NeuroImage 147 (2017) 658–668

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

A comprehensive review of group level model performance in the presence of heteroscedasticity: Can a single model control Type I errors in the presence of outliers?

Jeanette A. Mumford

Center for Healthy Minds, University of Wisconsin, Madison, United States

ARTICLE INFO

Keywords:
Robust regression
Ordinary least squares
Outliers
Heteroscedasticity

ABSTRACT

Even after thorough preprocessing and a careful time series analysis of functional magnetic resonance imaging (fMRI) data, artifact and other issues can lead to violations of the assumption that the variance is constant across subjects in the group level model. This is especially concerning when modeling a continuous covariate at the group level, as the slope is easily biased by outliers. Various models have been proposed to deal with outliers including models that use the first level variance or that use the group level residual magnitude to differentially weight subjects. The most typically used robust regression, implementing a robust estimator of the regression slope, has been previously studied in the context of fMRI studies and was found to perform well in some scenarios, but a loss of Type I error control can occur for some outlier settings. A second type of robust regression using a heteroscedastic autocorrelation consistent (HAC) estimator, which produces robust slope and variance estimates has been shown to perform well, with better Type I error control, but with large sample sizes (500–1000 subjects). The Type I error control with smaller sample sizes has not been studied in this model and has not been compared to other modeling approaches that handle outliers such as FSL's Flame 1 and FSL's outlier de-weighting. Focusing on group level inference with a continuous covariate over a range of sample sizes and degree of heteroscedasticity, which can be driven either by the within- or between-subject variability, both styles of robust regression are compared to ordinary least squares (OLS), FSL's Flame 1, Flame 1 with outlier de-weighting algorithm and Kendall's Tau. Additionally, subject omission using the Cook's Distance measure with OLS and nonparametric inference with the OLS statistic are studied. Pros and cons of these models as well as general strategies for detecting outliers in data and taking precaution to avoid inflated Type I error rates are discussed.

1. Introduction

When analyzing fMRI data, even with thorough preprocessing, it is likely that artifacts will prevail in some subject's data causing outlying blood oxygen level dependent (BOLD) contrast estimates in the group level analyses. This can be a concern when the group level model involves a continuous covariate, since outliers can easily influence the fit of a regression line. It can also be an issue with categorical covariates, although mean estimates are often less impacted by outliers than regression slopes. A drawback of the most common analysis strategy for imaging data is that it involves blindly applying a model in a voxelwise fashion, inspecting only the *p*-value maps. Comparatively, in a standard single regression analysis, say using behavioral data only, multiple plotting strategies and statistical assessments are used to study heteroscedasticity and other violations of regression assumptions. This practice is somewhat difficult in voxelwise analyses and so a common goal is to find a model, such as robust regression, that aims to detect and downweight the contribution of outliers in a regression analysis. Although there have been studies that focus on models that are robust to outliers (Wager et al., 2005; Fritsch et al., 2015; Woolrich, 2008), in each case only subsets of robust models have been compared over a somewhat limited set of heteroscedasticity scenarios and not all focused on performance with continuous regressors, but focus on group 1-sample *t*-tests. The purpose of this work is to examine the Type I error rate across a wide selection of regression models, including some that have not been considered in the context of fMRI analysis. Also, a larger set of heteroscedasticity settings, varying both the type and degree of heteroscedasticity are considered.

The most commonly used robust regression approaches rely on estimators of the regression slope that are robust to outliers. Another class of robust regression approaches, utilizing heteroscedastic autocorrelation consistent (HAC) estimators, also provide robust variance

E-mail address: jeanette.mumford@gmail.com.

http://dx.doi.org/10.1016/j.neuroimage.2016.12.058
Accepted 20 December 2016
Available online 25 December 2016

estimates. These will be referred to as “doubly robust” since both the slope and variance estimators are robust to outliers. In this work, the models compared are two types robust regression (singly and doubly robust), FSL's Flame 1 (similar to AFNI's MEMA), FSL's Flame 1 with outlier de-weighting, Ordinary Least Squares (OLS), which is equivalent to most commonly used model in SPM and AFNI, and Kendall's rank correlation. Improvements to OLS also considered are removing subjects according to the Cook's D metric and using nonparametric inference, which has fewer assumptions than parametric inference. All other approaches rely on parametric inference.

Due to the repeated measures nature of fMRI data, the variance structure has both a within-subject and a between-subject variance component and the outliers can be driven by heteroscedasticity in either of these variances. Past works only consider model comparisons with heteroscedasticity within one of these variance types, whereas here the comparison is across all models with heteroscedasticity in either variance component. Lastly, a wider selection of heteroscedastic variance patterns are considered, including univariate outliers, multivariate outliers and heteroscedasticity that correlates with the group model covariate (e.g. variance in BOLD contrast increases with an impulsivity measure of interest). Also, instead of only considering one level of outlying variance, a continuum of outlier degree is studied, illustrating how models perform with weak and strong outliers.

1.1. Heteroscedasticity

The residual plots (residual versus explanatory variable) in Fig. 1 illustrate the heteroscedasticity settings considered here. In the univariate outlier case (top row), the outlier is either in the explanatory variable or in the explained variable, while in the multivariate case (bottom left) both the explanatory and explained values are outlying. The final case, which has never been considered in robust regression studies of neuroimaging data, is when the variance increases along with the explanatory variable (bottom right). This will be referred to as heteroscedasticity without outliers, since there are no clear outlying values, but the variance is still heterogeneous.

