Model between Income and Health

Raoul Coccarda[1]

Department of Statistics. E-Campus University. Novedrate 22060 Italy

Abstract This manuscript This manuscript presents a study on the relationship between income and health status of a sample of 2107 individuals. The measurements were taken by a medical research on urinary symptoms and questionnaire respondents were also asked the income possessed. The author believes that, in this specific case, the income information are more reliable. It 'was then carried out an analysis of the basic model and a comparison in which is studied the relationship between income and level of education. It should be noted that the study is based on two conditions: first that there is a two-way relationship between income and health status, and the second requiring the "discretization" qualitative variables: health status and level of education

Keywords 1. Income 2. Health 3. Educational level 4. The relationship between health and income and between level educational and income. 5. “Discretizzazione” of variables: health and level of education.

1. Introduction

This research presents the relationship between income and health status of a sample of 2107 individuals (statistical units) to whom was administered a questionnaire aimed to identify medical targets and exactly referred to urinary disorders. There are no known author studies and research that have considered the relationship between health and income. It is not, therefore, can refer either to the prevailing doctrine, nor to an adequate and coherent bibliography. The research aims at modeling the relationship between health and income, the results of the data show a link between the two variables almost absent and negative in each case. Intuitively, one should note, however, a positive two-way relationship between health status and income, as, probably, the one who has more economic opportunities should be able to heal better and therefore to have a higher status of health. The "discovery" of a report absent or even negative relationship between health status and income emerged from the analysis of the sample, in our opinion, fairly representative, led the author to investigate in more detail the relationship. The study was conducted from a questionnaire administered to 2107 individuals appropriately labeled, which has been linked with a considerable number of variables mainly related to medical issues (i.e. disorders related to diseases such as diabetes, hypertension, etc..). The idea of the author is that the data on the two phenomena, income and health status, can be the subject of a study independent of the investigation of a medical nature. Proposed model[4] on the relationship between income and health status was compared with that between income and level of education in order to verify two conditions:1. the statistical validity of the analysis, the assumption that there is a correlation between two-way positive health status and level of education; 2. the same lack of homogeneity of the two phenomena represented by a quantitative variable (income) and two variables "nominal-ordinal" (health status and level of education).

Has represented the questionnaire from which it is clear that the questions put to respondents predominantly a medical reference, particularly with regard to issues related to urinary disorders. We proceeded, therefore, to perform statistical analysis describing the phenomenon of interest from the frequency distribution of the variables income, health status and level of education. Advance was necessary to specify the frequency distribution of the variables of sex and age groups because the sample is biased towards the male sex and in certain classes of mature age. The conclusions that have been reached must take account of this imbalance. Was carried out a suitable and comprehensive analysis of missing data to show that they do not affect the results of the study. In a second phase was carried out a detailed analysis on the association of two characters or variable health status and income, addressing issues related to the association between a phenomenon nominal-ordinal and a quantitative and identifying measures of association more consistent with the research objectives

We are analyzed also the inferential problems relating to the measures or indexes of association and hypothesis testing of stochastic independence with the calculation of confidence intervals for the observed health status and income. Was carried out a comprehensive analysis on the correlation between the two phenomena of interest and was finally presented a model of simple linear regression also studied from the point of view of inference.

1. Questionnaire

For the purposes of a better reading and interpretation of the search shows the structure of a part of the structured questionnaire administered to 2107 individuals suitably etichettati5.

DEMOGRAPHIC INFORMATION

A1. Sex

1. female

2. male

B3. Date of birth (/ year)?

SOCIAL INFORMATION

A3. Which of the following categories best corresponds to your level of education?

[Read steps 1-5 and circle one]

1. middle school 2. High school 3. Degree 4. Other 5. No formal education

HEALTH

B4. In general, your health is ...? (Circle one answer)

1. Excellent 2. Good 3. Fair 4. Poor

A6. Which category best describes your current occupation?

[Read steps 1-7 and circle one]

1. Agriculture / Fisheries 2. Crafts / construction 3. Transport / Communications

4. Trade 5. Service 6. Security 7. Other

F12. Which of the following groups is the family income of the last twelve months, adding the income of all members? Remember to consider all possible sources of income, such as salaries, pensions, aid for family, interests and so on:

Classes of reddito7

1 = less than 15000 euro 2 = 3 = 15000 to 25000 from 26000 to 40000

4 = 5 = 41000 to 70000 over 70000 LIFE STYLE 8

B15. Have you ever smoked? 1. No 2. Yes

B16. Are currently taking any of these medicines? [Read each drug before proceeding]

1.No 2.Yes 3. Does not Know.

B6a. Diuretics 1.No 2. Yes 3. Does not Know

B6b. Antidepressant drugs? 1.No 2. Yes 3. Does not Know

B7. Over the past twelve months have you seen a doctor for any reason? 1. No 2. Yes

B7A. How many times?

B13. How many pregnancies has?

