Influential Cases in Multilevel Modeling : A Methodological Comment

Tom Van der Meer, Manfred Te Grotenhuis and Ben Pelzer

American Sociological Review 2010 75: 173
DOI: 10.1177/0003122409359166

American Sociological Review
75(1) 173–178

DOI: 10.1177/0003122409359166
http://asr.sagepub.com

Influential Cases in Multilevel Modeling: A Methodological Comment

Tom Van der Meer,[a] Manfred Te Grotenhuis,[b] and Ben Pelzer[b]

A large number of cross-national survey datasets have become available in recent decades. Consequently, scholars frequently apply multilevel models to test hypotheses on both the individual and the country level. However, no currently available cross-national survey project covers more than 54 countries (GESIS 2009). Multilevel modeling therefore runs the risk that higher-level slope estimates (and the substantial conclusions drawn from these estimates) are unreliable due to one or more influential cases (i.e., countries).

This comment emphasizes the problem of influential cases and presents ways to detect and deal with them. To detect influential cases, one may use both graphic tools (e.g., scatter plots at the aggregate level) and numeric tools (e.g., diagnostic tests such as Cook's D and DFBETAS). To illustrate the usefulness and necessity of these tools, we apply them to a study that was recently published in this journal (Ruiter and De Graaf 2006). Finally, we provide recommendations and tools to detect and handle influential cases, specifically in cross-sectional multilevel analyses.

A CROSS-NATIONAL STUDY ON THE EFFECT OF RELIGION

In "National Context, Religiosity, and Volunteering: Results from 53 Countries," Ruiter and De Graaf (2006) raise the following question: To what extent do national religious contexts affect volunteering? One of their central hypotheses states that volunteer rates will be higher in devout countries than in secular countries. This hypothesis originates from two previous findings. First, religious citizens are more likely than nonreligious citizens to volunteer (Wilson and Musick 1997). Second, in devout societies, citizens are more likely to have active church members in their social networks (Kelley and De Graaf 1997). Because pro-civic norms and recruitment are more widespread, due to a higher share of religious citizens in social networks, the authors expect to find a positive effect of countries' degree of devoutness on individual volunteering for both religious and nonreligious citizens.

To test this hypothesis, Ruiter and De Graaf (2006) applied a hierarchical 3-level model to three waves of the European/World Values Survey (WVS): individuals at level 1 (N = 117,007), surveys from three waves at level 2 (N = 96), and countries at level 3 (N = 53). A crucial step in their test of the network explanation is the inclusion of a level-2 characteristic, namely country's average church attendance rate. This enabled them to test whether devout societies induce their citizens to volunteer more often

[a]University of Amsterdam/Netherlands Institute for Social Research
[b]Radboud University Nijmegen

Corresponding Author:
Tom Van der Meer, Department of Political Science, University of Amsterdam, OZ Achterburgwal 237, 1012 DL Amsterdam, The Netherlands
E-mail: t.w.g.vandermeer@uva.nl

