Catalogue no. 12-002-X — No. 2015001
ISSN 1710-2197

The Research Data Centres — Information and Technical Bulletin

Assessing the impact of potentially influential observations in weighted logistic regression

by Bridget L. Ryan, John Koval, Bradley Corbett, Amardeep Thind, M. Karen Campbell, and Moira Stewart

Release date: March 24, 2015

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

About the Research Data Centres Information and Technical Bulletin

The Research Data Centres Information and Technical Bulletin is a forum for current and prospective users of the centres to exchange practical information and techniques for analyzing datasets available at the centres. The bulletin is published twice per year, in the spring and fall. Additional special issues on timely topics may also be released on an occasional basis.

Aims

The main aims of the bulletin are:

- to advance and disseminate knowledge surrounding Statistics Canada's data;
- to exchange ideas among the Research Data Centres (RDC) user community;
- to support new users of the RDC program; and
- to provide an additional means through which RDC users and subject matter experts and divisions within Statistics Canada can communicate.

Content

The Research Data Centres Information and Technical Bulletin is interested in receiving articles and notes that will add value to the quality of research produced at the Statistics Canada RDCs and provide methodological support to RDC users.

Topics include, but are not limited to:

- data analysis and modeling;
- data management;
- best or ineffective statistical, computational, and scientific practices;
- data content;
- implications of questionnaire wording;
- comparisons of data sets;
- reviews on methodologies and their applications;
- problem-solving analytical techniques; and
- explanations of innovative tools, using surveys and relevant software available at the RDCs.

Those interested in submitting an article to The Research Data Centres Information and Technical Bulletin are asked to refer to the **Instructions for authors**.

The editors and authors would like to thank the reviewers for their valuable comments.

Author information

John Koval, M. Karen Campbell

Department of Epidemiology and Biostatistics
The University of Western Ontario
London, Ontario

Bradley Corbett

Statistic Canada, Microdata Access Division, Research Data Centres Program
University of Western Ontario Research Data Centre
The University of Western Ontario
London, Ontario

Amardeep Thind, Moira Stewart, Bridget L. Ryan

Departments of Family Medicine, and Epidemiology and Biostatistics
The University of Western Ontario
London, Ontario

Amardeep Thind

Schulich Interfaculty Program in Public Health
The University of Western Ontario
London, Ontario

Table of Contents

Assessing the impact of potentially influential observations in weighted logistic regression

Abstract

Influential observations in logistic regression are those that have a notable effect on certain aspects of the model fit. Large sample size alone does not eliminate this concern; it is still important to examine potentially influential observations, especially in complex survey data. This paper describes a straightforward algorithm for examining potentially influential observations in complex survey data using SAS software. This algorithm was applied in a study using the 2005 Canadian Community Health Survey that examined factors associated with family physician utilization for adolescents.

Introduction

Influential observations in logistic regression can be characterized as those observations that have a notable effect on certain aspects of the fit of the linear logistic model, such as the parameter estimates or fit statistics. Collett (2003) and Hosmer and Lemeshow (2000) provide extensive explanations about the identification of influential observations in the case of classical logistic regression. The use of datasets with large sample sizes (e.g. Statistics Canada survey data) is thought to mitigate concerns about potentially influential observations by minimizing the contribution of any given observation. However, influential observations can still arise in these large samples. For example, observations may exert influence if the observations have large weights resulting in a large impact on parameter estimates (Macnab et al., 2005). Therefore, it is important to identify potentially influential observations when conducting logistic regression using Statistics Canada data. Few papers contain information about influence diagnostics particularly for complex survey data (with Roberts, Rao and Kumar (Roberts et al., 1987) being one of these); unfortunately, the diagnostics developed in these papers are not available in any of the common statistical packages for analysis of complex survey data. However, Heeringa et al. (2010, p. 245), for example, recommend the following:

“Use one or more of the techniques described in Chapter 5 of Hosmer and Lemeshow (2000) to evaluate the fit of the model for individual patterns of covariates. If the complex sample logistic regression modeling program in your chosen software system (e.g., SAS PROC SURVEYLOGISTIC) does not include the full set of diagnostic capabilities of the standard programs, use standard programs (e.g., SAS PROC LOGISTIC) with a weight specification. As mentioned before, the weighted estimates of parameters and predicted probabilities will be identical, and serious breakdowns in the model for specific covariate patterns should be identifiable even though the standard program does not correctly reflect the variances and covariances of the parameter estimates given the complex sample design”.