1. Abortions, interruptions, extra-uterine pregnancy;

2. Dead at birth:

3. Live birth:

SECTION C: THE HISTORY OF URINARY DISORDERS

I'd like to ask you more questions about your health and urinary disorders that can or not they have experienced. I understand that you feel that these questions prove repetitive, but I was asked to porvele in a specific order, so be patient and try to answer doing your best.

C1. During a hike, how often you urinate? (Circle the answer)

A. No less than six hours

B. Every 5 or 6 hours

C. Every 3 or 4 hours

D. Every 1 or 2 hours

E. More than once per hour

C1.1 do you feel to urinate too often during the day?

1. NO 2. YES

C1.2 On a scale from 0 to 9, what feels annoyed by the frequency with which must empty the bladder? (0 = not at all disappointed, 9 = very disappointed). Circle your answer.

0 1 2 3 4 5 6 7 8 9

C4. Have you ever felt the need to urinate difficult to wait? (That is sudden and intense feeling to urinate that when you feel you must urinate immediately).

1. NO, 2 never 3 YES. 4.I do not know

2.Descriptiveanalysis

3.1.Frequencydistributions

Introductory functional analysis and descriptive statistics of the phenomena-health and income-is the representation of the frequency distributions of qualitative variables sex and age classes. The imbalance of the sample to the male and to the age groups above 25 years is taken as the assumptions of the model, best described later (Borra –Di Ciaccio 2008). Tables 1 and 2 show these distributions.

Table 1. Frequency distributions by sex

Sex					
		Freq. ass.	Freq. ass.%	% ob.val.	Freq. cum.%
Observation valid	male	1298	61,6	63,4	63,4
	female	748	35,5	36,6	100,0
	total	2046	97,1	100,0	
Missing data		61,6	2,9		
Total		2107	100,0		

Table 2. Frequency distributions by age

Age groups					
		Freq. ass.	Freq. ass.%	% ob.val.	Freq. cum.%
Observation valid	0-25 years	6	0,30	0,30	0,30
	26-50years	522	24,8	26,1	26,4
	51-70years	931	44,2	46,5	72,8
	>70	544	25,8	27,2	100,0
	total	2003	95,1	100,0	
Missing data		104	4,9		
Total		2107	100,0		

It can be observed that the gender composition of the sample is 63.4% for males and 36.6% for females, while that by age is significant from 25 years[10]. For the purposes of the research, which aims to analyze the relationship between health and income, this imbalance does not affect its reliability as the distribution of income earners coincides with that of the age group that it can be said that most part of the income is earned by individuals with more than 25 years. 10th.Similarly, the gender imbalance does not affect the study as the income received by individuals male respondents is definitely greater than the income received by women. Tables 3 and 4 are analyzed the frequencies of the phenomena of

interest income and health status. In Figures 1 and 2 are shown the histograms of frequency distributions of the phenomena of interest - income, health status - that show both positive asymmetry (right)[11].

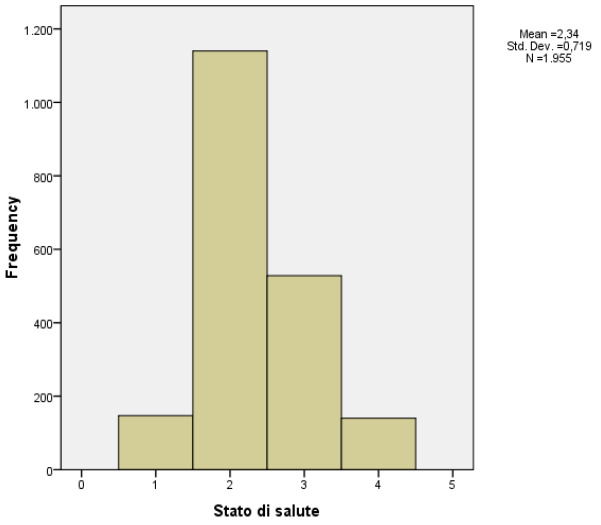


Figure 1. Frequency distribution of health

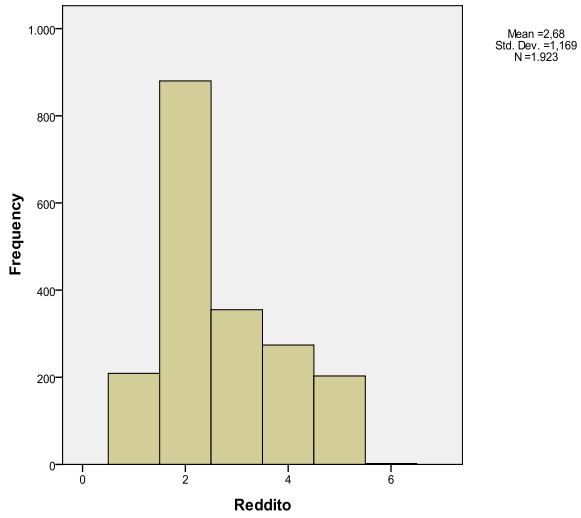


Figure 2 Frequency distribution of income

Figures 3 and 4 show the box-plots respectively of income classes and classes of health status.