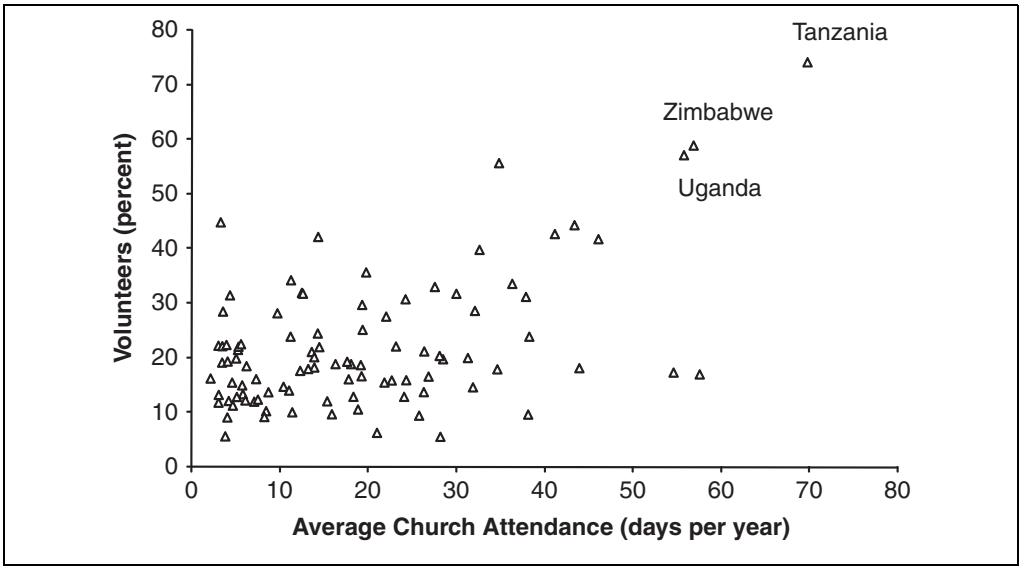

Figure 1. Scatter Plot for Average Church Attendance and Percentage Volunteers, in 96 Surveys Conducted in 53 Countries during Three Waves

than do citizens in secular societies. Ruiter and De Graaf found average church attendance to be significantly and positively related to volunteering.

GRAPHIC EVIDENCE FROM A SCATTER PLOT

To test whether their findings were robust, Ruiter and De Graaf (2006) re-estimated their model 96 times, leaving each survey out once, and compared the resulting estimates with those from the original model (these differences are known as DFBETA). Based on these comparisons, they found no influential cases. However, this method will most likely fail to detect a cluster of two or more influential cases that have a similar influence on the estimates. Furthermore, because DFBETA lacks standardization, it is hard to tell how large a difference should be to call a case too influential.

Because Ruiter and De Graaf (2006) were interested in the contextual effect of average church attendance, we will look for potential cases at level 2 that influence this effect in an undesirable way. To get some general clues about potential influential cases, we first inspect the bivariate scatter plot for volunteer rates and average church attendance for all 96 surveys (Belsley, Kuh, and Welsch 1980:8).

Figure 1 indicates a positive association between average church attendance and volunteering rates. However, it also reveals a cluster of three potentially influential cases: Tanzania, Zimbabwe, and Uganda. These countries are very devout and show high volunteering rates. Notably, they are three of the four sub-Saharan countries in the dataset, collected during the third survey-wave of the WVS. Exclusion of one of these three surveys does not affect the OLS regression slope estimate substantially. Simultaneous exclusion of Tanzania, Zimbabwe, and Uganda, however, causes the slope estimate to drop from .43 to .23.

NUMERIC EVIDENCE FROM A MULTIVARIATE HIERARCHICAL MODEL

Although the scatter plot is a good first indicator of influential cases, it is based on

Table 1. The Effect of Average Church Attendance after Eliminating Influential Cases

	Model 0[a]	Model 1	Model 2	Model 3
To neutralize their influence at the contextual level, the model includes dummies for:		Tanzania	Tanzania Zimbabwe	Tanzania Zimbabwe Uganda
Effect of average church attendance	.018 (.005)***	.014 (.005)***	.010 (.005)**	.007 (.006)
Highest DFBETAS[b]	.936	.560	.644	–.583
Survey with highest DFBETAS[c]	Tanzania	Zimbabwe	Uganda	Russia
Corresponding Cook's D[d]	.306	.185	.217	.247

Note: Standard errors are in parentheses. Estimates for all variables in models are available at http://www.ru.nl/mt/ic/downloads/.
[a]Model 0 replicates Model 4 in Ruiter and De Graaf (2006).
[b]All DFBETAS values exceed $2/\sqrt{n_2}$ (i.e., $2/\sqrt{96} = .204$ for Model 0; $2/\sqrt{95} = .205$ for Model 1; $2/\sqrt{94} = .206$ for Model 2; $2/\sqrt{93} = .207$ for Model 3), where n_2 represents the number of surveys minus the number of survey dummies.
[c]The surveys from Tanzania, Zimbabwe, and Uganda are from Wave 3. The survey from Russia is from Wave 2.
[d]Cook's D's in models 0 to 2 exceed $4/n_2$ (i.e., $4/96 = .0417$ for Model 0; $4/95 = .0421$ for Model 1; $4/94 = .0425$ for Model 2; $4/93 = .043$ for Model 3), where n_2 represents the number of surveys minus the number of survey dummies.
$^{**}p < .025$; $^{***}p < .001$ (one-tailed tests).

aggregated data that lacks individual and contextual control factors. The proof of the pudding is in multilevel, multivariate models.

