engineering proceedings

MDPI

Proceeding Paper

Modeling a Set of Variables with Different Attributes on a Quantitative Dependent Variable: An Application of Dichotomous Variables †

Gerardo Covarrubias * and Xuedong Liu *

Faculty of Higher Studies Aragon, National Autonomous University of Mexico, Netzahualcóyotl 57000, Mexico
* Correspondence: jgcovarrubiasl@economia.unam.mx (G.C.); xdong@comunidad.unam.mx (X.L.)
† Presented at the 10th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 15–17 July 2024.

Abstract: This study outlines the methodology employed to model the relationship among a set of dichotomous variables, which represent attributes, on a nominal scale. The objective is to elucidate their influence on a quantitative dependent variable measured on a ratio scale. This approach allows for the quantification of the impact of these attributes and their significance in shaping the behavior of the entity possessing them. The resolution method employed for the estimation is ordinary least squares. However, it is crucial to note that interpreting the estimators in the resulting model requires a nuanced perspective, distinguishing it from the conventional interpretation of slope or rate of change in a classic model. To clarify, these estimators align with the average behavior of the dependent variable concerning binary characteristics, and the outcomes are consistent with the analysis of variance.

Keywords: dichotomous variables; attributes; quantitative dependent variable; modeling

Citation: Covarrubias, G.; Liu, X. Modeling a Set of Variables with Different Attributes on a Quantitative Dependent Variable: An Application of Dichotomous Variables. *Eng. Proc.* **2024**, *68*, 9. https://doi.org/10.3390/engproc2024068009

Academic Editors: Olga Valenzuela, Fernando Rojas, Luis Javier Herrera, Hector Pomares and Ignacio Rojas

Published: 1 July 2024

1. Introduction

The significance of the developed methodology lies in quantifying the impact of a set of dichotomous variables that elucidate the behavior of a quantitative variable on a ratio scale. This methodology holds relevance not only in the realm of social sciences but also across various empirical study domains.

When referring to the attributes of a subject or entity, these dichotomous or binary variables assume values of 0 and 1, denoting the presence of the respective attribute. The methodology assesses the statistical significance of this variable set and compares the expected or average values of the dependent variable in the presence or absence of the specified attributes. This estimation goes beyond the traditional t-statistic and the interpretation of coefficients as a rate of change. The approach aligns with ANOVA (analysis of variance) but is distinctly developed from an econometric perspective.

A formal explanation of this methodology is found in [1–4]. However, for less specialized readers, we suggest [5,6]. Without a doubt, an excellent explanation is found in [7].

Beyond this introduction, the article is structured into three sections. The first offers a concise overview of the utility of dichotomous variables and their significance in such estimations. The second section articulates the methodology in formal terms. The third section presents a compelling example illustrating the application of this method. Finally, the conclusions are presented.

2. Importance of Applying Dichotomous Variables as Attributes in Modeling

Dichotomous variables are also known as indicator, binary, categorical or simply qualitative variables. They are nominal-scale variables that signify the presence or absence

of a specific attribute. To illustrate, consider a group of economies as subjects or entities, each possessing distinct attributes such as geographic area, the percentage increase in foreign trade, or membership in international organizations. The entity in this context could be a state or district, with categories including the party affiliation of its population or the increase in manufacturing production within certain limits.

For example, a qualitative variable representing the increase in manufacturing production might take the value of 0 if the increase is less than or equal to 2.5%, and of 1 if it exceeds 2.5%. This exemplifies how dichotomous variables allow for the representation of diverse attributes of a given entity.

Furthermore, this methodology extends its applicability across various empirical science domains, facilitating detailed analyses tailored to specific areas. Entities can vary, encompassing industries with attributes like sector affiliation or contribution to a country's production. Similarly, companies may be entities with attributes such as specialization in a particular market, eligibility for export subsidies, or surpassing a certain threshold of imports.

Moreover, the scope of entities broadens to include individuals, with attributes spanning race, gender, religion, or the quartile to which they belong concerning income perception. This versatile approach enables the application of dichotomous variables to capture a wide array of characteristics across different entities and sectors, enhancing the precision of empirical analyses.

In general, to mention some examples, dichotomous variables can use zero and one to express, in the case of economies, economic growth greater or less than one or more established parameters or affiliation or non-affiliation to a global organization. In the case of companies, they can express membership in the primary, secondary, or tertiary sector or company size as micro, small, or medium. Finally, in the case of individuals, they can represent masculine or feminine, single, married, or divorced, etc.