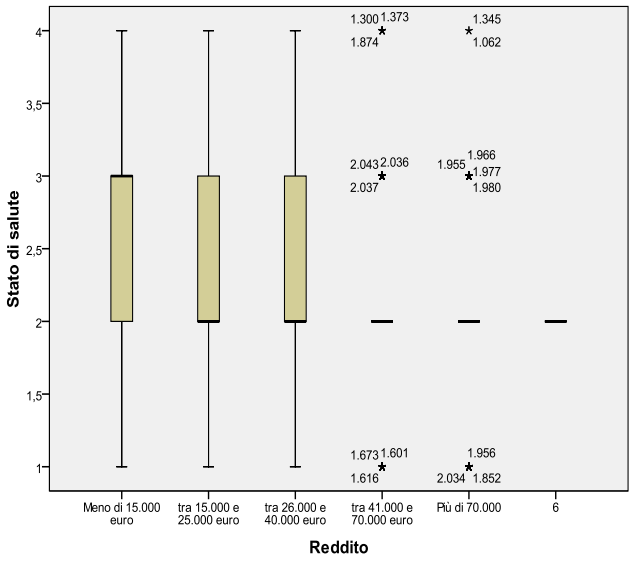
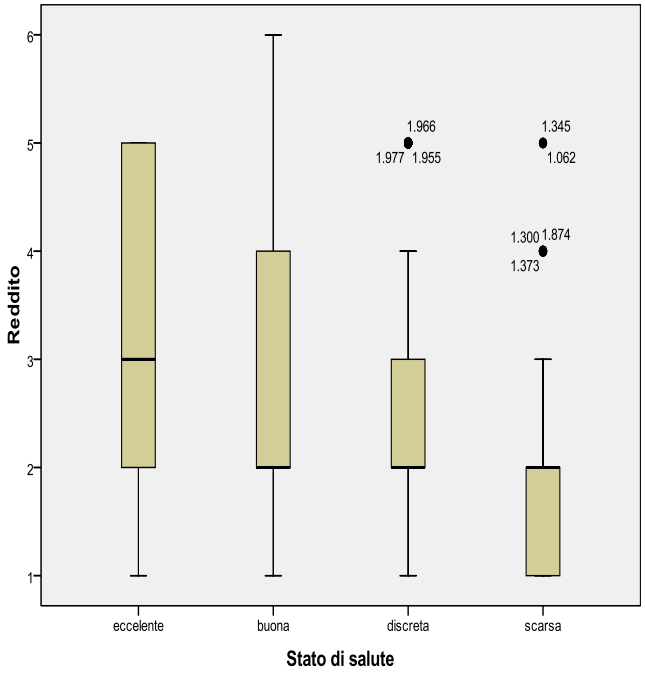


Figure 3. Boxplot income in classes and health

Figure 3[12] shows that there is a relationship in line with intuition only for the income class with less than 15,000 euro, whose values are distributed from the state of poor health to the discreet and no further. Since in this class is placed less than 10% of the sampling distribution, this % is not significant for the purposes of empirical evidence of the relationship between the two phenomena. Approximately 90% of the observations distributed in the other income classes refer to a state of health between good (the majority) and the excellent.

Figure 4. Box-plot health in classes and income



The comments made on the figure 3 shall also apply, as is easily seen, for figure 4[13].

2. Analysis of data gaps and anomalous data

Given that there is a significant% of missing data (missing value) should carry out an analysis that allows information to assess the impact of missing for the two variables of interest. At the same time groped in the first instance to analyze outliers each varied. In Table 5 is an excerpt of the model of missing data. In Table 6, however, shows the univariate statistics of missing data. In Table 7, finally, shows the table of patterns of missing values.

Table 5. Patterns of missing values (cases with missing data)

Case	Missing	Missing%	Patterns of missing extrems[a] values	
			health	income
397	1	50,0		
399	1	50,0		
422	1	50,0		
431	1	50,0		
456	1	50,0		
484	1	50,0		
522	1	50,0		
587	1	50,0		

Table 6. Mean and standard deviation of income and health status

				Missing data		N. Values of estremia[a]	
	N.	Mean		Number	%	Lower	Top
Inc	1923	2,68	1,169	184	8,7	0	205
Health	1955	2,34	0,719	152	7,2		

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Table 7. Part of Table missing data

Case number	Health	Income	Full if……[b]	
1845				1845
78	X			1923
74	X	X		2107
110		X		1955

Table 8. Descriptive statistics of income and health status

	Income	Health
Valid Observations	1923	1955
Missing data	184	152
Arithmetic mean	2,68	2,34
Median	2,00	2,00
Mode	2	2
Standard Deviation	1,169	,719
Variance	1,366	,518
Kurtosis	,642	,549
Quartiles:		
I 25%	2,00	2,00
II 50%	2,00	2,00
III 75%	3,00	3,00

The analysis of the data contained in Table 8 must be carried out taking into account the division into classes of phenomena of interest. The state of health is a qualitative (changeable) to which has been assigned an ordinal scale with an encoding from 1 to 4.

