Global Journal of Economic and Finance Research

e-ISSN: 3050-5348
p-ISSN: 3050-533X

Vol. 01(04): 53-57, September 2024

Home Page: https://gjefr.com/index.php/gjefr

Relationship Between Starting Salary and Categorical Variables

Luzolo Domingos Sanches-António
Department of Tourism, School of Hotel and Tourism, Agostinho Neto University, Luanda, Angola

KEYWORDS: Starting salary, Sex, Educational level, and Econometric model.
JEL Codes: C51, J31.

Corresponding Author:
Luzolo Domingos Sanches-António

Publication Date: 05 Sept.-2024
DOI: 10.55677/GJEFR/01-2024-Vol01E4

ABSTRACT

The present study aims to highlight the different phases of the econometric model construction process, namely: specification, data characterization, estimation, diagnosis, and interpretation of the estimated parameters, based on the analysis of the relationship between a discrete variable, the starting salary, and two categorical variables, sex and education level. Assuming that salary variations are expressed in percentages and that β is the coefficient that, also in percentage terms, explains the variations that occur per unit in the independent variables and that demonstrate the dependent variable, the logarithms of the salary were calculated initially, considering the magnitude of the values in which salaries are expressed, which are normally large. The asymptosis of the sample allowed us to obtain a normal distribution for Ln salary observations, guaranteeing the assumption of linearity between the variables analyzed, namely between starting salary, sex, and educational level.

1. INTRODUCTION

Although exogenous job satisfaction models can explain not only the performance of a given worker but also their perception of their work Shevchuk & Melnikova (2020), specific variables, which are not always easy to measure from a merely quantitative perspective, have been widely used in econometric studies to explain certain patterns in the behavior of quantitative variables of interest, such as the initial salary, which, according to Sanches-António, et al. (2024), is frequently found in the literature related to the study of the effect of subjective notions such as sex, race, place of residence, level of education and others, on the income of a given individual.

Thus, the possible relationship to be established between the behavior of certain variables, and how this same behavior affects the behavior of other variables, allows, on the one hand, to explain the type of relationship that exists between them (linearity or non-linearity) and, on the other hand, the strength with which some can explain the behavior of others (correlation), therefore, regression models can be seen as a computational device to estimate differences between a treated group and a control group (Chein, 2019, p. 9).

There are also authors such as Lim (2022), who reinforce the idea that multiple linear regression is a time series regression, which explains the performance of the explained variable over time, using characteristics associated with each explanatory variable. Therefore, the present study aims to highlight the different stages of the process of constructing the econometric model presented, namely: specification, data characterization, estimation, diagnosis, and interpretation of the estimated parameters, based on the analysis of the relationship between a discrete variable (initial salary) and two categorical variables (sex and level of education).

From this perspective and in light of the above, a two-stage analysis of the relationships between the initial salary and the level of education will be carried out, followed by the study of the relationship between the initial salary and sex, that is, a simple regression, for each proposed relationship, aiming to capture the individual effect of each variable on the variable that is intended to be explained.

In the second phase or stage, the relationship between the initial salary, the level of education, and gender will be analyzed, that is a multiple regression, where the aim will be to capture the combined effect between the explanatory variables and the explained variable.

2. PROCEDURES ADOPTED

Given the contextualization carried out and the availability of data obtained from: https://www.ibm.com/docs/da/spss-statistics/beta?topic=system-sample-files/employee.data.sav which is part of the experimental database of the SPSS *software*, the following procedures were adopted:

Starting from the fact that the values of variations in wages (increases or decreases) are expressed in percentages and that β is the coefficient that, also in percentage terms) explains the variations that occur per unit in the independent variables and that explain the dependent variable, the salary logarithms were calculated. This procedure was adopted since logarithmized models allow greater specification of models with variables given in percentages.

According to Goos & Jones (2011), a good experimental design allows the precise estimation of one or more unknown variables of interest, which is why in the present work the variables studied were three, with the main interest being in the present study. In this relationship, categorical variables can affect the starting salary. From this perspective, according to Murphy & Topel (2002), the estimation of models that contain unobservable, although estimable, variables, such as categorical ones, is now common in several areas of applied econometrics, so in the present study, the different phases of The process of building the final model, namely: specification, data characterization, estimation, diagnosis, and interpretation, will be presented in three parts considering the two initial models, and finally, the multiple regression model, as already mentioned.