This paper seeks to implement this recommendation for diagnostics for coefficient sensitivity by describing a straightforward algorithm and code for examining potentially influential observations with weighted data using SAS software (SAS Institute Inc., 2009).

Data and methods

Data source and sample

The algorithm and code described in this paper was applied in a study that examined the factors associated with family physician utilization for adolescents in Canada (Ryan et al., 2011). The study employed a cross-sectional design, conducting a secondary analysis of data for adolescent and young adult respondents to the 2005 (Cycle 3.1) Canadian Community Health Survey (CCHS) (Statistics Canada, questionnaire, 2005; Statistics Canada, User’s Guide, 2005). The sample sizes for the study were 4985 respondents for early adolescents (12 to 14 years old); 8718 for middle adolescents (15 to 19 years old); and 6681 for young adults (20 to 24 years old).

Permission was received from the Statistics Canada Research Data Centre (RDC) to access these data at The University of Western Ontario. Approval from The University of Western Ontario Health Sciences Research Ethics Board was not required because this was a secondary analysis of survey data with no possibility of identification of individual survey respondents.

Study analysis

The full study analysis, described elsewhere (Ryan et al., 2011) and summarized here, was conducted separately for each of the three age groups: early adolescence, middle adolescence; and young adulthood. Two logistic regressions were conducted for each age group resulting in a total of six regressions. Analysis used design-based software employing sampling weights to adjust the sample for the unequal probability of selection and bootstrapping to adjust the confidence intervals for the complex survey design effect. The binary outcome for the first regression was whether or not the adolescent had used family physician services within the last 12 months. Within those respondents who had used services, the outcome of the second logistic regression was whether the adolescent was a high user (4 or more visits) or a low user (1 to 3 visits). The independent variables were chosen according to Andersen's Behavioral Model of Health Services Use (Andersen, 1995). Wherever possible, the same variables were used for each of the three age groups to facilitate comparison across groups, and non-significant variables were left in the models to facilitate reporting across the age groups. Predisposing variables available and used were: age, sex, school attendance and educational attainment, ethnicity, community belonging, marital status (young adults), and work status (middle adolescents and young adults). Enabling variables used were: household income adequacy, living arrangement (young adults), having a regular medical doctor, and geography (urban or rural). Perceived need variables were: self-perceived general health, self-perceived mental health, opinion of own weight, and stress (available for middle adolescents and young adults only). Evaluated need variables were: BMI category, and the number of chronic conditions. Health practice variables used were: physical activity, smoking, sexual activity (available for middle adolescents and young adults only), and alcohol drinking. The CCHS does not provide health care system or external environment variables; however, province was used as a measure of context.

Identifying potentially influential observations

In the full study, each of the six logistic regression models were evaluated to determine whether any observations in each dataset had an undue influence on the parameter estimates from the logistic regression. The identification of potentially influential observations was conducted in SAS Version 9.1 (SAS Institute Inc., 2009). SAS PROC LOGISTIC will fit a logistic model using weights and can produce several of the diagnostic influence statistics described in Hosmer and Lemeshow (2000). While these statistics do not appropriately take all of the survey design into account (such as in how variance estimates are made), and it is too unwieldy to plot the values of these statistics for every data point (due to large sample size) they can still be useful in allowing the researcher to identify cases that have potentially undue influence on the parameter estimates using the weights in Statistics Canada survey data. It should be noted that currently SAS Version 9.3 provides PROC SURVEY LOGISTIC; however, the required diagnostic statistics are not available.

The examination of potentially influential observations focused on two main statistics, the confidence interval displacement diagnostic (C diagnostic) and the DFbeta diagnostic as suggested by Pregibon (1981). These statistics were calculated and output into separate datasets using commands that followed the logistic regression command. Appendix 1 contains the algorithm that was followed for the identification and examination of the potentially influential observations. Appendix 2 contains the SAS code that was used to identify the potentially influential observations. It should be noted that the formulas for calculating the C statistics and the DFbetas each contain variance elements which ideally should be estimated using the bootstrap method. As mentioned above, SAS will estimate the model using bootstrapping; however, it cannot calculate these statistics with bootstrapping.