1.Excellent 2.Good 3. Discreet 4. Poor fair

The first observation concerns the difference between mean, mode and median. While the fashion and the median are equal to 2, and then identify a health status of GOOD, the arithmetic mean is equal to 2.34 and therefore the health of the average observations for 1955, net of missing data, is between Good and Discreet.

Income is a quantitative character (variable) which has been assigned an interval scale with an encoding from 1 to 5.

1. Less than 15,000 euro;
2. Between 15,000 and 25,000 euro;
3. Between 26,000 and 40,000 euro;
4. Between 41,000 and70,000euro;
5. More than70,000five

While the fashion and the median are equal to 2, and then identify an average income falls within the class between 15000 and 25000 euro, the arithmetic mean is equal to 2.68 and the average income of 1,955 valid observations, net of missing data, we ranks among the class from 15000 to 25000 from 26000 and the 40000.
It may be noted that the measures more "robust" are represented by the median and fashion. As regards the variability of the two observed phenomena can easily be deduced that it is much stronger for the income variable with respect to the mutable state of health.

With a manual process were calculated variation coefficients that measure the relative variability of the two observed phenomena.

$$CV_R = \frac{\sigma_R}{\mu_R} = \frac{1,169}{2,68} x100 = 43,62\%$$

$$CV_S = \frac{\sigma_S}{\mu_S} = \frac{0,719}{2,34} x100 = 30,73\%$$

Which shows that the variability of the distribution of income is greater than that of the state of health, confirming the data of the indices of variability absolute.

5. Measures of association and independence stochastic

The analysis of the association observed between the two characters is crucial to the construction of the model and to define a valid law, though not all, of the relationship between them.

Is set analysis, shooting consistently in the model which will be discussed later, **giving the income function independent or explanatory variable (Y) and the health status of the dependent variable or response (X).**

Table 10. Measures towards

Index		ES asymptotic[a] value	T approssimativ[b]	Sig approssimativ[c]
Ordinal by ordinal	Somers'd Simmetric -,233	,019	-11,719	,000
	Dependent Health -,210	,018	-11,719	,000
	Dependent Income -,260	,022	-11,719	,000
Nominal range	Eta ,297			
	Dependent Health,273			
	Dependent Income,233			

a. without assuming the null hypothesis
b. the asymptotic standard error is used on the assumption of the null hypothesis ratio likelihood chi-squared probability

From the measurements of the direction given in Table 10 it can be deduced that there is a association[19] between the two characters observed.

The index "Age" for both shows a significant dependence on average close to 30%, verified by a p-value of 0. This goes to show that you can continue your research as, probably, the conclusions drawn from observations of the sample can be extended to the entire universe and the likelihood that the sample values differ from those of the population is very low and close at 0.

It should be noted, in support of this conclusion, that the sample size high almost always leads to reject the null hypothesis as the sample estimate of the indices is very precise and reliable and the asymptotic standard error is very low.

6. Correlation Analysis

The analysis of correlation between the two variables is the basis for the definition of the regression model and then the identification of a possible "law". There has been a double analysis of parametric and non-parametric correlations taking into account the missing cases (pairwise) and without missing data (lisewise) .

Tables 11 and 12 show the first significant to explore to the end that the negative correlation between income and health is highlighted by the index of Pearson for parametric analysis that the indices of Kendall and Spearman21 for non-parametric analysis .

Table 11. Parametric correlations with and without missing values.

		Income	Health
Income	Pearson correlation	1	-,270**
	P-value (two-tailed)		,000
	N	1923	1845
Health	Pearson correlation	-,270**	1
	P-value (two-tailed)	,000	
	N	1845	1955

**The correlation is significant at a level of 0.01 (two-tailed).

Table 12. Parametric correlations with and without missing values.[22]

		Income	Health
Income	Pearson correlation	1	-,270**
	P-value (two-tailed)		,000
	N	1923	1845
Health	Pearson correlation	-,270**	1
	P-value (two-tailed)	,000	
	N	1845	1955