Once this is highlighted, through ordinary least-squares regression and its corresponding assumptions, the significance and impact of these attributes on the expected value of a quantitative variable are found. It should be noted that the interpretation of the results obtained differs from the usual interpretation of the classic linear regression model, that is, the parameters do not represent a rate of change or marginal propensity. The interpretation is shown in the following section.

3. Formalization of the Method

Since dichotomous variables take on values of 0 and 1, the ordered pair (0, 1) can have any other value by using a linear function.

Let F be a linear function such that:

$$F = a + bD \tag{1}$$

$$\mathrm{b} \neq 0$$

$$\mathrm{a}, \mathrm{b} \in \mathrm{R}$$

When D takes the value of 1, we have:

$$F = a + b \tag{2}$$

When $D = 0$, we have:

$$F = a \tag{3}$$

Therefore, the ordered pair $(0,\ 1)$ becomes $(a,\ a + b)$, which would imply going from a nominal measurement scale of the variables to a ratio scale for the purposes of estimating the expected value.

Now, suppose the estimation of the following model:

$$Y_i = \beta_0 + \beta_1\ D_{1i} + \varepsilon_i \tag{4}$$

where

Y_i is a quantitative dependent variable on a ratio scale;
D_{1i} is a dichotomous variable that represents an attribute (0 for the absence of the attribute and 1 for its presence);
ε_i is the stochastic part of the error.

Assuming that the assumptions of the classic linear regression model by ordinary least squares are met, the expected value of Y_i is represented by:

$$E(Y_i|D_{1i} = 0) = \beta_0 \tag{5}$$

and

$$E(Y_i|D_{1i} = 1) = \beta_0 + \beta_1 \tag{6}$$

To exemplify, supposing that the attribute that characterizes i economies (entities) is the growth rate of their GDP $(dGDP)$, this attribute could acquire three different values:

(1) $dGDP \leq 1.5$
(2) $1.5 < dGDP \leq 2.5$
(3) $dGDP > 2.5$

It should be noted that the number of dichotomous variables included in a model is $k - 1$, where k is the number of possible outcomes to avoid perfect multicollinearity in model estimation and avoid falling into the dichotomous variable trap, that is, an exact relationship between the variables. In this sense, the model is written as:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \varepsilon_i \tag{7}$$

Furthermore, one way to avoid the trap is to include k qualitative variables while eliminating the constant term from the model.

In this sense, if the assumptions of ordinary least squares are satisfied, for the specific case of a GDP growth rate less than or equal to 1.5, both dichotomous variables assume a value of zero. This is considered the base category, where the expected value of the dependent variable Y_i is interpreted as:

$$E(Y_i|D_1 = 0, D_2 = 0) = \beta_0 \tag{8}$$

For the case in which the GDP growth rate is between 1.5 and 2.5, the expected value is:

$$E(Y_i|D_1 = 1, D_2 = 0) = \beta_0 + \beta_1 \tag{9}$$

Finally, for the case in which the growth rate of economy i is greater than 2.5, the expected value is:

$$E(Y_i|D_1 = 0, D_2 = 1) = \beta_0 + \beta_2 \tag{10}$$

But what if we aim to model multiple attributes or qualities? The estimate is presented below. Let us consider the following model:

$$Y_i = \beta_0 + \beta_1 d_{1i} + \beta_2 d_{2i} + \varepsilon_i \tag{11}$$

In this case, d was used to avoid confusion with Equation (7). Now, d_{1i} and d_{2i} represent two different attributes of the entity i, and each of them has only two possible values.

Without loss of generality, the results are:

$$E(Y_i\ |d_1 = 0, d_2 = 0) = \beta_0 \tag{12}$$

$$E(Y_i|d_1 = 1, d_2 = 0) = \beta_0 + \beta_1 \tag{13}$$

$$E(Y_i|d_1 = 0, d_2 = 1) = \beta_0 + \beta_2 \tag{14}$$

$$E(Y_i|d_1 = 1, d_2 = 1) = \beta_0 + \beta_1 + \beta_2 \quad (15)$$

The four expressions represent the expected value of Y_i under the following conditions: (a) the absence of both attributes (Equation (12)); (b) the presence of only the first attribute (Equation (13)); (c) the presence of only the second attribute (Equation (14); (d) the presence of both attributes (Equation (15)).