In Table 1, Model 0, the random slope model reported is virtually equal to Ruiter and De Graaf's (2006) Table 3, Model 4 (p. 201).[1] Model 0 shows the positive and significant effect of average church attendance that they found. Next, we compute two diagnostics to detect influential cases for all 96 surveys at level 2: Cook's D and DFBETAS. Cook's D measures the influence of one single case on all (or a subset of) level-2 estimates in the model, whereas DFBETAS measures a case's influence on each of the level-2 estimates separately.

Cook's D is defined as:

$$D_j = \frac{1}{r}(\hat{\beta} - \hat{\beta}_{(-j)})' \hat{S}^{-1}_{(-j)} (\hat{\beta} - \hat{\beta}_{(-j)}) \quad (1)$$

where r = number of fixed parameters, $\hat{\beta}$ = vector of estimates based on the full sample, $\hat{\beta}_{(-j)}$ = vector of estimates after unit j is excluded, and $\hat{S}_{(-j)}$ denotes the covariance matrix after unit j is excluded (Snijders and Berkhof 2008:158 [3.24]). Cook's D can be interpreted as the standardized average squared difference between the estimates with and without unit j.

DFBETAS is defined as:

$$\text{DFBETAS}_{jZ} = \frac{\hat{b}_Z - \hat{b}_{-jZ}}{se(\hat{b}_{-jZ})} \quad (2)$$

where $\hat{b}_Z - \hat{b}_{-jZ}$ represents the difference between the slope estimate $\hat{b}_Z$ of predictor Z based on the full sample and the estimate $\hat{b}_{-jZ}$ after excluding unit j, and $se(\hat{b}_{-jZ})$ denotes the standard error of $\hat{b}_{-jZ}$. Equation 2 is analogous to Belsley and colleagues (1980:13). One can interpret DFBETAS as the standardized difference between the estimate with and without unit j.[2]

To decide which cases are too influential, Belsley and colleagues (1980:28) propose using $4/n_x$ as the cutoff value for Cook's D, and $2/\sqrt{n_x}$ for the absolute value of DFBETAS (where n_x = number of units at level x).

Although subscript –*j* in Equations 1 and 2 suggests that all individuals of case *j* are effectively deleted, this is not the preferred option because it would mean deletion of all lower level units (i.e., individuals) nested in the influential higher-level unit (i.e., survey), thereby losing statistical power. Instead, we eliminate the *influence* of survey *j* from the slope estimate of each level-2 predictor and from the level-2 variance, but we maintain the individuals of survey *j* to estimate the level-1 parameters (Langford and Lewis 1998:125). We do this by including a fixed-effect dummy variable in the model (taking value 1 for individuals of survey *j* and value 0 for all others) and changing the intercept vector of the model to value 0 for individuals from survey *j*.

Ruiter and De Graaf (2006) were primarily interested in the level-2 effect of average church attendance, so we are mainly concerned with that effect as well. DFBETAS is especially useful for this purpose because it measures the extent to which cases influence a specific slope estimate (i.e., the effect of average church attendance). Were we interested in the *combined* influence on all (or a subset of) level-2 estimates, Cook's D might be more appropriate. For illustrative purposes, we will present both Cook's D and DFBETAS below.[3] In a first run, Tanzania turns out to have both the highest DFBETAS (.936) and the highest Cook's D (.306). To eliminate Tanzania's strong influence on the level-2 effects, we include a fixed-effect dummy variable at level 2 and exclude Tanzania from the intercept. Consequently, the effect of average church attendance drops from .018 to .014. However, single tests may not suffice when there is a cluster of influential cases. We therefore compute Cook's D and DFBETAS for all remaining 95 surveys in a second run. This time, Zimbabwe has the highest DFBETAS on average church attendance (.560), while its Cook's D (.185) is second to Russia. Excluding Zimbabwe's influence lowers the effect of average church attendance from .014 to .010. In a third run, Uganda has the highest DFBETAS on average church attendance (.644) and the second highest Cook's D, again second to Russia.[4] After eliminating Uganda's influence, the level-2 effect of average church attendance is no longer significant. In short, Tanzania, Zimbabwe, and Uganda (together with Russia)[5] are confirmed to be influential cases.

