Doctoral Thesis

Nonparametric M-estimation of a population mean

Author(s):
Hulliger, Beat

Publication Date:
1991

Permanent Link:
https://doi.org/10.3929/ethz-a-000596175 →

Diss. ETH No. 9443

Nonparametric M-Estimation of a Population Mean

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
Doctor of Mathematics

presented by
BEAT HULLIGER
Dipl. Math. ETH
born July 6, 1957
citizen of Aarau AG and Heimiswil BE

accepted on the recommendation of
Prof. Dr. F. R. Hampel, examiner
Prof. Dr. H. R. Künsch, co-examiner

1991 F. Hampel

Abstract

The mean of a variable over a finite population is an important indicator mainly due to its connection to the sum. It is a sensitive characteristic because it can be forced to take any desired value by moving only one observation. If there are outliers in a population there are two legitimate estimands: The population mean including the outliers or a mean of the "normal" observations. These two possibilities are reflected in two criteria for estimators, the representative and nonrepresentative mean squared error (MSE).

The sample mean is the natural estimator of a finite population mean or of the expectation of independent identically distributed random variables. It is always unbiased but it is nonrobust. If outliers may be sampled it is highly variable. Estimators with lower representative and nonrepresentative MSE are developed in the context of infinite populations. Location M-estimators are simple robust estimators. Usually they are biased estimators of the expectation but for a suitable choice of their defining function they may have lower MSE of both types than the sample mean due to a lower variance. The problem is to find a good defining function when the underlying distribution is unknown. A class of M-estimators whose defining functions are parametrized by one tuning constant is chosen. The representative MSE of an M-estimator is estimated by its asymptotic variance at the empirical distribution function plus the squared difference to the sample mean. The M-estimator which minimizes the estimated risk over the chosen class is the final estimator. It is called MER-estimator. Under certain conditions MER-estimators are scale equivariant and consistent for the expectation. But they are still not robust. Variants of the MER-estimator which are simpler or more robust are derived.

Biased M-estimators (BM-estimators) are M-estimators which are shrunken towards the sample mean by a penalty in their estimating equation. The weight of the penalty and the form of the two defining functions determine the robustness and asymptotic properties of BM-estimators.

In a simulation study a type of M-estimator which is similar to a trimmed mean, an ordinary M-estimator, two BM-estimators and five MER-estimators are investigated. For small to moderate sample sizes and skew populations MER- and BM-estimators may have considerably lower representative and much lower nonrepresentative MSE than the sample mean. On the other hand even in unfavourable situations they rarely have considerably higher representative MSE than the sample mean and almost never higher nonrepresentative MSE.

In complex samples from a finite population the Horvitz-Thompson estimator (HT-estimator) takes over the role of the sample mean. The HT-estimator is expressed as a weighted least-squares functional of a sample distribution function. The linear model needed is a regression through the origin (for unequal probability samples) or an analysis of variance (for stratified samples). The robustification of the HT-estimator becomes straightforward. An approximation of the sampling variance of the robustified HT-estimator is derived with a sampling sensitivity curve. Three estimators of this sampling variance are developed. They make MER-estimation applicable to sampling. In two examples the MER-estimator has lower representative MSE than the HT-estimator if outliers are present in the population, while it is practically as good as the HT-estimator if there are no outliers.

Zusammenfassung

Das arithmetische Mittel einer Variablen über eine endliche Population ist vor allem dank seiner Verbindung zur Summe ein wichtiger Indikator. Es ist aber auch unstabil, weil es genügt, einen einzelnen Wert zu verändern, um das Mittel einen beliebigen Wert annehmen zu lassen. Falls eine Population Ausreisser enthält, dann bestehen zwei legitime Zielgrössen: Das Mittel der gesamten Population einschliesslich der Ausreisser oder ein Mittel der "normalen" Beobachtungen in der Population. Diese zwei Möglichkeiten widerspiegeln sich in zwei Bewertungskriterien für Schätzer, dem repräsentativen und dem nichtrepräsentativen mittleren quadratischen Fehler (MQF).