4. Estimation of a Model to Exemplify the Methodology

Due to space constraints and a desire for simplicity, a straightforward yet highly practical case was selected to illustrate the methodology.

A sample of 42 students from two groups, who successfully completed a basic econometrics course at the Metropolitan Autonomous University Xochimilco in Mexico City, was selected. The subsequent model was then estimated, incorporating three attributes, each with two possible outcomes. The primary aim was to ascertain the impact of these three attributes on the final grades obtained by i students. The attributes under consideration included gender, the specific group in which the course was undertaken, and the grade obtained in the preceding course.

The estimated model is:

$$SCORE_i = \beta_0 + \beta_1\, GEN_i + \beta_2\, GROUP_i + \beta_3\, PREV\_SCORE_i + \varepsilon_i \quad (16)$$

where

$SCORE_i$ is the grade obtained in the completed course as a quantitative variable between 0 and 33 on a ratio scale.
GEN_i is the gender of the student $(woman = 0, man = 1)$ on a nominal scale.
$GROUP_i$ is the group in which student i took the course $(group\ A = 0,\ group\ B = 1)$ on a nominal scale.
$PREV\_SCORE_i$ is the grade obtained by student i in the previous course $(less\ than\ 70\% = 0,$ $and\ greater\ than\ or\ equal\ to\ 70\% = 1)$ on a nominal scale.

The results obtained by least squares are shown in Table 1.

Table 1. Ordinary least-squares regression results.

Variable	Parameter	Coefficient	Std. Error	*t*-Statistic	Probability
C	β_0	24.42715	0.675718	36.1490	0.0000
GEN	β_1	2.155909	0.902383	2.38913	0.0220
GROUP	β_2	2.256025	0.891076	2.53180	0.0156
PREV_SCORE	β_3	0.548145	0.789919	0.69393	0.4919

Care must be exercised when interpreting the results obtained, as they deviated somewhat from those based on the conventional interpretation due to the inherent nature of the methodology.

It is crucial to emphasize that the results presented in this document are not indicative of the cognitive characteristics of the students. In other words, being male or female does not imply greater or lesser capacity; the results are specific to a particular group of students, and any differences captured do not stem from gender-related factors.

Having clarified this point and as a consequence of the previous estimate, the values of the t-statistic and their respective probabilities were observed. Consequently, it was determined that only the variable corresponding to the previous grade was non-significant. The mentioned values affirmed the acceptance of the null hypothesis, indicating that the previous grade did not exert statistically significant effects at a 95% confidence level concerning the grade obtained in the recently completed course.

The model equation is expressed as follows:

$$SCORE_i = 24.42 + 2.15GEN_i + 2.25GROUP_i + 0.54PREV\_SCORE_i + \varepsilon_i \quad (17)$$

where $\beta_0 = 24.42$ represents the base category, that is, female students who belonged to group *A* and obtained a grade of less than 70% in the previous course. As can be seen in Table 1, gender and group had a marked statistical significance, since the value of the t-statistic was by far greater than 2, and the value of its probability was 0.0000, i.e., lower than the level of significance. In general, β_0 is interpreted as the average value of the grade obtained for the elements that are within the base category.

The remaining coefficients are defined as differential intercept coefficients and indicate the extent to which the value of the category, when the dichotomous variable obtains the value of 1, differs from the intercept coefficient corresponding to the base category β_0.

In this context, to establish the significance of gender in this study, the coefficient β_1 was examined, which, in this particular case, was 2.15. It is evident that the variable *GEN* was statistically significant, signifying that, at least in this study, a student's gender played a pivotal role in determining the final grade. On average, the male students in Group A obtained a grade of 26.57, calculated as the sum of β_0 and β_1.

Likewise, the coefficient β_2 corresponding to the group was examined. Similar to the earlier analysis, the *GROUP* variable was found to be statistically significant. Consequently, it can be inferred that a female student who took the course in group B achieved an average grade of 26.67, calculated as the sum of β_0 and β_2. Additionally, the male students who took the course in group B obtained an average grade of 28.82, calculated as the sum of β_0, β_1, and β_2.

It should be noted that it was possible to add the attribute corresponding to the previous course grade, which was 0.54; however, although it represents a numerical difference, it was not statistically significant.

In this context, with an R^2 of 0.40 in the model, the resulting residuals are illustrated in Figure 1.

