LORE methodological note 2014:6

Auto-coding versus manual coding of most important problem

Elias Markstedt

ABSTRACT

A manual coding of the open-ended *most important problem* question is compared with a keyword auto-coding routine. This evaluation shows that 92 percent of about 11,000 responses are coded the same across 35 categories. Correlations with party preference are very similar. Auto-coding can be a good alternative to manual coding when resources are scarce. It may even be more reliable than manual coding, since the coding is performed the same way each time.

Introduction

Open-ended responses have always posed problems for researchers. Is the produced data worth the resources that are required to code every response? The proliferation of web surveys has made this question even more relevant due to the ease by which researchers can now collect large amounts of free-text responses. Manual coding is impractical in many of these situations, which is why software applications with auto-coding features are tempting alternatives. This study examines how close an auto-coding routine might come to a manual coding in the case of a common measure of issue saliency in political science, namely "What is the *most important problem* facing the country today?" (MIP). It also evaluates how auto-coded codes compare to manual codes when correlated with party preference.

Data

The data is pooled from three annual self-administered paper surveys with nationally representative samples collected by the SOM Institute at the University of Gothenburg between 2011 and 2013. Response rates (RR5) reached around 50 percent in all three surveys, while item nonresponse for this specific question is about 19 percent, which resulted in about 11,362 answers. Manual coding was carried out each year using a coding scheme with 45 main code categories and 404 sub-categories. To facilitate the auto-coding only main categories are considered, and some of the main categories are merged, which results in 35 categories. The auto-coding is performed using Wordstat (a plugin

Series editor: Johan Martinsson

program for QDA Miner), which let researchers assign keywords to different coding categories.

Results

A comparison of the manual coding and the auto-coding shows congruence in 92 percent of the cases. 6 of the 35 categories constitute 78 percent of the cases (manual: 78.0, auto: 78.4 percent), categories which are all over 5 percent of the cases each. Table 1.a–1.f show the correlations between the six most common issue codes and party preference (responses to the question "Which party do you like best today?"). Correlations do not differ much across these categories using either coding strategy; the absolute difference usually ranges between 0 and 0.02. Almost as many of the reported manual coding coefficients are significant (30) as auto-coding coefficients (31). To see whether any of the strategies produce stronger correlations, the differences between the absolute coefficients are calculated, with a zero result (0.003 stronger correlations are produced using manual coding).

To sum up, auto-coding seems to be able to serve as a good substitute to manual coding when resources are scarce. Note that the featured question, MIP, might be a type of question that is especially suited for keyword coding, since respondents usually only type in one or two words (e.g. "Schools" or "High taxes"). To validate these results, longer, essay type questions should be similarly examined. A final point is that auto-coding has a considerable advantage over manual coding when dealing with reliability. Each case is treated equally each time. Furthermore, the process is very transparent because specific keyword coding schemes may be shared among researchers.

Table 1.a. Correlations: Labor market

	Auto	Manual	Diff (abs man - abs auto)	N-Auto	N-Man
Left Party	-0.01	0.00	-0.01	2,970	2,928
Social Democrats	**0.17**	**0.17**	0.00		
Centre Party	-0.05	-0.06	0.00		
Liberal Party	**-0.05**	-0.05	-0.01		
Moderates	0.01	0.01	0.00		
Christian Democrats	-0.05	-0.05	0.00		
Green Party	**-0.12**	**-0.12**	0.00		
Sweden Democrats	**-0.19**	**-0.18**	-0.01		

Comment: The correlations are tetrachoric. Each coefficient is the correlate between two dichotomous variables. Bolded numbers are significant at the .05-level.

Table 1.b. Correlations: Education

	Auto	Manual	Diff (abs man - abs auto)	N-Auto	N-Man
Left Party	-0.05	-0.05	0.00	1,488	1,573
Social Democrats	-0.02	-0.01	-0.01		
Centre Party	0.02	0.03	0.01		
Liberal Party	**0.19**	**0.17**	-0.02		
Moderates	**0.05**	**0.05**	0.00		
Christian Democrats	-0.01	0.00	-0.01		
Green Party	-0.02	-0.02	0.00		
Sweden Democrats	**-0.23**	**-0.23**	0.00		

Table 1.c. Correlations: Health care

	Auto	Manual	Diff (abs man - abs auto)	N-Auto	N-Man
Left Party	-0.06	**-0.08**	0.03	1,527	1,458
Social Democrats	**0.14**	**0.14**	0.00		
Centre Party	**0.07**	0.07	-0.01		
Liberal Party	-0.05	-0.03	-0.02		
Moderates	-0.03	-0.04	0.01		
Christian Democrats	0.02	0.03	0.01		
Green Party	**-0.12**	**-0.10**	-0.02		
Sweden Democrats	**-0.10**	**-0.11**	0.01		

Table 1.d. Correlations: Integration

	Auto	Manual	Diff (abs man - abs auto)	N-Auto	N-Man
Left Party	**-0.08**	**-0.09**	0.01	1,205	1,189
Social Democrats	**-0.25**	**-0.25**	0.00		
Centre Party	**-0.13**	**-0.15**	0.02		
Liberal Party	**-0.08**	**-0.09**	0.01		
Moderates	**-0.06**	**-0.06**	-0.01		
Christian Democrats	**-0.10**	**-0.10**	0.01		
Green Party	**-0.08**	**-0.09**	0.01		
Sweden Democrats	**0.64**	**0.64**	0.00		

Table 1.e. Correlations: Environment

	Auto	Manual	Diff (abs man - abs auto)	N-Auto	N-Man
Left Party	0.06	0.06	0.00	890	904
Social Democrats	**-0.20**	**-0.19**	-0.01		
Centre Party	**0.09**	**0.08**	0.00		
Liberal Party	**-0.09**	**-0.09**	0.00		
Moderates	**-0.19**	**-0.19**	0.00		
Christian Democrats	0.04	0.03	-0.01		
Green Party	**0.50**	**0.50**	0.00		
Sweden Democrats	**-0.29**	**-0.29**	0.00		

Table 1.f. Correlations: Economy

	Auto	Manual	Diff (abs man - abs auto)	N-Auto	N-Man
Left Party	**-0.13**	**-0.14**	0.02	795	770
Social Democrats	**-0.24**	**-0.24**	0.01		
Centre Party	-0.01	-0.03	0.02		
Liberal Party	**0.12**	**0.14**	0.02		
Moderates	**0.31**	**0.33**	0.02		
Christian Democrats	0.01	-0.05	0.06		
Green Party	**-0.16**	**-0.18**	0.02		
Sweden Democrats	**-0.18**	**-0.19**	0.01		

The Laboratory of Opinion Research (LORE) is an academic web survey center located at the Department of Political Science at the University of Gothenburg. LORE was established in 2010 as part of an initiative to strengthen multidisciplinary research on opinion and democracy. The objective of the Laboratory of Opinion Research is to facilitate for social scientists to conduct web survey experiments, collect panel data, and to contribute to methodological development. For more information, please contact us at:

info@lore.gu.se