The 'confidence interval displacement diagnostic' provides scalar measures of the influence of individual observations on the logistic regression parameter estimates. A scalar measure is one that provides a measure of the magnitude of the influence on estimates but not the direction of that influence. One C statistic is calculated for each observation

for the overall logistic regression. The C diagnostic is based on the same principle as the Cook distance in linear regression theory (SAS Institute Inc., 2009). Observations that have a C statistic greater than one are generally considered as possible influential observations (Hosmer, 2000, p.180). However, given that the variance has been estimated without appropriate bootstrapping to account for the design effect, the use of the suggested cut-off values must be employed with caution. It is important to examine any unusually large values as an indication of potential influence (Hosmer, 2000). Therefore, the code also includes PROC UNIVARIATE which will print out the five lowest and highest values regardless of absolute size.

The formula for the C statistic used by SAS (SAS Institute, 2008) is listed below. It is based on that developed by Pregibon (1981) but was modified specifically for logistic regression:

$$C_j = \chi_j^2 h_{jj}/(1-h_{jj})^2 \quad (1)$$

where

$$\chi_j^2 = \frac{w_j (r_j - p_j)^2}{p_j q_j}, \quad (2)$$

and

$$h_{jj} = w_j p_j q_j \left(1\ x_j'\right) \mathbf{V}\left(\hat{\mathbf{b}}\right) \binom{1}{x_j}. \quad (3)$$

Moreover,

r_j is the response (0 or 1),

w_j is the weight of the jth observation,

π_j is the probability of a response for the jth observation which is given by

$\pi_j = F(\beta_0 + \beta' x_j)$, where $F(\cdot)$ is the inverse link function,

$\mathbf{b}$ is the maximum likelihood estimate (MLE) of $(\beta_0\ \beta_1 \ldots \beta_s)'$,

s is the number of variables,

$\widehat{\mathbf{V}}_{\mathbf{b}}$. is the estimated covariance matrix of $\mathbf{b}$,

p_j is the estimate of π_j evaluated at $\mathbf{b}$,

and $q_j = 1 - p_j$.

A limitation of the C statistic is that it is a summary measure of change over all the coefficients in the model. Therefore, it is important to examine the changes in the individual coefficients (Hosmer, 2000, p. 181). The DFbeta is the standardized difference in the parameter estimate due to deleting each given observation. DFbetas are useful in detecting observations that are causing changes in coefficients (SAS Institute Inc., 2009). The underlying distribution of the DFbetas is unknown so there is no certain determination of what constitutes 'large'. The convention, therefore, is to use the value of 2 which coincides approximately with the usual critical value of the normal distribution (1.96). For any given variable, then, observations that have a DFbeta greater than two are considered as possible influential observations. As with the C statistic, the standard error has been estimated without appropriate bootstrapping, so the use of the suggested cut-off values must again be employed with caution. It is important to examine any unusually large values as an indication of potential influence. Therefore, the SAS code also includes PROC UNIVARIATE which will print out the five lowest and highest values regardless of absolute size.

Possible influential observations were identified using the following formula given by SAS (SAS Institute Inc., 2008) (developed by Pregibon, 1981).

$$DFbeta_{ij} = \frac{\Delta_i \mathbf{b}_j^1}{\sigma_i},\ i=0,\ 1,\ \ldots,s, \qquad (4)$$

Where

σ_i is the standard error of the ith component of $\mathbf{b}$,

$\Delta_i b_j^1$ is the ith component of the one-step difference, and

$$\mathbf{\Delta b_j^1} = \left(\frac{w_j(r_j - p_j)}{1 - h_{jj}}\right) \widehat{\boldsymbol{V}}_b \begin{pmatrix} 1 \\ x_j \end{pmatrix}. \qquad (5)$$

In other words, Δb_j^1 is an approximation to the change, b - b_j^1, in the vector of parameter estimates due to the omission of the jth observation.

Assessing potentially influential observations

After identifying potentially influential observations, the next step was to run logistic regressions excluding all observations identified as potentially influential by either statistic. Parameter estimates were compared between the regression with all cases and the regression without the potentially influential observations. Researchers must decide how large a change in parameter estimates is considered important for any given study (Rothman, 1998). Changes in parameter estimates of more than 10% were considered to be substantial changes for this study. In the case of substantial parameter changes, the observations should be examined carefully to determine if there might be any common covariance patterns associated with the influential observations.

Researchers must decide whether these observations are part of the study population or not. If they cannot be deemed outliers, and are in fact part of the study population, they should stay in the model.