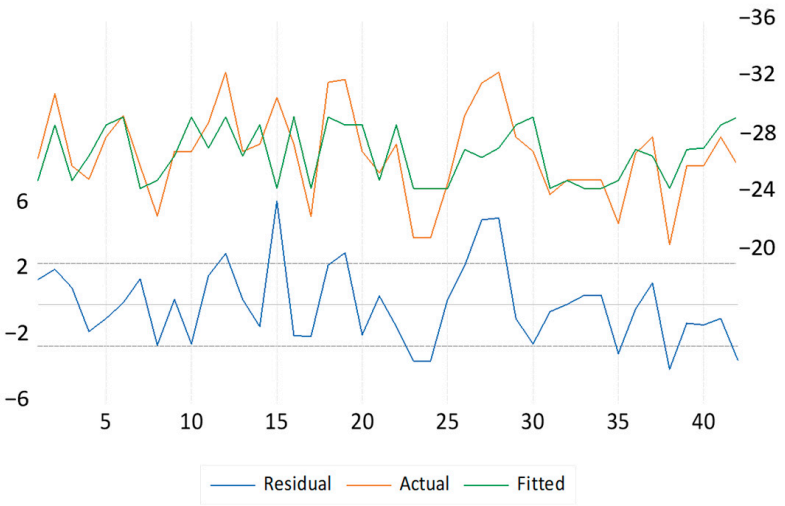

Figure 1. Model residuals.

These disturbances demonstrated a normal distribution, as illustrated in Figure 2, where a kurtosis of 3.08 and an asymmetry coefficient of 0.64 are observed. The Jarque–Bera statistic, registering a value of 2.88 with a probability of 0.23, lent support to accepting the null hypothesis of normality at a 95% confidence level.

Likewise, the errors exhibited constant variance, as validated by the White test, where the probability of the F-statistic was 0.08, indicating the acceptance of the null hypothesis of homoscedasticity.

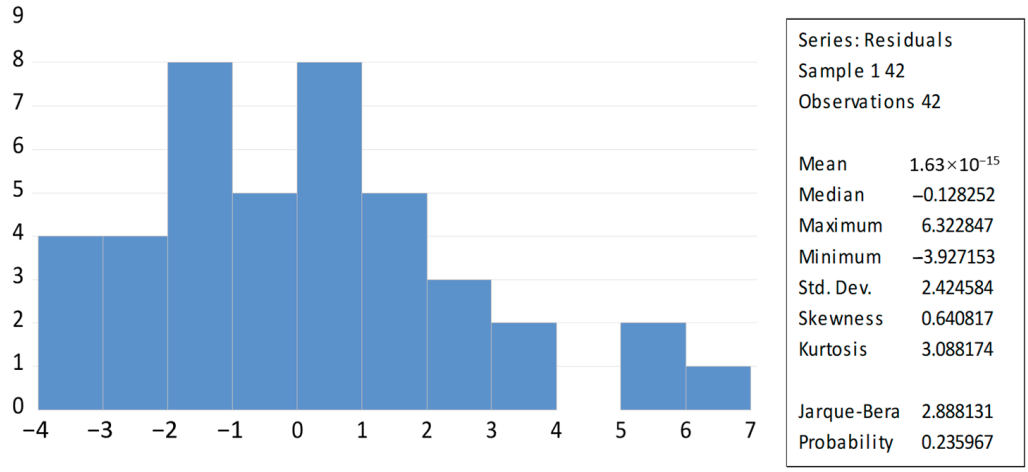

Figure 2. Jarque–Bera test for errors.

Furthermore, the errors in the model displayed the absence of both first-order and higher order spatial autocorrelation, as evidenced by a Durbin–Watson statistic of 1.62 and a probability associated with the F-statistic in the Breusch–Godfrey test with two lags of 0.33.

In summary, given the successful fulfillment of the model's assumptions, the estimators qualified as Best Linear Unbiased Estimators (BLUEs). As a result, making inferences, such as predicting a student's performance in the next course, becomes feasible. In this specific scenario, the university could identify potential external factors influencing the academic performance of students studying basic econometrics and implement appropriate measures.

5. Concluding Remarks

The methodology shown involves estimating the expected value of a quantitative dependent variable on a ratio scale and evaluating the statistical significance of specific characteristics or attributes represented by qualitative variables on a nominal scale, which allows for indicating the presence/absence of a given attribute. In the example developed in Section 3, it was possible to estimate the expected value of the grade obtained in the course by a set of students with different attributes, such as gender, the group where they were enrolled, and the grade obtained in the previous course.