3. SPECIFICATION OF THE ECONOMETRIC MODEL

Table I. Descriptive statistics

	Average	Deviation error	N
Ln Wo	$9.6694	$0.35284	474
Ln educ	2.5772	.22909	474
dummy for sex	.54	.499	474

Source: By the author from (SPSS, 2024)

Dependent variable: Salary Ln (Ln W_0)

Independent variables: Ln of educational level (Ln educ) and Sex_dummy.

Admitting a multiple regression analysis we have:

$$Ln\ W_0 = \beta_0 + \beta_1 * Ln\ educ + \beta_2 * Sex_{dummy} + \mu$$

The model's basic assumptions include the linearity and randomness of the residues.

Table II. Summary of model[b]

Model	R	R square	Adjusted R-squared	Standard error of the estimate
1	.721[a]	.520	.518	$0.24509

a. Predictors: (Constant), dummy for sex, Ln educ

b. Dependent variable: ln Wo

Source: By the author from (SPSS, 2024)

R^2: 52% of the variability in salary Ln is explained by educational level and sex;

$R^2_{Adjusted}$: Values of R^2 and $R^2_{Adjusted}$, are more suitable for comparing models with different numbers of independent variables, present very close values (almost equal) indicating that the independent variables practically explain the entire salary Ln, that is, they are highly relevant to the model.

The standard deviation of the regression: is 0.26, a value that tends towards 0, indicating a considerable level of adjustment between the variables under analysis.

Table III. ANOVA[a]

Model		Sum of squares	df	Medium square	Z	Sig.
1	Regression	30,595	2	15,298	254,674	,000[b]
	Residue	28,292	471	.060		
	Total	58,887	473			

a. Dependent variable: ln Wo

b. Predictors: (Constant), dummy for sex, Ln educ

Source: By the author from (SPSS, 2024)

A. Hypotheses:

This table presents the global test of the model, that is, the explained variability, from which hypotheses will be tested.

H_0: No independent variable explains the behavior of Ln Salary;

H_0: $R^2$of the population is equal to zero.

H_1: At least one independent variable explains Ln Salary, that is, $\beta \neq 0$.

Starting from the assumption of $\alpha = 1\%$ we have that:

$\rho < 0.000 < \alpha = 5\%$ -> The null hypothesis is rejected, meaning that it is assumed that at least one variable explains Ln Salary. The result is reasonable enough to reject it H_0, despite the significance level being less than 0%.

Table IV. Coefficients[a]

Model		Unstandardized coefficients		Standardized coefficients	t	Sig.
		B	Error	Beta		
1	(Constant)	7,556	.131		57,745	,000
	Ln educ	.762	.052	.495	14,660	,000
	dummy for sex	.274	.024	.388	11,489	,000

a. Dependent variable: ln Wo

Source: By the author from (SPSS, 2024)

B. Non-standard betas:

$$Ln\ \widehat{W_0} = 7{,}556 + 0{,}762\ (Ln\ educ + 0{,}274)\ Sex_{dummy}$$

Constant: 7,6 -> value of the dependent variable when the independent variables take on a value of zero -> $\beta_2 * \text{Sex}_{\text{dummy}}$.

Since sex is a nominal variable, the type of model in question is log-lin -> *ceteris paribus* (keeping the educational level constant).

On average, men earn 0.274% more than women, although the value is not significant.

Educational level: log-lin -> *ceteris paribus* (comparing salaries between individuals of the same sex).

It is also possible to verify that a 1% variation in educational level has an impact of 0.77% on salary.

C. Standardized coefficients:

The educational level variable has a relative importance almost twice as much (21) as the sex variable, considering its absolute values, that is, $\frac{0{,}495}{0{,}388} = 20{,}625$.

D. Hypothesis testing:

Constant: $\beta \neq 0$

$\rho < 0.000 < \alpha = 5\%$ -> decision: reject H_0, i.e., the constant must be introduced into the model.

Sex and Ln Salary:

