DePaul University

From the Selected Works of Fernando De Maio

2014

Regression analysis and the sociological imagination

Available at: <https://works.bepress.com/fdemaio/20/>

Regression Analysis and the Sociological Imagination

Fernando De Maio

DePaul University, Chicago, USA
e-mail: fdemaio@depaul.edu

Summary

Regression analysis is an important aspect of most introductory statistics courses in sociology but is often presented in contexts divorced from the central concerns that bring students into the discipline. Consequently, we present five lesson ideas that emerge from a regression analysis of income inequality and mortality in the USA and Canada.

Keywords:

Teaching; Income inequality; Social structure; Social justice; Dummy variables; Logistic regression.

INTRODUCTION

It is well known that social science students approach statistics with a certain degree of apprehension (Finney and Schraw 2003). This may be particularly prevalent in sociology, a discipline with a highly marked methodological divide that pits qualitative against quantitative (Onwuegbuzie and Leech 2005). This divide is presented in many introductory sociology textbooks and in the curriculum itself, with most universities courses on methods divided between ethnographic/qualitative/grounded theory and statistical approaches. When the statistics course is compulsory, students often view it as a burdensome task, rather than the development of an integral component of their sociological toolkit.

Introductory statistics courses in sociology often fall into one or two common traps: one trap is that they present statistical techniques in a way that seems divorced from the issues that interest sociology students. For instance, they may focus on hypothetical examples from microeconomics or even engineering and ask students to deal with fictional commodities. Other times, they 'dumb down' the material and present an oversimplified view of statistics. Although this is done in an attempt to make the material accessible, the end result can be that statistical analysis is not seen as useful in the eyes of sociology students – many of whom are drawn to the discipline in order to 'understand people, not numbers'. For many

students, statistical approaches do not seem to address the complexity of sociological theory or the fuzziness of real-world problems and issues. When taught in this way, statistical analysis takes on a reductionism that makes it irrelevant to many sociology students.

These pedagogical traps are incongruent with ongoing calls for reform in the statistical education research literature. Researchers in this area have published extensive analyses calling for the use of real examples in the classroom, particularly at the introductory level (Castro Sotos et al. 2009; Doane and Seward 2011; Howley 2008 and Valero-Mora and Ledesma 2011). The use of real-world examples is known to have a positive effect on students' attitudes towards statistical analysis (Vanhoof et al. 2011). Indeed, according to Cobb and Moore (1997), "In data analysis, context provides meaning" (p. 803); "...data are not just numbers, they are numbers with a context" (p. 801). This perspective recognizes that introductory courses need to stimulate and nurture understanding of statistical technique, but at the same time suggests that this is best approached from a pedagogical practice engaged with issues of data collection, production and interpretation of real-world examples (Wiberg 2009).

Many sociology students come to introductory statistics with ideas of statistics as dehumanizing, as limited for trying to apply a natural science model to the social world, and as giving a false sense of precision. Some hold that statistical

analysis of relationships between variables creates a static view of social life that is independent of people's lives. From that perspective, a focus on statistical technique leaves little energy or time to deal with history and culture and politics – the prime concern for sociologists (Mills 1959).

This attitude echoes a position from a former president of the American Sociological Association, Joe Feagin (Feagin 1999, p. B4), who decried “one reason why sociology does not currently have more social impact is its over-emphasis on advanced statistical methods and a neutrality toward society's marked inequalities. Like other social scientists, too many sociologists have lost touch with the moral and practical concerns from which our field emanated”. Elsewhere, Feagin has argued for a sociology that places social justice at the centre of its gaze (Feagin 2004). He calls for innovative, creative and rigorous research that studies the ‘big social questions’ of our time – the costs of racism, sexism and other forms of oppression, of environmental crises, economic marginalization and the concentration of wealth and of the violation of human rights. Feagin clearly sees incongruence between the task of sociology (with a focus on social justice) and the methods of sociology (which he notes with derision have become too dependent on statistical analysis).