**The correlation is significant at a level of 0.01 (two-tailed)

5. The Model "Health and Income".

Carried out the descriptive analysis, association and correlation between the two phenomena of interest - Income and Health Status - setting a classical linear regression model $Y=a+bX+\varepsilon_i$ As a general approach should be primarily specify correctly the hypothesis that the model is subject and to identify the relevant procedure, which, it should be specified, will represent the reality observed in an approximate way, taking into account the impossibility of an accurate and precise the complex phenomenal reality of the observed variables. The estimated results from a model specified with scientific rigor help us to understand, in general, an economic and social phenomenon and allow to obtain evidence useful, especially at the level forecasting. In the case study aims to establish a "law" generally derived from empirical observation of the relationship between income and health status, rather than a survey-type forecasting. The inclusion of a random variable in the model meets several requirements of:1.asistemics related to human behavior;2.description of the joint effect of variables not measurable, related to measurement errors. The correct specification of the model requires a thorough analysis of the phenomenon under study and the identification of the variables that influence it. The choice of them involves the risk of taking into account the variable is not determinant of the statistical relationship and of neglecting some determinants. Therefore, the specification of the model requires a careful and thorough search of the trade-off between response or dependent variable and independent variable or explanatory .In this case study the model takes into consideration the dependent variable or response "the status of health" and as an independent variable or explanatory "the income". The assumptions that limit the practical application of the model are as follows: 1. the income is considered the phenomenon that should explain the state of health;2. the assumption of income explained by health status does not change the relationship between the two;3. the limitation of typological heterogeneity represented by two phenomena: the income, expressed as a quantitative variable and health status, expressed as a mutable quality ordered;4. the classification of the mutable-state-health as an ordinal cyclic measured on an ordinal scale;5. the transformation of conventional mutable character in the same amount through a numerical coding for class excellent, good, fair and poor - to be taken as a quantitative measurement. The output of the simple linear regression must be read in the light of the above limitations. The surprising result that emerges from the regression analysis is confirmed by the association between the two phenomena observed. Is confirmed a negative regression between the two phenomena is observed which can take as an explanatory variable income or health status. As often happens in statistical analysis refutes what intuitively one might think. Namely that as the income received by classes of mature age does not increase the health status of the same. In Tables 13, 14, 15 and 16 shows the output of the regression analysis with the inferential.

Table 13. Index of model[b]

Model	R	R^2	R^2 correct	Error standard	Durbin-Watson Index
1	,270	,073	,072	,0695	1,953

Explanatory variable: Income Dependent Variable: Health Status

Table 14. Coefficients

	Unst.ed coeffic		St.ed coefficients			Conf int 99%	
Model	Beta	MSE	Beta	T Student	P-value	Lower limit	Upper limit
Intercept	2,782	,040		68,774	,000	2,678	2,886
Income	-,166	,014	-,270	-12,043	,000	-,201	-,130

a.Dependent Variable: Health Status

Table 15. ANOVA

Model	Sum of squares	Degrees of freedom	Mean square	F di Fisher	P-value
Regression	69,981	1	69,981	145,023	,000[a]
Residue	889,344	1843	,483		
Total	959,325	1844			

Explanatory variable: Income; Dependent Variable: Health Status

Table 16. Residues Analysis

	Minimum	Maximum	Mean	Std. Dviation	N
Predicted Value	1,79	2,62	2,34	,195	1845
Residues	-1,616	2,048	,000	,694	1845
Std. Predicted Value	-2,818	1,440	,000	1,000	1845
Std. Residues	-2,326	2,948	,000	1,000	1845

Dependent Variable: Health Status

In Table 17 shows the estimated parameters of Model.

Table 17. Estimated Parameters of Model

Model						Est. Param. Correl.	
	R^2	F di Fisher	D.F. 1	D. F 2	P-value	Intercept	Coeff
Regression linear	,073	145,02	1	1843	,000	2,782	-,166
Regression exponential	,065	127,16	1	1843	,000	2,683	-,070

Explanatory variable: Income Dependent Variable: Health Status

In the figure 5, 6 and 7 shows the observed and expected data (with fit curve), the residues and the scatter plot of regression between health status and income

Figure 7. Scatter Plot

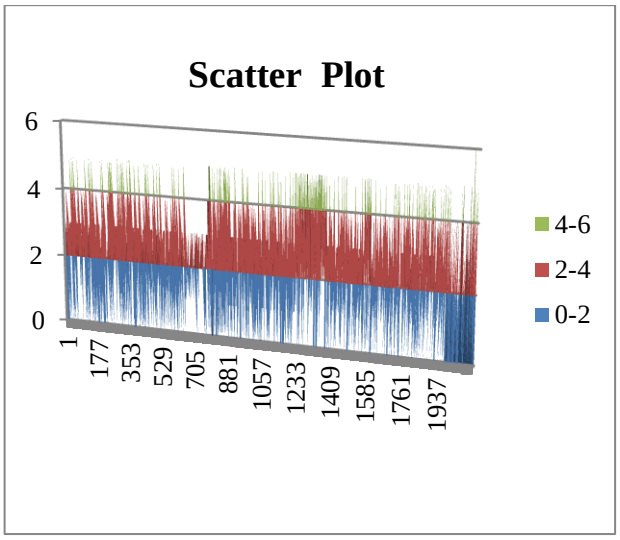
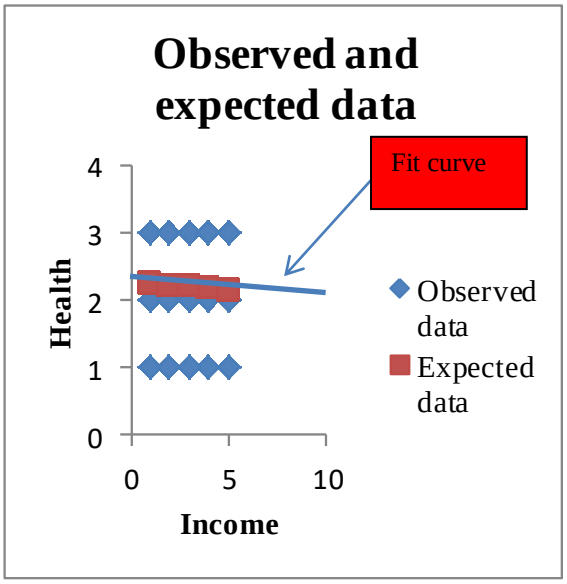