Das Stichprobenmittel ist der natürliche Schätzer eines Populationsmittels oder des Erwartungswerts von unabhängigen, identisch verteilten Zufallsvariablen. Es ist immer unverfälscht, aber es ist nicht robust. Falls Ausreisser in der Stichprobe auftreten können, hat das Stichprobenmittel eine grosse Varianz. Schätzer mit tieferem repräsentativen und nichtrepräsentativen MQF als das Stichprobenmittel werden für unendliche Populationen entwickelt. Lokations M-Schätzer sind einfache robuste Schätzer. Sie sind gewöhnlich nicht erwartungstreu für den Erwartungswert, aber sie können bei geeigneter Wahl ihrer definierenden Funktion dank einer kleinen Varianz kleineren MQF beider Art haben als das Stichprobenmittel. Das Problem ist, wie man eine geeignete definierende Funktion finden kann, ohne die unterliegende Verteilung zu kennen. Dazu wird eine Klasse von M-Schätzern, deren definierende Funktionen durch eine Abstimmungskonstante parametrisiert werden, gewählt. Der repräsentative MQF eines M-Schätzers wird geschätzt durch die asymptotische Varianz an der empirischen Verteilungsfunktion plus das Quadrat der Differenz zum Stichprobenmittel. Der M-

Schätzer, welcher das geschätzte Risiko über die ausgewählte Klasse minimiert, ist der endgültige Schätzer. Er wird MER-Schätzer genannt. MER-Schätzer sind unter gewissen Voraussetzungen Skalen-äquivariant und konsistent für den Erwartungswert. Aber sie sind noch nicht robust. Varianten von MER-Schätzern, welche einfacher oder robuster sind, werden entwickelt.

Verfälschte M-Schätzer (BM-Schätzer) sind M-Schätzer, die durch eine Bestrafung in der Schätzgleichung zum Stichprobenmittel hin geschrumpft werden. Das Gewicht der Bestrafung und die zwei definierenden Funktionen bestimmen die Robustheits- und asymptotischen Eigenschaften von BM-Schätzern.

In einer Simulations-Studie wird eine Art M-Schätzer, welcher einem gestutzten Mittel ähnlich ist, ein gewöhnlicher M-Schätzer, zwei BM-Schätzer und fünf MER-Schätzer untersucht. Für kleine bis mittlere Stichprobengrössen und bei schiefer Verteilung haben MER- und BM-Schätzer oft beträchtlich kleineren repräsentativen und viel kleineren nichtrepräsentativen MQF als das Stichprobenmittel. Andererseits haben sie auch in ungünstigen Situationen selten beträchtlich höheren repräsentativen MQF und fast nie höheren nichtrepräsentativen MQF.

Bei komplexen Stichproben aus endlichen Populationen übernimmt der Horvitz-Thompson Schätzer (HT-Schätzer) die Rolle des Stichprobenmittels. Der HT-Schätzer wird als Kleinste Quadrate Funktional an einer Stichproben-Verteilungsfunktion ausgedrückt. Das benötigte lineare Modell ist eine Regression durch den Nullpunkt (für Stichproben mit ungleichen Wahrscheinlichkeiten) oder eine Varianzanalyse (für geschichtete Stichproben). Die Robustifizierung des HT-Schätzers wird dann einfach. Eine Approximation der Stichprobenvarianz des robustifizierten HT-Schätzers wird mit Hilfe einer Stichproben-Sensitivitätskurve hergeleitet. Drei Schätzer dieser Stichprobenvarianz werden entwickelt. Sie machen MER-Schätzung auf komplexe Stichproben anwendbar. In zwei Beispielen erreicht der MER-Schätzer tieferen repräsentativen MQF als der HT-Schätzer, falls in der Population Ausreisser vorhanden sind, während er praktisch gleich gut wie der HT-Schätzer ist, falls keine Ausreisser vorhanden sind.