A different Approach to Regression - Correlation in the Social Sciences

Othmar W. Winkler
Georgetown University
Washington, DC 20057 – USA
<winklero@georgetown.edu> and <winklero@msb.edu>

Regression-Correlation Analysis, considered to be one of the most useful investigative statistical tools, will be discussed in the following specifically for statistics in the social and economic sphere. It should be noted that frequently the fact is ignored that the name STATISTICS refers to three quite different kinds of statistics. Despite some overlap, they differ substantially in origin, purpose, methods and results. The first and oldest use of the word STATISTICS referred to the registration of the population in a political-territorial entity, a Astate@- hence the name Astatistics@- . This eventually included economic activities as well as the living and social conditions of their inhabitants. Although mainly of practical-political importance, some important theoretical development originated here. To mention a few: index numbers, sampling, interviewing techniques, aggregation, and the national accounts. The second use of the term STATISTICS occurred in the natural sciences. The origin and purpose as well as the nature of their data are quite different, leading to correspondingly different methods such as the design of experiments, survival models, and multivariate methods. This second type of STATISTICS developed independently of statistics in the socio-economic field. The third use of the term STATISTICS occurred in mathematics. Although originating in the problems of measurement and measurement errors in astronomy, mathematical statistics took off on its separate, independent tangent. Mathematicians pursue e.g. distribution theory and abstract probabilistic models without regard for the peculiar needs of potential fields of application, such as its indifference to the aggregate nature of the data in the social sciences

Regression Analysis with its terminology and implicit assumptions was introduced in the social sciences from the bio-sciences, despite the fact that conditions in the social sciences are quite different. The phenomena in society, in its population, its economy and sociology are of a decidedly temporal-spatial , that is, short-lived and regional character. In other words, the socio-economic phenomena should be expected to vary substantially from period to period, and from one region to another. This is probably the most important characteristic by which the phenomena and data in the social sciences differ from those in the natural sciences. Another important difference, the subjective self-reporting instead of the objective observation by an outside investigator, does not affect the data I am using in the following to demonstrate the suggested different, better suited, approach to Regression-Correlation Analysis in socio-economic phenomena. The data I am presenting represent the characteristics of residential real-estate properties that were sold in the Washington, D.C. area during a recent two year period, obtained from the official records kept by the county registrar in each district. These data are free of the usual problems afflicting economic and social data, that are obtained through self administered questionnaires or personal interviews, such as memory lapses, intentional misrepresentation or non-response.

The regression line, least squares or otherwise, linear or non-linear, arithmetic or logarithmic, in one or two-steps, simple or multiple, is assumed to reveal the underlying relationship between >variables.= Individual observations that depart from this regression line or surface, that is believed to represent the underlying, common relationship or force, are considered deviations, errors due to random influences. When it is believed that differences in the time of year or differences in the regional districts account for some of the dispersion, helping to explain those deviations, dummy or binary variables have been introduced for the time periods, e.g. trimesters, and/or regions. Instead of such an insensitive procedure I propose to recognize this peculiar feature of economic and social statistics by centering Regression Analysis on the short-lived and regional nature of its phenomena.

Before proceeding further, I should draw attention to the difference in the interest of the abstract, academic Regression-Correlation research, and the much more concrete interest of practitioners in Regression Analysis. These latter include forecasters, and all who are involved in shaping the actual economic and social life of the population in the region. Academic research, interested in the underlying, abstract Alaws of economics@ or of a given social phenomenon, will be inclined to disregard the regional differences

and changes over time. The interest of practitioners, business people and persons in public administration and politics, in contrast, is more concretely focused on the short-lived, regionally diverse quantitative information regarding that phenomenon. Such different interests in Regression-Correlation Analysis, of academicians and of practitioners, should lead to different statistical approaches.

When dealing with the typically practical economic problem, real estate sales in the different parts of a country, at different times of the year or of the business cycle, the three most important feature are humoristically given as A1. location (of the property to be sold) 2. location and 3. location.@ In other words, the location of the economic entity to be researched, is of paramount importance. It should therefore also be treated prominently in Regression Analysis, not as an afterthought, assigning it an insensitive dummy variable, with value of >1' if the info is from region A, and >0' if it is from another region. That may satisfy a bio-scientist for whom location may be of minor interest, but does not satisfy the social scientist.

I therefore propose a twofold approach to the regional and also temporal differentiated data.

The academic approach to social phenomena treat the data for each region and time period separately. One hopes that the region and time span, for which data are collected, are sufficiently narrow so as to contain data that are as similar as possible, e.g. sales districts of similar housing and population characteristics, and time periods, e.g. monthly data, that are sufficiently short to capture changes in the sales within a month. Regression Analysis is to be carried out separately for each region-time grouping. Such analysis will yield interesting information about how the relationship of the investigated variables changes over time and in districts that are contiguous and in those that are further apart. This first step will be of particular interest to practitioners. Instead of dummy variables for region and for time period, the Regression Analysis is carried out separately for each of these period-regions. The academic researcher will continue with a second step: convert the data for each region-period into standardized data, then pool them into one single data set, adjusted for regional differences and for changes over time, hoping that an underlying relationship may emerge.

The practitioner, on the other hand, will run the Regression Analysis separately for each region/time period, and accept the differences he encounters as valuable information to be noted and to be worked with. He then can place these findings in their historic and geographic context to discover the bigger >picture.= The academic researcher, on the other hand, will use these only as a first step in his pursuit of abstract, timeless, general socio-economic >laws,= while it is precisely these differences that are of interest to the practitioner

Sommaire

Le terme >statistique= couvre trois genres différentes de théorie et d=application: 1. Le compt de personnes, leurs activités et niveau de vie 2. Les statistiques appliquées à la biologie et aux sciences naturelles, et 3. Une forme de mathématiques pures. Ces trois types de statistiques se sont developpées séparément et aujourd=hui ils ont peu en commun. - Ensuite je discute de la méthode de l=analyse de régression avec des données réelles de vente de biens immobiliers durant deux ans dans la région de Washington D.C. Parceque ces phénomènes socio-économiques changent rapidement, à un rythme différent par région, il faut calculer la régression pour chaque région et pour chaque période séparément pour les besoins pratiques des politiques et des hommes d=affaires. Cette procédure doit être superieure à la pratique courante d=utiliser >dummy variables= par régions et par périodes. Par contre, les chercheurs académiques doivent normaliser les données dans chaque région-période, puis faire l=analyse de régression de l=ensemble de ces données abstraites.