Figure 5. Observed and expected data. Fit curve



8. Model comparison "Health and Education ."

The analysis of a model of comparison between the variable income and mutable nominal level of education carried out on the same sample is useful for demonstrating the validity of the basic one. Even in the comparison model was employed to explain the hypothesis that the income level of education. The results show that there is a positive correlation between the two quantities or that the higher the level of income increases, albeit slightly, the level of education. This was expected for the relationship between income and health status. But we have seen that the analysis shows an inverse relationship to be further discussed. In the following tables 19 and 20 we present the results of descriptive and inferential on the relationship between level of education and income.

Figure 6 . Residues Plot

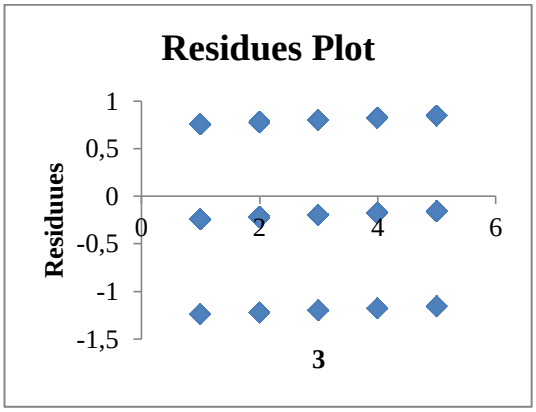


Table 18. Descriptive statistics on the level of education

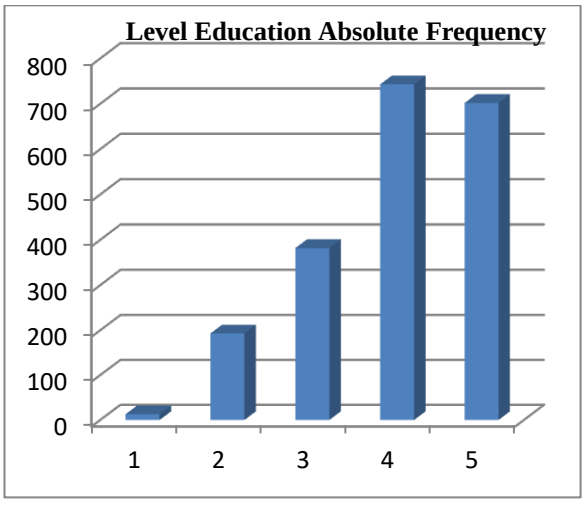
N. Valid Observation	2033
N. Missing data	74
Arithmetic Mean	3,95
Errore standard by mean	,022
Median	4,00
Mode	4
Deviazione standard	,984
Variance	,968
Asimmetric	-,657
Errore standard by asimmetric	,054
Kurtosis	-,409
Errore standard by curtosi	,109
Interval	4

Table 19. Frequency distributions of level education

Levels of education (Code)		Freq. ass.	Freq. ass.%	% ob.val.	Freq. cum.%
Observation valid	None (1)	13	,6	,6	,6
	elementary school (2)	193	9,2	9,5	10,1
	middle school (3)	382	18,1	18,8	28,9
	Vocational school/ High school (4)	743	35,3	36,5	65,5
	University (5)	702	33,3	34,5	100,0
	Total	2033	96,5	100,0	
Missing data		74	3,5		
Total		2107	100,0		

In the figure 8 shows the frequency distribution of level education

Figure 8. Frequency distribution of level education



As it was easy to predict the frequency distribution of the level of education is rather skewed to the left and emphasizes an average fairly high and almost equal (3.95) to the degree (4).

In Tables 20,21,22,23 and 24 are presented data regression between income and level of education.

Table 20. Index of model[b]

Model	R	R^2	R^2 correct	Error standard	Durbin-Watson Index
1	,461[a]	,213	,212	,866	1,687

Explanatory variable: Income Dependent Variable: Level of education

Table 21. Coefficients

	Unst.ed coeffic		St.ed coefficients			Conf int 99%	
Model	Beta	MSE	Beta	T Student	P-value	Lower limit	Upper limit
Intercept	2,925	,050		58,983	,000	2,797	3,053
Income	,385	,017	,461	22,725	,000	,341	,429

a.Dependent Variable: Level of education

Table 22. ANOVA

Model	Sum of squares	Degrees of freedom	Mean square	F di Fisher	P-value
Regression	387,313	1	387,313	516,445	,000[a]
Residue	1432,423	1910	,750		
Total	1819,736	1911			

Explanatory variable: Income; Dependent Variable: Education

Table 23. Residues Analysis

	Minimum	Maximum	Mean	Std. Dviation	N
Predicted Value	-3,850	1,690	,000	,866	1912
Residues	-1,440	2,836	,000	1,000	1912
Std. Predicted Value	-4,446	1,951	,000	1,000	1912
Std. Residues	-3,850	1,690	,000	,866	1912

Dependent Variable: Level of education

In Table 17 shows the estimated parameters of Model.

