Regression with Ordinal Variables

and irreaslism with correlation

measurements

Abstract

The paper is based on lectures and subsequent mini-symposiums in bio-statistics at Örebro University. Hence, it is interpretative of lectures and lessons learnt thereof.

The essay is striving to highlight the loopholes in using regression model analysis in measuring reality as it is used in social sciences. Social reality is a multiple-plane of actions some random, other mere events others real and continuous how realistically possible is quantitative methodology applicable?

Usually, social scientist have a desire to seek understanding of society disposition, social agents and their actions by looking at correlation of given variables as causes of characteristics assumed to be representative of human nature and her actions. It's for this reason, therefore, the thesis is regression with ordinal variables and irrealism in correlation measurements.

I'll suggest alternative statistical tools as opposed to the usual regression models. How possible could social science methodology adequately be done without a greater distortion and mis-representation of the social reality and facts of substance, is the subject matter of the study.

Introduction

Firstly, I'll have to point out methodological relat-ed social science measurement errors, by posing a question, 'should social science be based on an awk-ward way of doing science?'

It is very important for social scientists, to real-ise the loopholes in regard to the *nature* of human beings in relation to their practical and social or-der, within their respective societies and academics fields. Social science methodology distortion of facts will certainly distort fields where social sci-ences are applicable namely human geography, technol-ogy and engineering, economic and ecology, industrial studies and other fields in medicine, biology and history itself.

Sociology for example tends to ask:

1. What is the *socio-economic* environment pointing to the socio-economic preconditions determining the possibilities of life?

   instead of asking;

2. What are the possibilities of a quality life or pre-conditions in a particular natural environment, necessary for human nature in order to make life pos-sible?

Statistically, the first presumption poses a rigid and ready-made abstracting and modelling error. It reduces and condenses too, a complex social environ-ment (reality) into two objective dimensions the (a). Social independent of its natural properties and (b). Economic dichotomy or spheres. Human beings are natu-ral beings too and indeed the social as well as the economic sphere is small parts of a complex web of society, social agent's and their action's normative-ly. It is not a social sciences purpose to reduce life possibilities but making it better and abundant.

Since human beings are natural beings, therefore are not only objective but are also subjective to propensities of their complex environment. For instance a closed urban environment will certain generate criminal structures which effect is categorical extrinsic to human dispositions- people are not born criminals.

Human societies therefore, are part and intersection of a complex environment (*social being and human na-ture*) hence can't be detached or split from their own natural being, (1). as social agents (2) as a society and (3). as their actions in isolation; any split leads to a regressive moment in (a) practical order
(2) social order (3) natural order.

Now, let us suppose the first question changed and always asks, what is the social natural environment? Such a question is composed of a social moment, which induces a practical moment, which is a natural structure.

Will such a question change the way methodology in social science is done?

Certainly.

Social sciences should not be used scientifically to factor out nature, from the way society social agents and their actions are. Factoring or reduction of HUMAN NATURE to BEING, or simply to the simplest components statistically poses a greater danger to the good intentions of social theory, than can be thought. Notice over dependence on any of the moments generates bias in methodology and applied science. E.g. the social variable studied in isolation generates over simplification of matters of facts i.e. the social sphere becoming a determining factor in both practical and natural order. It has so happened with the physicalisation of social science, in the mechanistic conception of science culminating into rigid probabilistic and mechanistic statistics over application on social sciences.

Hence, social sciences study the natural generality

of social life. It is in these light social sciences, studies *ordinal categories* of social life for natu-ral beings or objects of nature.

Studying e.g. drug, slumisation, crime problems as socio-economic issues, devoid of the respective environmental (natural) perimeters urban enclosures for example, gives wrong formulation of what society, social agents and their actions are in space-time in relation to place and distance. It is likewise true, criminality is a reflection of degrees of human natural environment but not a *person's* disposition. There is scientific evidence to these fact i.e. historical data about rural sociology and differing social characteristic in different social groups within Emil Durkheimian sociology. What generate crime is quite well known in sociology and criminality for example.

Ordinal data

Ordinal categorical data with possible ordered structures are classed namely as (a). Discontinuous (b). Continuous data. In simple terms ordinal data covers events, experiences or mechanism in the real, actual or empirical categories.

