical without diploma	0.296***	0.032	0.235	0.357
Technical with diploma	0.378***	0.012	0.353	0.403
Technological without diploma	0.363***	0.055	0.255	0.471
Technological with diploma	0.510***	0.016	0.479	0.541
Bachelor without diploma	0.417***	0.034	0.350	0.485
Bachelor with diploma	0.879***	0.014	0.852	0.906
Postgraduate without diploma	1.295***	0.077	1.144	1.446
Postgraduate with diploma	1.383***	0.017	1.350	1.416
n	27,981			
R^2	0.4136			
Adjusted R^2	0.4132			
F-Statistic	1,315.0*** (gl = 15; 27,965)			

Note: Statistical significance at the 10/5/1% significance level indicated with */**/***, respectively.

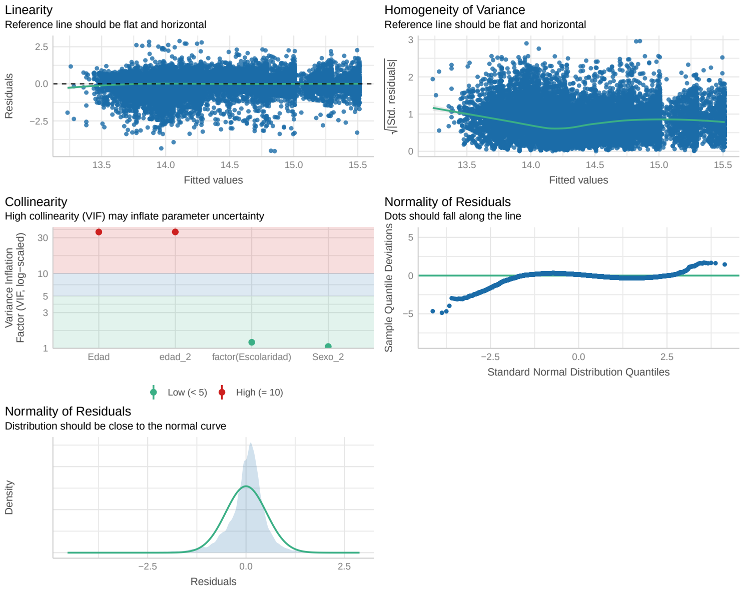

Figure A1: Linear model validation.

Appendix B. Estimates for GAMLSS

Table A2: Estimates of μ equation for GAMLSS.

Explanatory variable	Estimate	Standard error	Lower limit 95%	Upper limit 95%
Intercept	14.183***	0.021	14.141	14.224
Sex (Male = 1)	0.007	0.004	−0.002	0.016
Age	0.003**	0.001	0.001	0.005
Age^2	−0.000	0.000	−0.000	0.000
None	−0.083***	0.022	−0.126	−0.041
Preschool	−0.088	0.116	−0.316	0.140
Elementary	−0.001	0.008	−0.017	0.015
Middle School	0.047***	0.007	0.033	0.061
Technical without diploma	0.066***	0.022	0.023	0.109
Technical with diploma	0.102***	0.009	0.085	0.118
Technological without diploma	0.066	0.046	−0.024	0.157
Technological with diploma	0.136***	0.012	0.113	0.160
Bachelor without diploma	0.158***	0.028	0.104	0.211
Bachelor with diploma	0.657***	0.010	0.637	0.676
Postgraduate without diploma	0.972***	0.070	0.834	1.110
Postgraduate with diploma	1.160***	0.014	1.133	1.188
n		27,981		
Link function		log(·)		

Note: Statistical significance at the 10/5/1% sinificance level indicated with */**/***, respectively.

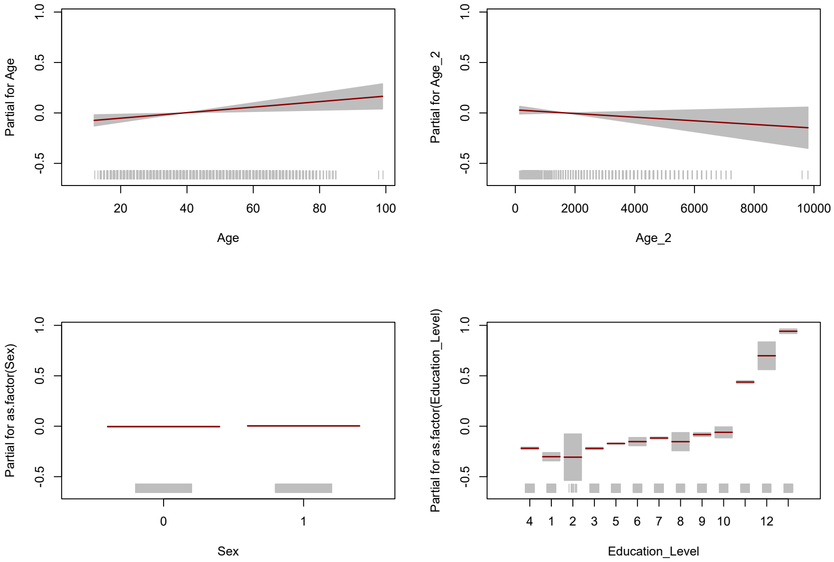

FIGURE A2: Partial terms chart for μ.

TABLE A3: Estimates of σ equation for GAMLSS.