Table 24. Estimated Parameters of Model

Model						Est. Param. Correl.	
	R^2	F di Fisher	D.F. 1	D. F 2	P-value	Intercept	Coeff
Regression linear	,213	516,44	1	1910	,000	2,925	,385
Regression exponential	,191	451,89	1	1910	,000	2,839	,110

Explanatory variable: Income Dependent Variable: Education

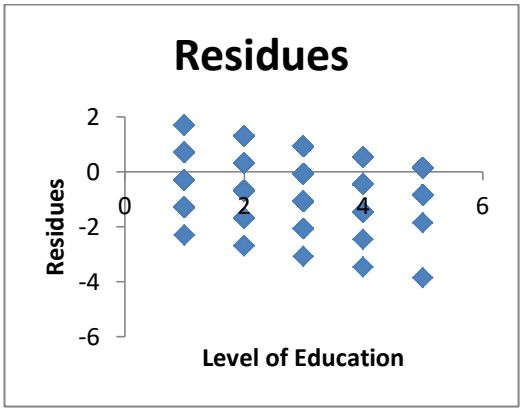


Figure 9. Observed and expected data with fit curve

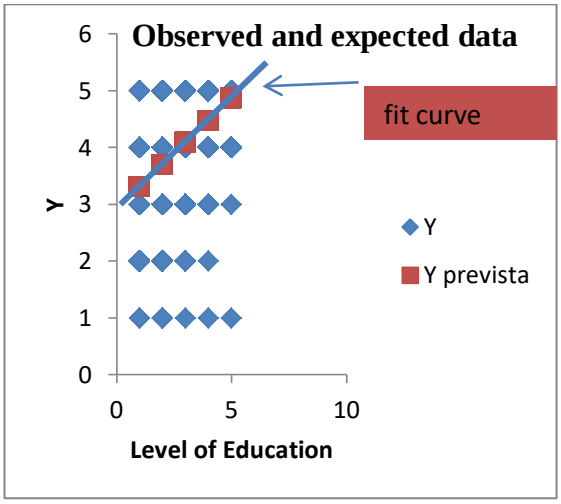
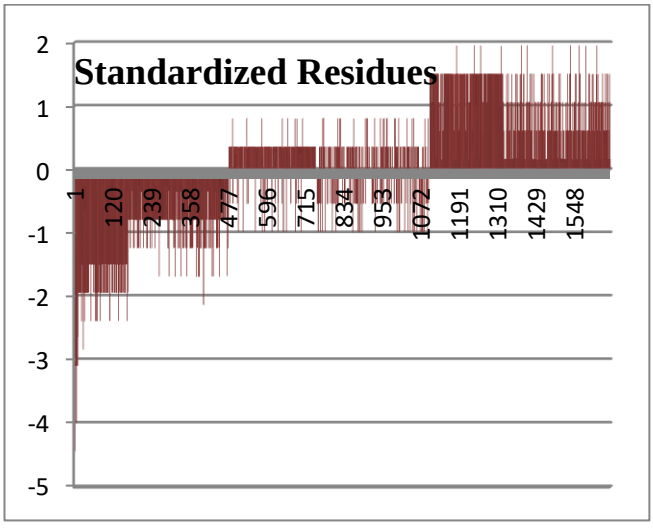


Figure 9. Standardized Residues



The analysis of the curves estimated between the degree of income statements and highlights what had previously been observed that there is a clear positive correlation.

CONCLUSIONS

This study is not intended to be comprehensive nor does it presume to have established a "law" or a "theorem" on the relationship between the two phenomena observed - health status and income. It 'only a contribution to the quantitative analysis of two empirical magnitudes between them is not homogeneous, whose homogeneity has been "forced" the assumption of quantization through numerical codes. The doctrine gives a consistent validation of a regression model in the presence of two quantitative variables. In this research was, however, further study descriptive and inferential statistics in terms of lack of homogeneity between characters. Descriptive analysis of the sample shows an imbalance in the characters sex and age that does not affect the analysis, but which must be taken into account. Inferential analysis emerges a problem of heteroscedasticity of the errors, as can be seen from the graph of standardized residuals, which affects the final results and, therefore, must be considered critically. It should be pointed out, however, that in comparison model, even in the presence of heteroskedastics errors, the relationship between level of education and income follows a trend in line with reliable results on an intuitive level. Of these problems, the author aims to carry out a more extensive and thorough. It is possible, however, to a good approximation, that there is a non positive correlation between health status and income but rather absence of a link, refuting, as often happens in statistical studies, which intuitively one might think, namely that the state of health can improve a function of an economic availability greater

REFERENCES

[1] The questionnaire was extracted by searching "Study of urinary symptoms in Italy" by A. Nicolosi (MD, M.D. Ph.D.) and P. Moi (Ph.D.) - Department of Epidemiology and Medical Informatics - Institute of Biomedical Technology National Research Council.