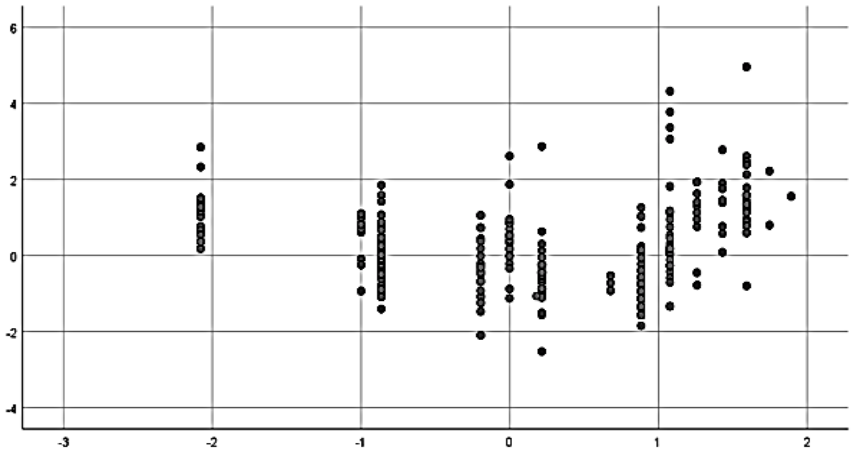
In both cases, $\rho < 0.000 < \alpha = 5\%$ -> both sex and educational level explain Ln Salary, i.e., it is rejected H_0. This demonstrates that educational level has a great explanatory capacity about Ln Salary.

E. Diagnosis:

Graph I. Scatterplot

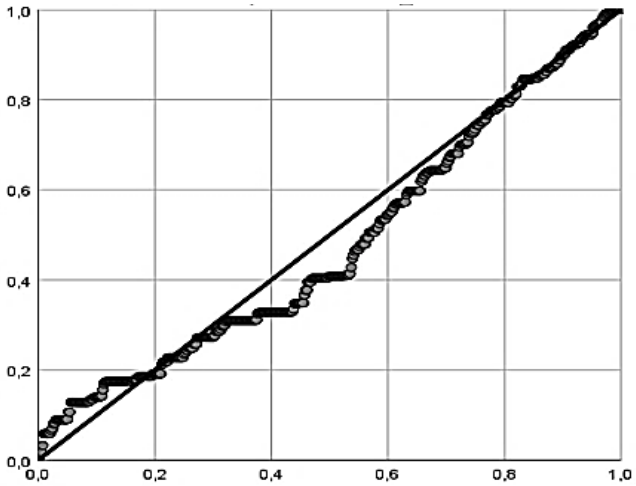
Dependent variable Ln Wo

Standardized predicted value regression



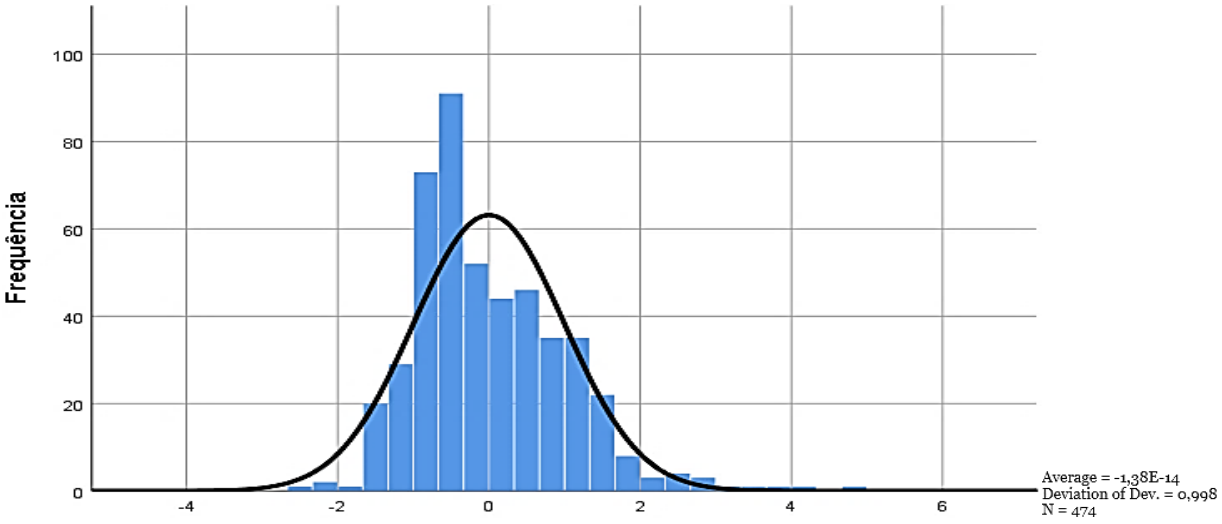
Source: By the author from (SPSS, 2024)

Graph II. Normal PP graph of standardized residual regression
Dependent variable Ln Wo
Cumulative probability



Source: By the author from (SPSS, 2024)

Graph III. Histogram
Dependent variable Ln Wo
Standardized residual regression



Source: By the author from (SPSS, 2024)

The set of observations is grouped very uniformly around the line originating at zero, that is, it presents a very balanced number of observations both above and below the regression line. Thus, the sample's linearity, constant variance, and normality are admitted considering its size, which is large, and asymmetry and kurtosis coefficients provided by the frequency distribution table.