Results

While all six regressions from the full study were examined to identify potentially influential observations, only one of these is reported here for illustration; the regression for the young adult age group with the outcome of whether or not the respondent had used family physician services. Appendix 3 provides an annotated example of output for the C statistics. Perturbation was used to alter the observation numbers and C statistics to protect confidentiality.

For the C statistic, eleven cases were identified with large C statistics. This suggests possible influential observations and warrants further investigations. The DFbetas were then reviewed and no cases had large DFbetas for any variable. The lack of cases with large DFbetas suggests that there were no potentially influential observations causing undue instability in the parameter estimates. The eleven potentially influential observations based on the C statistics were removed and the logistic regression was run again. Three non-significant parameters changed by greater than 10%; however, none of these changed from non-significant to significant in the regression. Therefore, the decision was made to include all cases in the reported regression model.

Discussion

This paper provides an algorithm and SAS code that can be easily applied to analyses using complex survey data such as the Canadian Community Health Survey in order to identify possible influential observations in logistic regression models. Caution is advised in using automatic cut-off values for identifying potentially influential observations. Kleinbaum, Kupper, and Muller (1988, p. 201) state that "some observation must be the most extreme in every sample. It would be silly to delete automatically this most extreme observation, or some cluster of extreme observations, based on statistical testing procedures. The goal of regression diagnostics in evaluating outliers is to warn the data analyst to examine more closely such extreme observations. Scientific judgment is more important here than statistical tests, once influential observations have been flagged". Rather the researcher must make a decision on handling potentially influential observations informed by knowledge of the study population and a careful examination of the data as described herein. Having made this decision, this should be reported in the results section of a manuscript and the discussion should explain the reasoning and the possible effects of the decision.

References

Andersen RM. 1995. “Revisiting the behavioral model and access to medical care: does it matter?”. J Health Soc Behav. Vol. 36. no. 1, 1-10.

Collett D. 2003. Modelling Binary Data. [2nd ed]. Boca Raton, FL: Chapan & Hall/CRC Press.

Gagné, C., Roberts, G., and Keown, L.A. 2014. Weighted estimation and bootstrap variance estimation for analyzing survey data: How to implement in selected software. The Research Data Centres Information and Technical Bulletin, (Winter) Vol. 6 no. 1, 5-70. Statistics Canada Catalogue no. 12-002-X

Heeringa S, West B, Berglund P. 2010. Applied Survey Data Analysis. Boca Raton, FL: Chapman and Hall/CRC Press.

Hosmer DW, Lemeshow S. 2000. Applied Logistic Regression. [2nd ed]. New York: Wiley.

Kleinbaum DG, Kupper LL, Muller KE. 1988. Applied Regression Analysis and Other Multivariable Methods. [2nd ed]. Belmont,CA: Duxbury Press.

Macnab JJ, Koval JJ, Speechley KN, Campbell MK. 2005. “Influential observations in weighted analyses: examples from the National Longitudinal Survey of Children and Youth (NLSCY)”. Chronic Dis Can 2005; Vol. 26, no. 1, 1-8. See www.ncbi.nlm.nih.gov/pubmed/16117839. (accessed January 31, 2012).

Pregibon D. 1981. Logistic Regression Diagnostics. Annals of Statistics. Vol. 9. no. 4, 705-724.

Roberts G, Rao JNK, Kumar S. 1987. “Logistic-Regression Analysis of Sample Survey Data”. Biometrika. Vol. 74, no. 1, 1-12. See biomet.oxfordjournals.org/content/74/1/1.full.pdf. (accessed January 31, 2012).

Rothman KJ, Greenland S. 1998. Modern Epidemiology. [2nd ed]. Philadelphi, PA: Lippincott Williams & Wilkins.

Ryan BL, Stewart M, Campbell MK, Koval J, Thind A. 2011. “Understanding adolescent and young adult use of family physician services: a cross-sectional analysis of the Canadian Community Health Survey”. BMC Family Practice. Vol. 12, no. 118, 1-10.

SAS Institute Inc. 2009. SAS/Stat 9.1 Software. Cary, NC.

SAS Institute Inc. 2008. Regression Diagnostics, SAS 9.1 Online Documentation. (Path - SAS/STAT; SAS/STAT User’s Guide; The Logistic Procedure; Details; Regression Diagnostics).