Explanatory variable	Estimate	Standard error	Lower limit 95%	Upper limit 95%
Intercept	2.992***	0.037	2.920	3.063
Sex (Male = 1)	−0.256***	0.009	−0.274	−0.238
Age	−0.025***	0.002	−0.029	−0.021
Age^2	0.000***	0.000	0.000	0.000
None	0.306***	0.035	0.238	0.373
Preschool	0.045	0.181	−0.310	0.401
Elementary	0.210***	0.017	0.177	0.243
Middle School	−0.131***	0.015	−0.161	−0.100
Technical without diploma	−0.207***	0.055	−0.314	−0.099
Technical with diploma	−0.473***	0.018	−0.508	−0.438
Technological without diploma	−0.600***	0.093	−0.783	−0.417
Technological with diploma	−0.674***	0.024	−0.721	−0.627
Bachelor without diploma	−0.789***	0.047	−0.881	−0.697
Bachelor with diploma	−0.864***	0.018	−0.900	−0.829
Postgraduate without diploma	−0.949***	0.102	−1.148	−0.750
Postgraduate with diploma	−0.804***	0.023	−0.849	−0.759
n		27,981		
Link function		log(·)		

Note: Statistical significance at the 10/5/1% significance level indicated with */**/***, respectively.

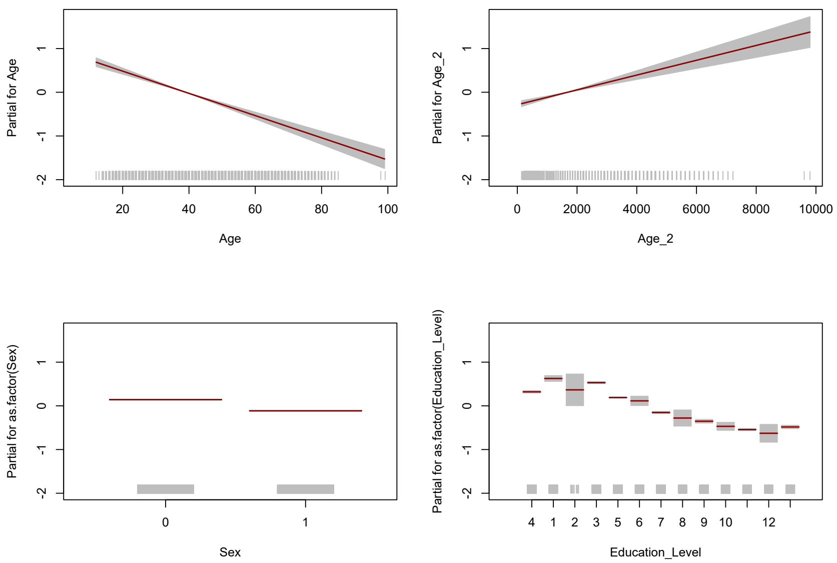

FIGURE A3: Partial terms chart for σ.

TABLE A4: Estimates of ν equation for GAMLSS.

Explanatory variable	Estimate	Standard error	Lower limit 95%	Upper limit 95%
Intercept	−3.510***	0.060	−3.628	−3.392
Sex (Male = 1)	0.612***	0.012	0.588	0.636
Age	0.084***	0.003	0.078	0.090
Age^2	−0.001***	0.000	−0.001	−0.001
None	−0.767***	0.061	−0.886	−0.648
Preschool	−0.041	0.031	−0.102	0.020
Elementary	−0.438***	0.027	−0.490	−0.386
Middle School	0.319***	0.022	0.275	0.363
Technical without diploma	0.709***	0.073	0.566	0.852
Technical with diploma	0.921***	0.026	0.870	0.972
Technological without diploma	1.012***	0.027	0.959	1.066
Technological with diploma	1.213***	0.033	1.149	1.278
Bachelor without diploma	0.946***	0.060	0.828	1.064
Bachelor with diploma	1.143***	0.127	0.894	1.393
Postgraduate without diploma	0.868***	0.031	0.808	0.928
n		27,981		
Link function		log(·)		

Note: Statistical significance at the 10/5/1% significance level indicated with */**/***, respectively.

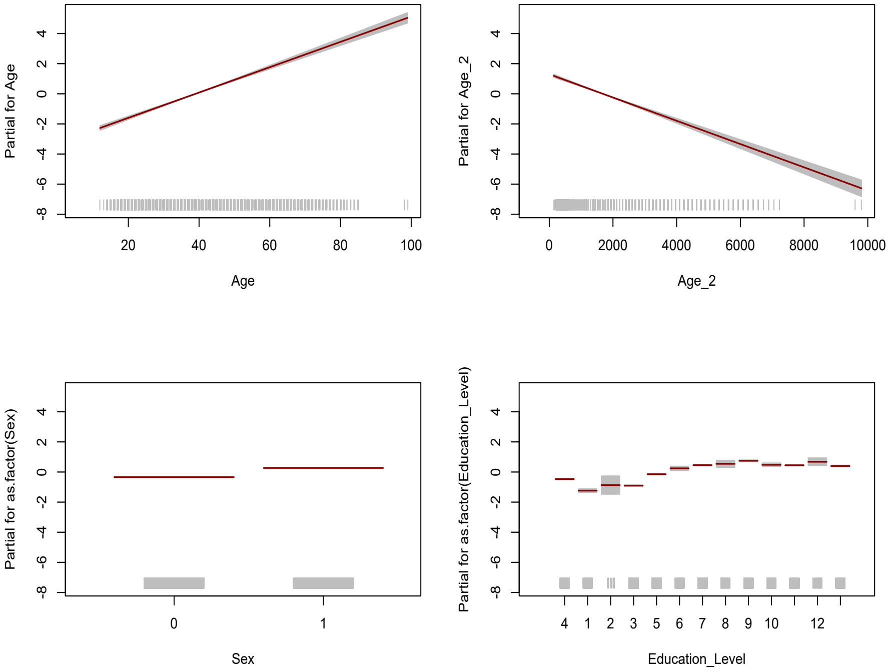

Figure A4: Partial terms chart for ν.