Related to this, feminist work in the social sciences has drawn attention to the power inequalities embedded within the research process (Taylor 1998). In particular, feminist critiques of survey methods as exploitative and hierarchical have led many sociologists to equate quantitative methods with a traditional mode of operation with which they take issue (Bryman and Teevan 2005); as a result, many sociologists who rightfully value meaningful and reciprocal relationships with research participants (or ‘subjects’, to use a somewhat anachronistic term) have come to place little value in quantitative methods.

The approach that I have found is useful in helping students to overcome these anxieties and see that the potential of statistical analysis in sociology revolves around the notion of ‘radical statistics’ – the idea that statistical analysis can be used to not only describe but also challenge inequities. This questions Feagin's assessment of the role of statistical analysis in sociology, and importantly, this entails a particular approach to teaching the subject. Fictional examples are kept to a minimum, and instead, students work with challenging examples of measuring human rights (Landman and Carvalho 2009), measuring race/ethnic inequalities in exposure to environmental hazards (Downey and Hawkins 2008) and measuring

inequalities reflective of the social determinants of health (De Maio and Kemp 2010).

CLASSROOM TIPS

The notion of ‘radical statistics’ is useful when introducing sociology students to regression models. Some of the sociological literature that uses regression utilizes advanced techniques based on multilevel analysis, but a great deal of it uses relatively simple models that can be used quite effectively with introductory students. Below, I outline five points central to effectively teaching sociology students about regression analysis.

1. Build discussion from real data

An article that I have found very useful in this area is the Ross et al. (2000) analysis of the ‘income inequality hypothesis’. This has been an area of rich methodological and theoretical debates, linking social epidemiology to seminal sociological theories of Marx (on alienation and conflict), Max Weber (on status) and Emile Durkheim (on social solidarity/cohesion) (for comprehensive overviews, see De Maio 2010 and Wilkinson and Pickett 2009). Along with theoretical ties, the literature on the income inequality hypothesis can be used with students to analyse questions of epistemology, as much of the literature has questioned the positions of positivism and critical realism (Coburn 2004). The hypothesis broadly asserts that an individual's health is influenced not only by their own income but also by the level of inequality in the area where they live.

Both ecological and individual-level studies have been published. The analysis in Ross et al. (2000) is at the ecological level, and this feature of the study can be used in the classroom to stimulate discussion about the ecological fallacy (Diez-Roux 2002). Their research compared the health effects of income inequality in the USA and Canada, both at the state/provincial level and at the level of metropolitan areas. Income inequality was defined as the percentage of the total household income received by the poorest 50% of households, a commonly used indicator – this is referred to as ‘median share’ in the regression models discussed below. In contrast to other measures of income distribution such as the Gini coefficient and the Atkinson index, the median share of income has an attractive conceptual clarity that is useful in the classroom. The dependent variable (population health) was operationalised as age-adjusted

all-cause mortality. The clear validity of this variable as an indicator of population health also helps in classroom discussion.

Building on bivariate correlation analyses, Ross et al. (2000) developed an ordinary least squares regression model to estimate the impact on mortality of a 1% increase in the proportion of income earned by the poorest 50%. Their results suggest that such an increase leads to a decline of 21 deaths per 100,000 per year for the working age population. More important, this finding was robust to the inclusion of a city's median income as an additional independent variable (with the coefficient associated with median share changing from -21.71 to -21.80 in their USA-only models; see Table 1 for selected results). Ross et al. (2000) also detected important differences between the USA and Canada, with the Canadian data showing overall lower income inequality and lower mortality than the US data. Their combined USA–Canada analyses (which contained a dummy variable for country) indicated a strong pathogenic effect of income inequality (resulting in an R^2 of 0.51; see Table 1).