Statistics Canada. 2005. Canadian Community Health Survey (CCHS) Cycle 3.1. (questionnaire). www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SurvId=1630&InstaId=22642&SDDS=3226 [Accessed from: URL: www23.statcan.gc.ca/imdb-bmdi/instrument/3226_Q1_V3-eng.pdf]

Statistics Canada. 2005. Canadian Community Health Survey (CCHS) Cycle 3.1. User’s Guide. www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SurvId=1630&InstaId=22642&SDDS=3226 [Accessed from: URL: www23.statcan.gc.ca/imdb-bmdi/document/3226_D7_T9_V3-eng.pdf]

Appendices

Appendix 1 – Algorithm for assessing potentially influential observations in weighted logistic regression in SAS 9.1

1. Develop logistic regression model in a design-based software using population weights and bootstrapping.
2. Determine the potentially influential observations in SAS 9.1. (Appendix 2).
    i. Run logistic regression with standardized weights (Gagné, Roberts, and Keown, 2014).
    ii. Output into a temporary file Confidence Interval Displacement Statistics (C statistic) from the logistic regression.
    iii. Run proc univariate for the confidence interval displacement statistic. Examine the extreme values. If all five are greater than one (suggesting there may by more than five), then proc print can be used to print all cases where the C statistic is greater than one.
    iv. Output into another temporary file DFbeta statistics from the logistic regression.
    v. Run proc univariate for the DFbeta statistics. Examine the extreme values. If all five are greater than two (suggesting there may by more than five), then proc print can be used to print all cases where the DFbeta is greater than 2.
3. Remove the influential observations and determine the effect on the model.
    i. Create a duplicate dataset and delete the potentially influential observations identified by either the C statistics or the DFbetas.
    ii. Run the logistic regression again using standardized weights[1] on this duplicate dataset.
    iii. Create an Excel spreadsheet with a column for the parameter estimates (all cases used) and a second column for the parameter estimates (potentially influential observations deleted). In third and fourth columns, calculate the absolute difference and percentage difference for each parameter estimate between the two logistic regressions.
    iv. Determine how large a change in parameter estimates constitutes influential for the particular study; for example, a change in a parameter estimate of 10%.
    v. Compare the differences to see if removal of potentially influential observations affected parameter estimates. Flag percentage differences greater than 10% as parameter estimates that were significantly changed by influential observations.
    vi. Examine reasons for differences such as large weights, possibly miscoded data, or unique covariance patterns.

Appendix 2 – Assessing potentially influential observations in weighted logistic regression in SAS 9.1

***Code written in bold refers to headings;**
**Code written in italics refers to file and variable names that will vary depending on dataset and variables being used;*

Syntax for the Confidence Interval Displacement values

```
*CREATE CONFIDENCE INTERVAL DISPLACEMENT (C) VALUES;
*Standardize population weight;
proc means data=Mylib.datafilename noprint;
        var wtse_m;
        output  out=wts
        mean=mnwt;
data Mylib.datafilename;
        if _n_=1 then set wts;
        set Mylib.adolage3nooutliers;
        standardized_weight=wtse_m/mnwt;
        drop _freq_ _type_;

*Logistic regression;
proc logistic data=Mylib.datafilename;
title1 'Logistic regression';
model Outcome_variable=independent_var1 independent_var2 independent_var3;
      output out=Temporary_output_file_name c=c;
      weight standardized_weight;
run;

*EXAMINE CONFIDENCE INTERVAL DISPLACEMENT (C) VALUES;
*Univariate statistics;
*Provides 'C' five lowest and highest values;
proc univariate data= Temporary_output_file_name;
        var c;
title 'Proc univariate for logistic regression C values';
run;

*If all five values are greater than 1, run proc print to identify all extreme values;
proc print data=Temporary_output_file_name;
title 'Proc print for logistic regression C values greater than 1';
        var c;
        where c>=1;
run;
```

Syntax for the DFBeta Values

```
*CREATE DFBETAS;
*Logistic regression;
proc logistic data=Mylib.datafilename noprint;
model Outcome_variable=independent_var1 independent_var2 independent_var3;
        output out=Temporary_output_file_name dfbetas=dfbeta_independent_var1 defbeta_
      independent_var2 defbeta_independent_var3;
        weight standardized_weight;
run;

*EXAMINE DFBETAS;
*Below is code for one variable, this would be done for each variable as needed;

*Provides 'DFBETA' five lowest and highest values;
proc univariate data= Temporary_output_file_name;
var dfbeta_independent_var1;
title 'Proc univariate for logistic regression DFBETA values';
run;

*If all five values are greater than 2, run proc print to identify all extreme values;
proc print data=Temporary_output_file_name;
title 'Proc print for Name of logistic regression DFBETA values';
        var dfbeta_independent_var1;
        where dfbeta_independent_var1>=2;
run;
```

Appendix 3 – Assessing potentially influential observations in weighted logistic regression in SAS – SAS output

The output from the logistic regression will appear first in its usual format followed by "The Univariate Procedure" output and/or the "Proc Print" output as shown below.