[2] Borra S., A. Di Ciaccio "Statistica – Metodologie per le scienze economiche e sociali", McGraw-Hill., 2008.

[3] Cicchitelli G., "Probabilità e Statistica", Maggioli Editore (II^Ed.), 2003.

[4] Freedman D., R. Pisani, R. Purves "Statistica." Milano, McGraw-Hill, 1998.

[5] Frosini B. V. "Connessione Regressione Correlazione", Celuc Libri, 1993.

[6] Moschese G., , "Data Mining: Tecniche di trasformazione dei dati" (first and second part), Apogeonline, 2009.

[7] Zenga M., "Inequality curve and inequality index based on the ratios between lowe and upper arithmetic means", in "Statistics and Applications", 5, no. 1, 2007

APPENDIX THEORETICAL

Inferential Analysis

The model is determined by the following notation

$$Y=a + bX + \varepsilon_i$$

where Y represents the state of health, X the income and ε_i a random variable.

It is assumed that the v.c. ε_i is normally distributed with mean μ and variance $\sigma^2 \sim N(\mu; \sigma^2)$.

Consequently, the observations y_i in the specified model are realizations of v.c. Normal Y_i with mean value or expectation:

$$E(\hat{y}_i) = \hat{a} + \hat{b}x_i$$

and variance σ^2, whose notation is represented by

$$Y_i \sim N(\sigma^2, \hat{a} + \hat{b}x_i)$$

The inference analysis is based on the study of the confidence intervals for the regressors and testing hypotheses.

Confidence Intervals

The methodology for the determination of the confidence intervals for the two regression coefficients (regressors) and estimated *a* e *b* at a level of significance 1 – α are given by the following notation:

$$\hat{a} \pm t_{\alpha/2} s(\hat{a}) \text{ e } \hat{b} \pm t_{\alpha/2} s(\hat{b})$$

"Where $t_{\alpha/2}$ indicates that value for which the probability of observing values of the t-Student, with n-2 degrees of freedom, greater than or equal to $t_{\alpha/2}$ is equal to α / 2".

Variance Analysis

From the least squares line can be deduced that the difference between the observed values and the estimated results of the embodiments of the Y, which express the total deviance, can be expressed as the sum of:

$$\sum_{i=1}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \bar{y}) + \sum_{i=1}^{n}\hat{e}^2_i$$

DT = DS + DR

i.e. the total deviance (DT) and decomposed between the explained variation (DS) and the residual sum of squares (DR).

These values are obtained by observing the relationship breakdown

DT / DS = n / n + DR / n

n-1 = 1 + (n-2) = n-1 ►degrees of freedom

or, for the inverse relationship

DT / n - DS / DR = n / n

(n-1) - 1 = (n-2) ► degrees of freedom

The values of DT, DS and DR are defined, also, sum of squares. If these values are related with the number of degrees of freedom is obtained the average of the square defined by the respective notations:

$$MDT=DT/(n-1)= \sum_{i=1}(y_i \quad \bar{y})^2 / (n-1)$$

$$MDS=DS/1= \sum_{i=1}^{n}(y_i \quad \bar{y})^2 / 1$$

$$MDR=DR/(n-2)= \sum_{i=1}^{n}\hat{e}^2_i / (n-2)$$

Source of variation	Sum of squares	Degrees of freedom	Mean of squares	Test F (o di Fisher)
Regression	MDS	1	MDS=DS/1	F = MDS/MDR
Residues	MDR	n-1	MDR=DR/(n-2)	
Total	MDT	n	MDT=DT/(n-1)	

1.3 Analysis of residues

The MSE of the regression is given by the following notation:

$$\text{err std regression(MSE)} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

The MSE of the slope is given by the following notation:

$$\text{err std coeff.corr.} = \text{err std regressione} * \sqrt{(X_i - \bar{X}_i)^2}$$

The MSE intercept is given by the following notation:

$$err \text{ std intercept} = \text{err std regression} * \sqrt{\frac{1}{n} + \frac{(\frac{\sum_{i=1}^{n} X_i}{n})^2}{(X_i - \bar{X}_i)^2}}$$

Hypothesis testing

For the hypothesis testing of the regression coefficients (regressors) the model must specify the null hypothesis or of interest, for which the regressors estimated and are

placed equal to predetermined values, for example respectively c and d, then the test statistics are explained by notations:

$$t = \frac{\hat{b} - c}{s(\hat{b})} \quad \text{e} \quad t = \frac{\hat{a} - d}{s(\hat{a})}$$

They are distributed, under the null hypothesis, as a Student t with n-2 degrees of freedom.

Analysis of variance takes into account the statistical F-test (or Fisher), whose value is a measure for the acceptance (or rejection) or rejection (or acceptance) of the null hypothesis:

$$H_0 : \hat{b} = 0$$

The higher the value of F is close to 1, the more you tend to accept (not reject) the hypothesis of interest H_0.

As far as the test statistic F was much larger than 1 tend to reject (or accept) the null hypothesis H_0 and accept (or reject):

$$H_1 : \hat{b} \neq 0$$