Ordinal data where quantified is therefore, not the same as qualitative data. Notice that usually there is a struggle to reach the later in social science studies.

Arithmetically, numbers are not the same as the words or deeds done. Notice, social science tends to make observations of social agents, society or their actions, which later are quantified to represent reality as statistical models figuratively- yet some of the actions might be just random. A cumbersome job if the nature of human beings is not well understood. How realistic are such models to social science?

In the above respect, social science should not be operationalising *intransitive* as *transitive* (epistemological) facts of matter as classical statistical

modelling error. There is a difference on these two dimensions one being the essence of thing of nature and the later being the historical nature of things.

Seeking to establish the *correlation* between what social agents and society does in relation to their general natural actions via their essences require another statistical modelling than regression tools can offer. There is certainly a difference between measuring say one hydrogen and two oxygen atoms in a water molecule and frequency of a particular crime and its cause in a slum or city. Crime in this case is based on intransitive factors like morality, religion, culture which are not enumerative inductable.

I have said above there is a fundamental difference between *human nature* and *human being* in all fields of human inquiry. Social science investigation of a *hu-man being* has a highly streamlined pattern in rela-tion to the *socio-economic sphere* and basically in-stinctive, while *human nature* is universal and dis-cretely patterned in relation to *being (*socio-economic). Being is neither a complement nor a sub-stitute to human nature i.e. biological needs or en-vironmental conditions, which might generate an array of social practices. Where it does it is time con-fined.[83]

Consequently if wrong statistical tools are used to understand social agents, society and their actions, social science will catapult social agents, society and their actions into a survival TRAP and general POVERTY not only in material terms but culturally, mor-ally etc,. (Thurow Lester 1976)

Her poverty implies de-construction of given human structures namely; power relationships (Children/ parents, ethnic, political, gender, employee-employer relationships etc,) discursive and communicative structures and lastly normative and moral imperatives[84].

---

83 Brain Ellis:2001 Scientific Essentialism pg.17-23.

84 Bhaskar Roy; Dialectic The pulse of freedom, pg.161.

The last point can be interpreted as 'bad' society laws. Notice inadequate social science methods, usu-ally generate the problem of *judgmental value* inter-preted as a sociological problems of choice and also as a question of choosing a proper social science methodology for object studying.[85]

Problems

What the above actually indicates, is that by statis-tical measurement using regression models, social scientists are trying to reach an *equidistant scaling* or *intervals* (a logical connection) in the data char-acteristics let that be uni-dimensional or multidi-mensional in the way society, social agents and their actions are.

Can this be possible?

If it was possible, with the thoroughness statistical methods offers, social science would have established a consistent of social life, however as of now there is no scientific evidence to prove such a theory among social agents, societies and their actions does exist, not least among highly homogenous societies.

Besides, there are clear indicative *real properties* and *relations* as identities of what social scientists do investigate into. Notice here too *identities* are not causes. I have written about appearances in structures, which does not automatically imply the nature of things distinct from their intrinsic na-ture. Social sciences tend to take appearances as given hence causative to social effects.

Statistically the relationship of real properties is an assessment of the *concordance* between discrete and continuous scaling of sociological variables. The level between two fallible and interchangeable scales, is a measurement of their consistency, one acting as a surrogate for the other.

---

[85] Holme I. M. and Bernt K. S.; Forskning metodik om kvalitativa och kvantitativa metoder pg.107.

Transposing crime with urban structures is not establishing consistence since urban crime is varied and has different generative mechanisms in the how ex-trinsic factors impact on intrinsic ones. We can therefore assume crime is associated to oppressive urban realities i.e. unemployment, high cost of liv-ing, desolation etc.

The method elucidate here upon, is dependent on cross or contingence table cell frequencies. Note, where there are tied observations, are corrected as the t error or correction factor.

Social Science statistical based studies:

Social sciences analogically should not be analysing surface changes in observed characteristics or identities. These are usually misrepresented as behaviour assumptions rather than inner mechanism i.e. dispositions, properties or propensities. Dispositions, properties or propensities can be power internal relationships; attitudes, inner demographic changes or family structures and preconditions generating them, pedagogical structures derived from social mechanism, disease spatiality etc. Therefore one condition can only be scientifically explained as a set or subset of another but not as a cause of the other - that will only generate a lapse in knowledge generation.