Theoretically, this challenges the network explanation offered by Ruiter and De Graaf (2006).[6] Methodologically, this illustrates that excluding a single survey and comparing absolute differences between estimates (DFBETAS) does not suffice. To signal a cluster of influential cases, repeated tests of DFBETAS or Cook's D are necessary.

EXPLANATIONS FOR INFLUENTIAL COUNTRIES

Tanzania's, Zimbabwe's, and Uganda's average scores on volunteering and church attendance are, by far, the highest among all 96 surveys. It is therefore legitimate to question the validity of these figures. Measures of volunteering may not be cross-culturally equivalent and there is no hard evidence to support the validity of the WVS data in these three countries (Govaart et al. 2001). Moreover, even if the data are valid and cross-culturally equivalent, these countries are too influential. From a strictly methodological point of view, that is reason enough to neutralize their influence on the estimates. An analysis in which a small number of cases determine the outcomes does not offer a satisfactory test of a theory—especially when a theory is unconditional and influential cases are grouped geographically, as is the case here.

Rather, influential cases offer a good starting point for theory refinement. By studying influential cases more closely, scholars may specify cultural or institutional conditions under which a theory holds. Post-hoc, in-depth studies may offer more conclusive remarks on both the country-specific validity of quantitative measures in cross-national

surveys and the micro-level mechanisms that operate. With regard to religiosity and volunteering, studies of other sub-Saharan countries suggest that volunteering is primarily a form of reciprocal support, necessitated by economic uncertainty due to decolonization and stimulated by the church (Govaart et al. 2001).

RECOMMENDATIONS AND CONCLUSIONS

This comment signals the problem of influential cases in cross-national multilevel research. With the increased use of cross-national survey data, this age-old problem is of renewed importance in assessing the reliability of slope estimates. Below, we formulate some recommendations to detect and handle influential cases.

As a first step, bivariate scatter plots are highly useful in detecting possible influential cases (Belsley et al. 1980:8). Partial residual/regression plots (not shown here) may be even more instructive as they take into account potential confounding effects (Belsley et al. 1980:30).

DFBETAS and Cook's D should be used as numeric diagnostic tests. DFBETAS is most useful in evaluating the reliability of specific estimates separately. Cook's D is more suitable when evaluating the reliability of all (or subsets of) higher-level estimates simultaneously.

If absolute values of DFBETAS exceed the cutoff value $2/\sqrt{n_x}$, or Cook's D exceeds $4/n_x$, the case should be considered too influential (Belsley et al. 1980:28).

Although the cutoff values are useful rules of thumb for detecting influential cases, they are not the only stop criteria. According to Belsley and colleagues (1980:29), a gap between subsequent absolute values of DFBETAS is an additional suitable stop criterion. Moreover, we advise caution when DFBETAS or Cook's D reveals many influential cases. This may point to a misspecification of the model; for example, a nonlinear relationship is not modeled adequately. This should not lead to the elimination of all influential cases, but to a better specification of the model.

To detect and handle influential cases statistically, one must eliminate their impact. We do not advocate for deletion of higher-level influential cases altogether in multilevel analysis, because that would lead to a loss in statistical power. Instead, researchers should include fixed-effect dummy variables at higher levels and adapt the intercept vector for individuals within the influential higher-level units (Langford and Lewis 1998:125).

A single run of diagnostic tests will not suffice when there is a cluster of outliers. Repeated runs are required to assess the reliability of the estimates. For reasons of parsimony, iterative elimination of cases with the highest DFBETAS or Cook's D is preferable to elimination of all influential cases in a single step.

To run these diagnostic tests in multilevel models and eliminate the impact of influential cases, we offer scripts on our Web page for the MLwiN and R packages (http://www.ru.nl/mt/ic/downloads/).

In summary, influential cases are a potential threat to every study with a limited set of observations. This includes state-of-the-art multilevel studies with a relatively small number of observations at higher levels. Single tests to uncover influential cases may not suffice. To detect a cluster of influential cases, repeated runs of DFBETAS or Cook's D are required.

Acknowledgments

We would like to thank the anonymous reviewers for their useful comments. We are grateful to Rense Nieuwenhuis who translated our MLwiN macro to detect influential cases in multilevel models into the R-module 'Influence.ME.'