Their analysis has been particularly useful in the classroom – their final model is relatively simple (three independent variables) with a clear and easy-to-understand dependent variable (number of people per 100,000 who die every year) which changes with the independent variables in a readily interpretable fashion. The article can be used with students to help them relate results presented in a table (listing intercepts and coefficients, as Ross et al. 2000 do), with the implied regression equation, for example,

$$\text{Estimated Mortality} = a + b(\text{Median Share})$$

In particular, their initial model, which only contained median share as the independent variable, for the USA, was

$$\text{Estimated Mortality} = 848 - 21.71(\text{Median Share})$$

Their second model then added city median income as a second independent variable

$$\text{Estimated Mortality} = 838 - 21.80(\text{Median Share}) + 0.4(\text{Median Income})$$

Their third and final model added Canadian cities to the analysis, along with a dummy variable for country

$$\text{Estimated Mortality} = 826 - 20.92(\text{Median Share}) + 0.2(\text{Median Income}) - 67(\text{Canada})$$

(Here, as standard, the variable 'Canada' is equal to 1 for cities in Canada and 0 for cities not in Canada.) Classroom discussion of these results can emphasize the idea of 'model building'. Students can see how the effect of median share did not change substantially over the models (leading the authors to claim a robust effect of income distribution). By writing out the basic equations through 'plugging in' coefficients from Table 1, students can see – with a relatively simple example – how models can be expanded. Classroom discussion can include discussion of how to further refine the model by adding more variables, and in more advanced courses, examining how interaction effects may be included. Classroom discussion or follow-up written assignments can integrate these statistical lessons with notions of 'theory building'; in this context, students could be introduced to debates between the psychosocial and neo-material explanations of the health effects of income inequality (De Maio 2010). Perhaps, most important for sociology students, this example can easily be integrated with discussion of 'radical statistics' – quantitative research that is attuned to major questions of inequality and social structure.

2. Discuss all elements of the regression equation

The discussion above can be used to meaningfully examine how to interpret a regression coefficient. The Ross et al. (2000) article can also be used in the classroom to introduce the idea of centring independent variables (more interesting, something that the authors did not do in their analysis). For example, discussion can be steered towards interpretation of the intercepts in Table 1 (848 in

Table 1. Selected regression results – income inequality and mortality, adapted from Ross et al. 2000

<i>Model</i>	<i>Intercept</i>	<i>Median share (%)</i>	<i>Median income (US \$1,000)</i>	<i>Country dummy</i>	<i>R²</i>
USA only	848 *	-21.71 *	–	–	0.33
USA only with income added	838 *	-21.80 *	0.40	–	0.34
USA–Canada with dummy	826 *	-20.92 *	0.20	-67^*	0.51

* p -value < 0.05 .

model 1, 838 in model 2 and 826 in model 3). Students are often puzzled by interpretation of the intercept – for example, in the Ross et al. (2000) study, the first model intercept reflects a hypothetical place where the bottom 50% makes 0% of the income (the initial model thus predicts that 848 per 100,000 working age people will die in that city every year). What if, instead of median share, the authors had first created a 'centred' variable by subtracting the mean from all observations? The intercept would then reflect a far more meaningful situation – the expected mortality in a place with average levels of income distribution. For sociology students, this holds great appeal. It brings the results closer to empirical reality and allows students to see more clearly that the results reflect what C. Wright Mills (1959) called the sociological imagination – the linking of biography and history, of personal troubles and public issues. The sociological imagination allows us to see how ill health – one of the most personal of personal troubles – is shaped by public issues; in this case, how mortality among working age people is influenced by relatively small changes in income distribution.

3. Work with dummy variables

Another appealing aspect of the Ross et al. (2000) study is its use of both continuous and dummy independent variables. This is critically important in social science, because many of the variables that we work with are categorical. Students need to be able to work with dummy variables, and this is unfortunately often neglected in introductory textbooks for our students. In the Ross et al. (2000) analysis, the country dummy variable has a substantial and statistically significant effect (with a coefficient of -67). Classroom discussion can be devoted to understand how the dummy variable works in the equation, what it signifies at a conceptual level and how it might be 'unpacked' by adding more variables to the equation. In a more contextual sense, the country dummy could be discussed for what it represents: in this case, a more universal and equitable health care system, more generous unemployment benefits and a host of other social policies that differentiate the Canadian welfare state from that of the USA.