Table V. Collinearity diagnosis[a]

Model	Dimension	Eigen value	Condition Index	Variance proportions (Constant)	Ln educ	dummy for sex
1	1	2,665	1,000	.00	.00	.04
	2	.332	2,834	.00	.00	.88
	3	.004	27,166	1.00	1.00	.08

a. Dependent variable: ln Wo

Source: By the author from (SPSS, 2024)

Despite the existence of a correlation between Ln educ and the constant (in the same dimension), resulting in a logarithm that compressed the salary and educational level values, there is no evidence of multicollinearity since the condition index is lower than 30. Equally important is the fact that the correlation does not occur between the independent variables.

Table VI. Waste statistics[a]

	Minimum	Maximum	Average	Deviation Error	N
Predicted value	$9.1406	$10.1506	$9.6694	$0.25433	474
Predicted value error	-2.079	1,892	,000	1,000	474
Standard error of predicted value	.015	.033	.019	.004	474
Adjusted predicted value	$9.1318	$10.1464	$9.6691	$0.25451	474
Residue	-$0.61913	$1.21518	$0.00000	$0.24457	474
Residual error	-2.526	4,958	,000	.998	474
Waste Waste	-2.533	4,978	.001	1,001	474
Of study.	-$0.62226	$1.22487	$0.00034	$0.24622	474
Study Waste.	-2.547	5,109	.002	1.006	474
Mahal. distance	.836	7,644	1,996	1,634	474
Distance from Cook	,000	.066	.002	.005	474
Centered leverage point value	.002	.016	.004	.003	474

a. Dependent variable: ln Wo

Source: By the author from (SPSS, 2024)

There are no extreme observations (outliers) since the observation with the greatest Cook distance has a value less than 1, that is, 0.066.

4. CONCLUSIONS

Considering the results of the diagnosis carried out, it can be concluded that the asymptosis of the sample allowed obtaining a normal distribution for the Ln salary observations, guaranteeing the assumption of linearity between the variables analyzed, namely between the starting salary, sex, and educational level.

As the ρ value is greater than 0.05, there is no indication of a statistically significant relationship between the variables at the 95% confidence level, which eliminates the need to remove variables from the specified model and reinforces the existence of a strong correlation between the model variables.

The study's objective was fulfilled, as the different phases of the final model construction process, namely: specification, data characterization, estimation, diagnosis, and interpretation, were fulfilled.

The assumptions underlying the regression models, namely the linearity of the function, as well as the non-correlation of the model variables with the error terms, cannot be fully assured, considering the reduced number of explanatory variables included in the model presented, which suggests that in future studies, variables such as race or ethnic origin can be considered.

REFERENCES

1. Chein, F. (2019). *Introdução aos Modelos de Regressão Linear: Um passo Inicial para Compreensão da Econometria como uma Ferramenta de Avaliação de Políticas Públicas.* Brasília - DF: Enap.
2. Goos, P., & Jones, B. (2011). *Optimal Design of Experiments: A Case Study Approach.* Rotterdam: Wiley & Sons.
3. Lim, K. G. (2022). *Theory and Econometrics of Financial Asset Pricing.* Lithuania: Walter de Gruyter GmbH & Co KG. doi:10.1515/9783110673951-201
4. Murphy, K. M., & Topel, R. H. (2002). stimation and Inference in Two-Step Econometric Models. *Journal of Business & Economic Statistics*, *20*(1), 88-97. doi:10.1198/073500102753410417
5. Sanches-António, L. D., Sanjimbi, A. S., Silvano, A. I., Lili, J. F., Sassimba, J. A., Sachivango, S. M., & da Silva Pascoal, A. (2024). O estado civil e a raça como determinantes do salário em Boston (EUA). *Brazilian Journal of Business*, *6*(2), 1-15. doi:10.34140/bjbv6n2-012
6. Shevchuk, I. A., & Melnikova, T. B. (2020). Exogenous model of job satisfaction. *J. Sib. Fed. Univ. Humanit. Soc. Sci.*, *13*(5), 818-830. doi:10.17516/1997-1370-0529
7. SPSS, S. (11 de 01 de 2024). *www.ibm.com*. Obtido de https://www.ibm.com/docs/da/spss-statistics/beta?topic=system-sample-files/employee.data.sav