These types of estimates are very useful; once the expected values are estimated, they help forecasting and decision-making related to the various entities, considering their associated attributes, which may be diverse.

However, it is worth mentioning that we must be very careful in interpreting the results, as these are empirical in nature, that is, they capture the characteristics of the entities according to a priori information. That is to say that, in the example shown, it would not be appropriate to assert that female students always tend to obtain a lower grade, as in the first instance, due to gender issues; furthermore, it would be risky to make the same assertion in the case of all the female students of the aforementioned university. In other words, specifically, the female students who took that course earned a lower grade than the male students. With these results, university authorities could examine what happened to this specific group of female students and take it into consideration for subsequent courses.

Author Contributions: Conceptualization, G.C. and X.L.; methodology, G.C. and X.L.; software, G.C.; validation, X.L.; formal analysis, G.C. and X.L.; investigation, G.C.; resources, X.L.; data curation, X.L.; writing—original draft preparation, G.C.; writing—review and editing, X.L.; visualization, G.C.; supervision, X.L.; project administration, G.C.; funding acquisition, G.C. and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this research were generated through a survey of students at the Metropolitan Autonomous University and are available in Appendix A.

Acknowledgments: Conahcyt is widely recognized, specifically, the Program of National Postdoctoral Stays and the Faculty of Higher Studies Aragón of the National Autonomous University of Mexico, which provided the means to carry out this research.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

The table shown below shows the data used in the model that exemplifies the methodology derived from a survey carried out on two groups of students who successfully completed the basic econometrics course at the Universidad Autonoma Metropolitana in Mexico City.

As mentioned in Section 3, SCORE refers to the grade obtained in the course as a quantitative variable on a ratio scale, with values between 0 and 33 and, in the model, serves as a dependent variable.

The independent variables are GEN, which refers to the gender of the student (0 if female and 1 if male), GROUP, which indicates the group in which the student took his/her course (0 for group A and 1 for group B), and finally, PREV_SCORE, which refers to the grade obtained by the student in the previous course, (0 if it was less than 70% and 1 if it was greater than or equal to 70%).

Obs.	Score	Gen	Group	Prev_Score
1	26.5	0	0	1
2	31	1	1	0
3	26	0	0	1
4	25.05	0	1	0
5	26	1	1	0
6	29.5	1	1	1
7	30.5	0	0	0
8	22.5	0	0	1
9	27	0	1	0
10	27	1	1	1
11	29	0	1	1
12	32.5	1	1	1
13	27	0	1	0
14	27.5	1	1	0
15	30.75	0	0	0
16	27.5	1	1	1
17	20.5	0	0	0
18	31.8	1	1	1
19	32	1	1	0
20	23.5	1	1	0
21	25.5	0	0	1
22	27.5	1	1	0
23	20	0	0	0
24	20	0	0	0
25	24.75	0	0	0
26	29.5	1	0	1
27	31.75	1	0	0
28	32.5	0	1	1
29	28	1	1	0

Obs.	Score	Gen	Group	Prev_Score
30	27	1	1	1
31	24	0	0	0
32	25	0	0	1
33	25	0	0	0
34	25	0	0	0
35	18.7	0	0	1
36	26.85	1	0	1
37	28	0	1	0
38	20.5	0	0	0
39	26	1	0	1
40	26	0	1	1
41	28	1	1	0
42	26	1	1	1

References

1. Terza, J. Estimating Count Data Models with Endogenous Switching. *J. Econom.* **1998**, *84*, 129–139. [CrossRef]
2. Greene, W. Functional Form and Heterogeneity in Models for Count Data. *Found. Trends Econom.* **2007**, *1*, 113–218. [CrossRef]
3. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2013.
4. Fox, J. *Applied Regression Analysis and Generalized Linear Models*, 3rd ed.; SAGE Publications, Inc.: New York, NY, USA, 2016.
5. Esteban, M.V.; Moral, P.; Orbe, S.; Regulez, M.; Zarraga, A.; Zubia, M. *Econometría Básica Aplicada con Gretel*; Facultad de Ciencias Económicas y empresariales, Universidad del país Vasco: Leioa, Spain, 2008.
6. Matilla, M.; Pérez, P.; Sanz, B. *Econometría y Predicción*, 2nd ed.; McGaw-Hill: Interamericana de España, Spain, 2017.
7. Gujarati, D.; Porter, D. *Econometría*, 5th ed.; McGaw-Hill Interamericana: Mexico City, Mexico, 2009.