A. Output for the Extreme Confidence Interval Displacement values

Proc univariate for *logistic regression C* values page #

Time, Day, Month Date, Year

The UNIVARIATE Procedure
Variable: c (Confidence Interval displacement C)

Extreme Observations

--------Lowest-------		-----Highest-----	
Value	Obs	Value	Obs
4.57384E -08	4000	1.37861	4800
7.39697E -08	1000	1.54967	3500
8.43858E -08	1500	1.83458	1200
1.23856E -07	1750	2.18468	2575
1.34758E -07	2500	5.39574	3600

B. Output for the Proc Print Confidence Interval Displacement values greater than 1

Proc print for *logistic regression C* values greater than 1 page #

Time, Day, Month Date, Year

c	Obs
1.37861	4800
1.54967	3500
1.83458	1200
2.18468	2575
5.39574	3600
1.33748	1450
1.21385	4150
1.11285	4375
1.24473	2750
1.22847	1850
1.18695	3700

Instructions for authors

The Information and Technical Bulletin will accept submissions for articles that address methodological or technical topics related to the datasets that are available at the Research Data Centres.

Language of material

Manuscripts may be submitted in English or French. Accepted submissions will be translated into both official languages for publication.

Length of submissions

The maximum length of submitted articles should not exceed 20 pages, double-spaced, excluding programs and appendices. In addition to in-depth explanations of technical issues, the Bulletin also accepts short (3 page) submissions that provide quick solutions to analytical problems and commentary from fellow researchers about material previously released in the Bulletin.

File formats and layout of text

Manuscripts must be submitted in Microsoft Word (.doc) and may be sent by regular mail on a disk or CD or by email.

Manuscripts must have a cover page showing the names of the authors, their primary institution of affiliation, and the contact information (telephone number, mailing address and e-mail address) of the lead author. Manuscripts must be prepared in 12pt Times New Roman, double-spaced, with 1-inch (2.5 cm) margins. Titles should have sentence-case capitalization (e.g., Bootstrapping made easy…).

Boldface type should only be used for headings. Underlining and italics are not to be used for headings.

Footnotes and references should be single-spaced and formatted according to *The Canadian Style: A Guide to Writing and Editing.*

File formats and layout of tables and charts

Tables and charts must be submitted in Microsoft Excel worksheets (.xls) or in comma-separated value (.csv) format. Each file must be clearly named table1, chart6, etc.

Tables and charts may be sent by regular mail on a disk or CD, or by e-mail.

Do not insert tables or charts into the text, but indicate their location in the text by inserting the title, followed by the filename in parentheses, e.g.

Chart 6 Chocolate consumption by children, Canada, 2000 (chart6)

Mathematical expressions

All mathematical expressions should be set out separate from paragraph text. Equations must be numbered, with the number appearing to the right of the equation flush with the margin.

Style guide

Please follow *The Canadian Style: A Guide to Writing and Editing*. It is available for purchase by contacting Government of Canada Publications, Public Works and Government Services Canada.

Address for submission

Manuscripts and all correspondence relating to the contents of the Bulletin should be sent to the Editorial Committee

- by email to MAD-HOOU@statcan.gc.ca

The review process

The editorial committee conducts the initial article review process. Editors may solicit past authors of the Bulletin or subject matter experts to participate in the process. The articles submitted to the Bulletin are reviewed for accuracy, consistency, and quality.

Upon completion of the initial review, the articles undergo both peer and institutional review. Peer reviews are conducted in accordance with Statistics Canada's Policy on the Review of Information Products. Institutional reviews are be conducted by members of senior management within Statistics Canada in order to ensure that the material does not compromise the Agency's guidelines of standards, or reputation for non-partisanship, objectivity and neutrality.

For more information about the review process, please contact the Editorial Committee at the address above.