I will mention and concretises too here again, human beings are part and parcel of human nature (social, practice and nature). Subsequently, there is an *association*[86] (not correlation*)* within human nature's way of being in a given location/s, in time and subsequently generated structures given to the parameters

---

[86] the measure of discordance or disordered (D) observation, where the empirical disorder $D = \frac{2\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} x_{ij}\, x_{ij}^{ul}}{n(n-1) - t}$, where t is

$$= m_1 \; m_2 \; \left(\; \right)$$

$$\sum_{i=1}^{x_{ij}} \; \sum_{j=1}^{x_{ij}-1} .$$

social scientists try to analyse in order to isolate changes within society, social agents and their actions.

E.g. low child mortality rate associated with industrial structures is not a correlation or causal ef-fect of women reproductive properties, but rather mechanism implicit in configuration of social life in industrial societies. It is apparent non-industrial societies can equally generate the same social dispo-sitions, properties or propensities.

In the above connection, social sciences as per the purpose of this paper, does not to methodically seek *causal connection*[87] between or among objects of social inquiry. Causality does not answer the question, which so often is not asked- what causes the causes?

For instance, in society today, there is a said sociologically causal connection between higher education and employment possibilities. What is education in this regard? Will the above causal relationship apply if one took into consideration for example society or social agents ability and experiential knowledge to work[88] as human properties, dispositions and propensi-ties?

Extreme formalisation of social facts leads to statistical errors resulting into social science misinterpretation of reality.

Regression models

Let is take a proposition that there is a strong correlation i.e. between housing and employment. On the outset, I'll point out that this is a deterministic relationship in a closed social system not an open one.

As per the above classical statistics, it might be finally concluded therefore that homelessness is a

87 ibid. 73- 83

88 Thurow Lester 1976 **Generating Inequality pg.** 80 - 81.

result of unemployment. At least if we were to use regression model the state of affairs reflected from data and how society will it be informed by how the nature of society is and should be. What are the counterfactual propositions? If we were to argue, in a give city all people who are employed can get plac-es of abode due to cost, will the first proposition still stand as a sufficient explanation of homeless-ness?

What I want to show here above are namely;

1. Statistical deterministic relationship to facts of matter derived from regression modelling.

2. The problem of regression models and causation.

3. The problem of regression and correlation.

What is statistically true does not imply, it is a social scientific fact and it implies the way societies, the social agents and their actions are or should be.

The conjecture above is a reflection of classical industrial social structures and/or class social relationship. In other words, society and social agents irrespective of their social actions and status, can built and therefore distribute houses.

Statements above implies the *population* mean of housing (Y) *distribution* given employment status(X) is functionally related to or simply that there is an average value in housing distribution to employment status[89]. Statistically, the assumption above can be held true, as long as the industrial social structure is the only possible social system and organisation possible and there is into an open world.

Notice, if all employed have better housing it is judgementally predictive to suggest employment status is functionary related to housing distribution. That

[89] $e(Y|X_i) = \beta_1 + \beta_2 X_i$, thus the concept of population regression function (PRF)

is to say, those who have no employment in regard to the above assumption can't have good housing facilities.

What happens to the above sociological assumption when we find there are societies where employment status doesn't determine housing distribution? There will certainly occur a negative correlation, which in real life, does not exist.[90]

Svensson's method

Svensson argues correlation does not *measure* the level of *agreement* and *interchangeability* between two assessments. A strong correlation does not indicate two assessments produce equivalent results.[91]

I have illustrated what Svensson is indicating above in regard to housing employment status etceteras.

Sociologists by doing any type of methodological rating are trying to scale quality of life, ability, need and want in most cases based on comparative studies of *classes, groups, ethnicities, gender* ra-ther than the nature of human societies.

Crompton Rosemary has highlighted the problems of class analysis[92]. There is shifting group positioning given to indicative ruling economic pointers.

Indeed, class shift does not mean for example drug misuse, quality life levels, crime and other social ills etceteras, are limited to lower class groups alone.