Notes

1. Model 0 replicates Model 4 in Ruiter and De Graaf (2006). Our estimates differ slightly from those of Ruiter and De Graaf due to the use of a different software package. We estimated our models in MLwiN 2.02 (PQL, 1st order), whereas Ruiter and De Graaf

used HLM. These very small differences are not relevant for the diagnostic tests.
2. Note that squared DFBETAS = Cook's D when only one parameter is considered.
3. We calculate Cook's D from the estimates for all three level-2 predictors in Model 4 in Ruiter and De Graaf (2006).
4. Russia has the highest Cook's D (i.e., the strongest overall influence) in Models 1 through 3. Along with rather high DFBETAS on average church attendance, Russia exceeds the cutoff value for DFBETAS on level of democracy, which is not of central interest here.
5. In the fourth and fifth runs, the two survey-waves in Russia had the highest DFBETAS on average church attendance (–.583 and .710). After eliminating both their influences, the effect of average church attendance remains nonsignificant (b = .0063, s.e. = .0057). Furthermore, there were no big gaps anymore between the absolute values of DFBETAS (for details, see http://www.ru.nl/mt/ic/downloads/).
6. The network theory might still hold *within* countries with a common religious tradition (Borgonovi 2008).

References

Belsley, David A., Edwin Kuh, and Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley.

Borgonovi, Francesca. 2008. "Divided We Stand, United We Fall: Religious Pluralism, Giving, and Volunteering." *American Sociological Review* 73:105–28.

GESIS. 2009. *Tabular History of International Comparative Survey Research Projects*. Compiled March 29, 2009 (http://www.gesis.org/en/services/data/portals-links/comparative-survey-projects/#%291).

Govaart, Margriet-Marie, Henk J. Van Daal, Angelika Münz, and Jolanda Keesom. 2001. *Volunteering Worldwide*. Utrecht, Netherlands: NIZW.

Kelley, Jonathan and Nan Dirk De Graaf. 1997. "National Context, Parental Socialization, and Religious Belief: Results from 15 Nations." *American Sociological Review* 62:639–59.

Langford, Ian H. and T. Lewis. 1998. "Outliers in Multilevel Data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 161:121–60.

Ruiter, Stijn and Nan Dirk De Graaf. 2006. "National Context, Religiosity, and Volunteering: Results from 53 Countries." *American Sociological Review* 71:191–210.

Snijders, Tom A. B. and Johannes Berkhof. 2008. "Diagnostic Checks for Multilevel Models." Pp. 457–514 in *Handbook of Multilevel Analysis*, edited by J. De Leeuw and E. Meijer. New York: Springer.

Wilson, John and Marc Musick. 1997. "Who Cares? Toward an Integrated Theory of Volunteer Work." *American Sociological Review* 62:694–713.

Tom Van der Meer is a researcher at the Netherlands Institute for Social Research (SCP) and at the Department of Social Science Research Methodology of the Radboud University of Nijmegen. He is an affiliate of the Interuniversity Centre for Social Science Theory and Methodology (ICS). He studies the association between state institutions, networks of participation, and trust relationships. Recent publications include articles in *Comparative Political Studies*, *European Journal of Political Research*, *European Sociological Review*, *European Societies*, and *Scandinavian Political Studies*.

Manfred Te Grotenhuis is an assistant professor of methodology at Radboud University of Nijmegen, the Netherlands, and an affiliate of the Interuniversity Centre for Social Science Theory and Methodology (ICS). His main interests are longitudinal data analysis, simulation techniques, and age-period-cohort models. Recent publications include articles in *Acta Psychiatrica Scandinavica*, *American Journal of Sociology*, *European Sociological Review*, *European Societies*, *Journal for the Scientific Study of Religion*, *Sociology of Religion*, *Review of Religious Research*, and *Psychiatric Services*.

Ben Pelzer is an assistant professor of quantitative research methods at the Radboud University of Nijmegen, the Netherlands. His main interests are (repeated) cross-sectional data analysis, multilevel data analysis, nominal data analysis, and causal models. Recent publications include articles in *Clinical Trials*, *Political Analysis*, *Quality & Quantity*, and *Statistica Neerlandica*.