1.2. Within- and between-subject variance

Here it is assumed that each subject has a single functional run of data and in this case the standard modeling approach is the two-stage summary statistics model (Mumford and Nichols, 2006). The first stage models the time series data and, for subject i, results in a within-subject estimate of the BOLD contrast, $\hat{\beta}_i$, as well as the within-subject variance of the contrast, which will be denoted $\sigma^2_{w,i}$. The second stage model combines the within-subject contrast estimates and their variances in a group model. This model results in a group contrast estimate, γ, as well as a between-subject variance, σ^2_b, which is combined with the within-subject variance to form the mixed effects variance, $\sigma^2_{w,i} + \sigma^2_b$. Specifically, for subject i, let $\hat{\beta}_i$ be the level 1 contrast estimate, W_i is the group level covariate value (assumed to be a scalar), and γ is the group-level parameter (regression slope) then

$$\hat{\beta}_i \sim N(W_i\gamma, \sigma^2_{w,i} + \sigma^2_b). \quad (1)$$

Given this structure, it is clear that outliers in the $\hat{\beta}_i$ can be driven either by inflated within- or between-subject variance. To be clear, the focus here is on outliers in the first level parameter estimates ($\hat{\beta}_i$) and not in the time series data, which are not directly studied in this work. Of course it could be the case that a subject with multiple outliers in their time series data, say due to motion, may have an inflated value for $\sigma^2_{w,i}$. The following section describes the various estimation strategies

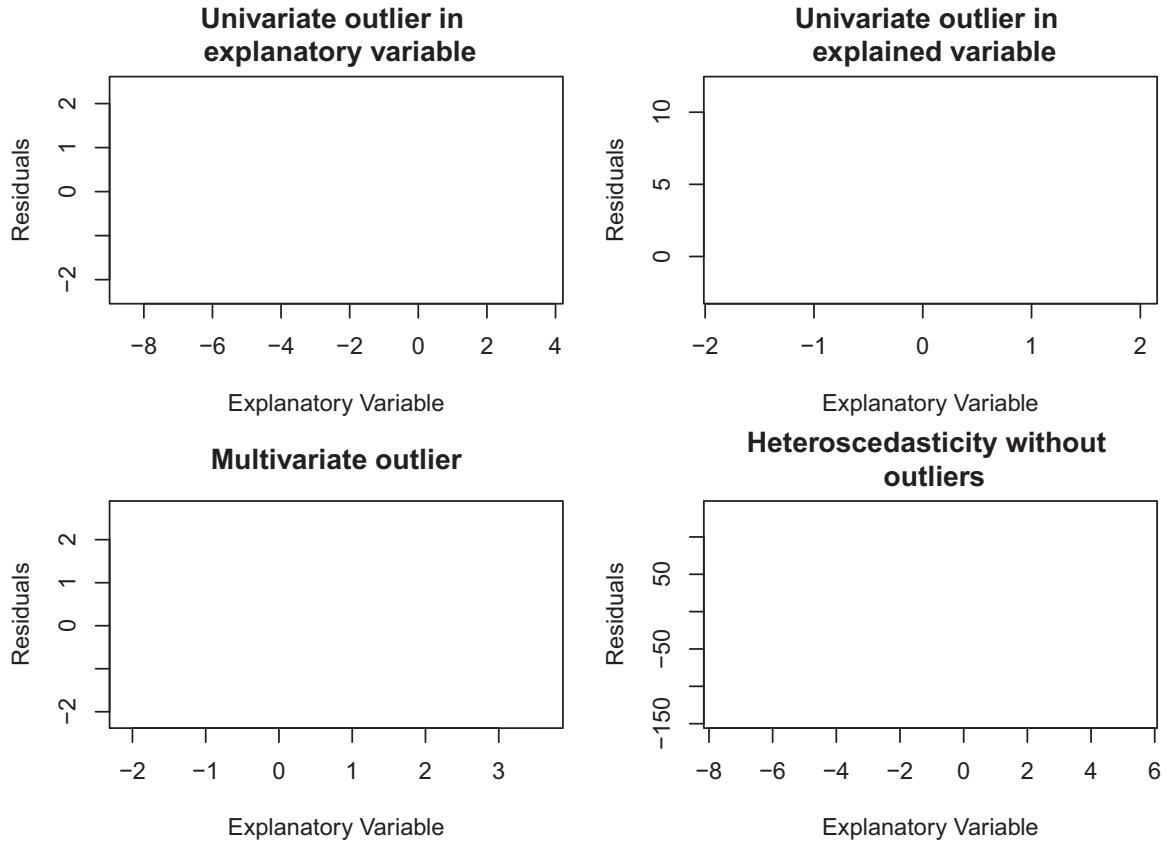

Fig. 1. Residual plots illustrating different types of heteroscedasticity. The top two plots represent univariate outliers where the outlier is either in the explanatory variable (left) or explained variable, shown by the large residual (right). The bottom left shows a multivariate outlier where both the explanatory variable and residual are inflated. The bottom right shows an example of heteroscedasticity without outliers, where the variance gradually decreases with the explanatory variable.