When it comes to such sociological problems Svens-son's statistical method is strong enough, since it strives to establish the *agreement* between assess-

---

90 Holme I. M. and Bernt K. S.; Forskning metodik om kvalitativa och kvantitativa metoder pg. 248-249

*see also Elisabeth Svensson: 2000 Guidelines to statistical evaluation of data from rating scales and questionnaires- Örebro University.

91 Ibid 48.

92 Crompton Rosemary: Class and stratification an introduction to current debates pg. 49-75.

ments remembering, some of the measurements have non-metric properties of data from rating scales.

Housing distribution might depend on other factors other than employment wage scales with which regression models are usually done.

The example of class stratification is a strong example in this case.

Social Science methodology

Since in most cases, data used in social science inquiries are basically non-metric, it might also have ordered structure and a number of categories.

Svensson's method is based on rank-invariant method for evaluation of paired ordered categorical data, which in social sciences do occur so frequently.

Methods of measurements focuses on systematic disagreement, in categorisation, separately from disagreement in individual or random classification. I have given three such categorisation namely practi-cal, social and natural order. Reflected as real, experiences or simply mechanism transposed upon the domains of the actual, the real or the empirical.

Therefore for quantitative social scientific inquiries, a social scientist is actually studying inter-rater reliability in histological arrangement and this is not independent of other quantitative in-quiry.

Rank-invariant evaluation of reliability is established by assessments of m x m cross-classification tables where m denotes the number of ordered categories.

Up-per Left (*ul*)		
	ij: th cel l	
		Low er rig ht *lr*

The percentage agreement (PA) is therefore;

(PA) is defined by $\frac{1}{n}\sum_{i=1}^{m} x_{ii}$,

where the *ij* th cell frequency is denoted X_{ij}, I and j = 1, …..m.

Systematic disagreement

The method also allows plotting two sets of cumula-tive relative frequencies for the marginal distribu-tion, which yields relative or receiver operating characteristic curve (ROC).

The subsequent s-shaped curve (convex or concave), is a sign of systematic disagreement in position or the concentration of the categories.

Empirical measure of relative position (RP) and rela-tive concentration (RC) are consistent estimates of two set categorical marginal distribution, as in case of employment (X) and housing distribution (Y), where v = 1……m is the cumulative frequencies of the two sets of categorical marginal distribution.

RP = p_0- p_1, where

$$p_0 = \frac{1}{n^2}\sum_{v=1}^{m}[y_v\ C(x)_{v-1}]$$

and

$$p_1 = \frac{1}{n^2}\sum_{v=1}^{m}[y_v\ C(y)_{v-1}]$$

and

$$RC = \frac{1}{Mn^3}\{\sum_{v=1}^{m}[y_v\ C(x)_{v-1}\ (n - c(\ )_v) - x_v C(Y)_{v-1}\ (n - c(Y)_v)]\}$$

where

$M = \min\{(p_0 - p_0^2),\ (p_1 - p_1^2)\}$

$p_0, p_1 \neq 0$

M is a normalising constant of RC.

Notice that RP, RC and ROC curve according to this method describe the systematic disagreement defined by marginal distributions.

Problems of social stratification is also solved, us-ing what is referred to as Random disagreement of paired ordered categorical classification often dis-persed from the rank – transformable pattern of agreement which can be clearly seen in cross tabula-tion.

The value measure of the random part, the disagree-ment is therefore the relative rank variance (RV).

$$RV = \frac{6}{n^3}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} x_{ij}\left(R_{ij(x)} - R_{ij(Y)}\right)^2$$

Conclusion

Svensson's method is sociologically adequate for application without mystification of the data required in identification of a phenomenon under quantitative investigation.

Systematic changes for a group(s) marginal	ROC
Heterogeneity	
% agreement in categories	PA
In position	RP
In concentration	RC
Random Invid. Changes	RV
Measure of disordered	D(SE)

Notice too, visual analogue scale (VAS) can be used in the measurement non-linear properties of assess-ment. These are inter-scale consistency between dis-crete scales.