4. Extend to logistic models

As shown above, the Ross et al. (2000) article works very well for students as an introduction to ordinary least squares regression. However, sociological research guided by the idea of 'radical statistics' often takes the form of logistic regression – a technique that often receives very superficial

(if any) treatment in introductory textbooks for social scientists. And this is understandable, given the complexity of the technique. However, it is quite possible to develop discussion around odds, odds ratios and confidence intervals. The income inequality literature is replete with examples that use logistic regression – these can be introduced after analysis of the Ross et al. (2000) article in productive ways. Studies in this area have used logistic regression to study the relationship between income inequality and a variety of dichotomous outcomes, including self-rated health (Craig 2005), adverse birth outcomes (Auger et al. 2009), mental health (Kahn et al. 2000) and obesity (Subramanian et al. 2007).

5. Work with predicted values, not just coefficients and odds ratios

Last, an aspect of regression analysis that holds particular promise for engaging with the 'sociological imagination' involves moving from discussion of coefficients/odds ratios and their statistical significance towards analysis of predicted values. For example, De Maio et al. (2008, 2009) have presented an 'average, deprivation, inequality' framework. The average perspective predicts an overall value – for example, the average rate of a disease in a particular place. The deprivation perspective uses regression results to identify the worst-off social group and predict their expected level of, say, diabetes or asthma. The inequality perspective uses regression results to identify the best-off group and then compares that group to the worst-off group. With longitudinal data, this can be used to track how patterns of inequity change over time. This allows us to evaluate not only how the prevalence of a particular condition is changing but also how the gap between the worst-off and the best-off is changing as well. This is particularly important for sociology students interested in health inequalities, as public health interventions have a history of increasing differences between the best-off and the worst-off (Bartley 2004).

CONCLUSION

Regression analysis holds an important place in most introductory statistics courses for sociology students. Unfortunately, the technique is often approached in ways that are divorced from the concerns of sociology students. The lesson ideas described in this article focus on a regression analysis of income inequality and mortality. On the basis of a relatively simple model, these lesson ideas offer the potential for developing students' capacity to interpret regression results.

Classroom discussion based on the income inequality hypothesis can give sociology students the necessary context from which to engage in statistical analysis. Such practice would reflect recent developments in statistics education research (Zieffler et al. 2008), which emphasizes the need to focus on statistical reasoning, statistical thinking and statistical literacy, rather than on mathematics (Ben-Zvi and Garfield 2004; Castro Sotos et al. 2009 and Garfield 2002). Showing the real-world applicability of regression models encourages students to develop a more nuanced and critical understanding of statistical analysis.

Acknowledgements

Dr Christine Allen and two anonymous reviewers provided helpful suggestions on a draft of this article.