Examples:

Analysis of change

Scale a < b < c < d

X	Y
a	a
a	a
a	b
a	b
a	c
b	b
b	c

b	d
b	d
c	b
c	c
c	c
c	d
c	d
c	d
d	c
d	c
d	d
d	d

Frequency distribution (marginal distribution)

Categories	a	b	c	d	total
X	5	5	6	4	20
Accumulated ratio	0.25	0.50	0.80	1.00	
Y	2	4	6	8	20
Accumulated ratio	0.10	0.30	0.60	1.00	

Systematic change on the scale is therefore

Measure: RP ('relative position') ($-1 \leq \gamma \geq 1$)
Parameter: $\gamma = p(X_l < Y_k) - P(Y_l < X_k)$

Empirical Formula: $RP = p_o - p_1$

$$p_0 = \frac{1}{n} \sum_{v=1}^{m} \frac{1}{2} y_v \left[C_{(x)_{v-1}} \right]$$

and

$$p^1 = \frac{1}{n^2} \sum^{m} \left[y_v C (y)_{v-1} \right]$$

Frequency distribution (marginal distribution)

Categories	a	b	c	d	total
X_i	5	5	6	4	20
$C(X)_i$	5	10	16	20	
Y_i	2	4	6	8	20
$C(Y_i)$	2	6	12	20	

$$p_0 = \frac{1}{n^2} \sum_{v=1}^{m} y_v \left[C_{(x)_{v-1}} \right]$$

$= 1/20^2[4*5 + 6*10 + 8*16] = (20 + 60 + 128)/400 = 0.52$

$$p_1 = \frac{1}{n^2} \sum_{v=1}^{m} [y_v C(y)_{v-1}]$$

$= 1/20^2[5*2 + 6*6 + 4*12] = (10 + 36 + 48)/400 = 0.24$

$RP = P_0 - P_1 = 0.52 - 0.24 = 0.28$

$$RC = \frac{1}{Mn^3} \{ \sum_{v=1}^{m} [y_v C(x)_{v-1} (n - c(x)_v) - x_v C(Y)_{v-1} (n - c(Y)_v)] \}$$

where

$M = \min\{(p_0 - p_0^2), (p_1 - p_1^2)\}$

$p_0, p_1 \neq 0$

M is a normalising constant of RC.

Therefore $M = \min \{(0.52 - 0.52^2 = 0.80), (0.24 - 0.24^2 = 0.30)\}$

RC = $1/(0.30*20^3)$* [6*6(20-16)- 6*10(20-12)]+ [5*2(20-10)-4*5(20-6)]

= 1/2400* (144-480) + (100-80) = -0.132

Rank Transformation Patterns in tabulated form

		X				
		a	**b**	**c**	**d**	**Total**
	d			4	4	8
	c		4	2		6
y	**b**	3	1			4
	a	2				2
	Total	5	5	6	4	20

PA = (2+1+2+4=9)/20= 0.45 (45%)

Observed changes

		X				
		a	**b**	**c**	**d**	**Total**
	D		3	3	2	8
	C	1	1	2	2	6
Y	**B**	2	1	1		4
	A	2				2
	Total	5	5	6	4	20

PA = 7/20 = 0.35 (35%)

Agreement ration PA/PAmax = (0.35/0.45)= >0,70

- **Reference** -

-Bhaskar Roy:1993 **Dialectic** The pulse of freedom, Verso.

-Brain Ellis:2001 **Scientific Essentialism** – Cambridge University press.

-Crompton Rosemary: **Class and stratification** an introduction to current debates. Blackwell Publishers - Holme I. M. and Bernt K. S.: 1986 **Forskning metodik** om kvalitativa och kvantitativa metoder Tano Oslo. - Elisabeth Svensson: 2000 **Guidelines to statistical evaluation of data from rating scales and questionnaires**- Örebro University.

-Elisabeth Svensson: 2000 **Comparison of the Quality of Assessments Using Continuous and Discrete Ordinal Rating scales**- Biometrical Journal 42-4, 417-434. Dept. Of Mathematical Statistics-Chalmers University of technology and Göterborg University.

-Elisabeth Svensson: 1997 **A coefficient of Agreement Adjusted for Bias in paired Ordered categorical Data**; Biometrical Journal 29(1997)6, 643-657.

-Thurow Lester 1976 **Generating Inequality**. *The Macmillan press Ltd*