References

- Auger, N., Giraud, J. and Daniel, M. (2009). The joint influence of area income, income inequality, and immigrant density on adverse birth outcomes: a population-based study. *BMC Public Health*, **9**, 237.
- Bartley, M. (2004). *Health Inequality: An Introduction to Theories, Concepts and Methods*. Cambridge: Polity.
- Ben-Zvi, D. and Garfield, J. (2004). *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*. Netherlands: Kluwer Academic Publishers.
- Bryman, A. and Teevan, J. L. (2005). *Social Research Methods: Canadian Edition*. Don Mills: Oxford University Press.
- Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W. and Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests? *Journal of Statistics Education*, **17**(2).
- Cobb, G. W. and Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, **104**(9), 801–823.
- Coburn, D. (2004). Beyond the income inequality hypothesis: class, neo-liberalism, and health inequalities. *Social Science & Medicine*, **58**(1), 41–56.
- Craig, N. (2005). Exploring the generalisability of the association between income inequality and self-assessed health. *Social Science & Medicine*, **60**(11), 2477–2488.
- De Maio, F. G. (2010). *Health and Social Theory*. Basingstoke: Palgrave Macmillan.
- De Maio, F. G., Corber, S. J. and Joffres, M. (2008). Towards a social analysis of risk factors for chronic diseases in Latin America. *LASA Forum*, **39**(2), 10–13.
- De Maio, F. G. and Kemp, E. (2010). The deterioration of health status among immigrants to Canada. *Global Public Health*, **5**(5), 462–478.
- De Maio, F. G., Linetzky, B. and Virgolini, M. (2009). An average / deprivation / inequality (ADI) analysis of chronic disease outcomes and risk factors in Argentina. *Population Health Metrics*, **7**(8).
- Diez-Roux, A. V. (2002). A glossary for multilevel analysis. *Journal of Epidemiology and Community Health*, **56**(8), 588–594.
- Doane, D. P. and Seward, L. E. (2011). Measuring skewness: a forgotten statistic? *Journal of Statistics Education*, **19**(2), 1–18.
- Downey, L. and Hawkins, B. (2008). Race, income, and environmental inequality in the United States. *Sociological Perspectives*, **51**(4), 759–781.
- Feagin, J. R. (1999). Soul-searching in sociology: is the discipline in crisis? *The Chronicle of Higher Education*, **46**(8), B4–B6.
- Feagin, J. R. (2004). Social justice and sociology: agendas for the twenty-first century. In: W. K. Carroll (ed.) *Critical Strategies for Social Research*, pp. 29–43. Toronto: Canadian Scholars' Press Inc.
- Finney, S. J. and Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology*, **28**(2), 161–186.
- Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, **10**(3).
- Howley, P. P. (2008). Keeping it real, keeping them interested and keeping it in their minds. *Journal of Statistics Education*, **16**(1).
- Kahn, R. S., Wise, P. H., Kennedy, B. P. and Kawachi, I. (2000). State income inequality, household income, and maternal mental and physical health: cross sectional national survey. *British Medical Journal*, **321**(7272), 1311–1315.
- Landman, T. and Carvalho, E. (2009). *Measuring Human Rights*. London: Routledge.
- Mills, C. W. (1959). *The Sociological Imagination*. New York: Oxford University Press.
- Onwuegbuzie, A. J. and Leech, N. L. (2005). Taking the 'Q' out of research: teaching research methodology courses without the divide between quantitative and qualitative paradigms. *Quality and Quantity*, **39**, 267–296.
- Ross, N. A., Wolfson, M. C., Dunn, J. R., Berthelot, J. M., Kaplan, G. A. and Lynch, J. W. (2000).

- Relation between income inequality and mortality in Canada and in the United States: cross sectional assessment using census data and vital statistics. *British Medical Journal*, **320**(7239), 898–902.
- Subramanian, S. V., Kawachi, I. and Smith, G. D. (2007). Income inequality and the double burden of under- and overnutrition in India. *Journal of Epidemiology and Community Health*, **61**(9), 802–809.
- Taylor, V. (1998). Feminist methodology in social movements research. *Qualitative Sociology*, **21**(4), 357–379.
- Valero-Mora, P. M. and Ledesma, R. D. (2011). Using interactive graphics to teach multivariate data analysis to psychology majors. *Journal of Statistics Education*, **19**(1).
- Vanhoof, S., Kuppens, S., Castro Sotos, A. E., Verschaffel, L. and Onghena, P. (2011). Measuring statistics attitudes: structure of the Survey of Attitudes Toward Statistics (SATS-36). *Statistics Education Research Journal*, **10**(1), 35–51.
- Wiberg, M. (2009). Teaching statistics in integration with psychology. *Journal of Statistics Education*, **17**(1).
- Wilkinson, R. G. and Pickett, K. E. (2009). *The Spirit Level: Why More Equal Societies Almost Always Do Better*. London: Allen Lane.
- Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K. and Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the college level? a review of the literature. *Journal of Statistics Education*, **16**(2).

Copyright of Teaching Statistics is the property of Wiley-Blackwell and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.