HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2018 February 15.

Published in final edited form as:

Neuroimage. 2017 February 15; 147: 658–668. doi:10.1016/j.neuroimage.2016.12.058.

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

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.

Keywords

robust regression; ordinary least squares; outliers; heteroscedasticity

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 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 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 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.

Heteroscedasticity

The residual plots (residual versus explanatory variable) in Figure 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.

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 for this group model and their corresponding assumptions. Specifically, some models will simplify the variance structure by assuming the within-subject variance is constant across subjects.

Group models considered and previous work

The specific details of the group models will be given in the methods, but will be broadly discussed here. The simplest group model is OLS, where the within-subject variance is assumed to be constant across subjects, simplifying the mixed effects variance to a single parameter, σ^2, so Equation 1 becomes $\hat{\beta}_i \sim N(W_i\gamma, \sigma^2)$. The estimate of this model is found by least squares, which is the value of γ that minimizes

$$\sum_{i=1}^{N}\left(\hat{\beta}_i - W_i\gamma\right)^2. \quad (2)$$

This estimator is commonly used in the SPM software package (www.fil.ion.ucl.ac.uk/spm/) and in AFNI (afni.nimh.nih.gov/afni/).

In comparison, FSL's (fsl.fmrib.ox.ac.uk/fsl) Flame 1 model (Woolrich et al., 2004) and AFNI's MEMA model (Chen et al., 2012) do not relax the assumption of Equation 1, and allow for heteroscedasticity in the within-subject variance. Univariate outliers driven by the within-subject variance, alone, and the impact on Type I error and power between this style of model and OLS was compared in Beckmann et al. (2003) and Mumford and Nichols (2009). The model of interest was the 1-sample t-test and it was found that Type I error was preserved for both approaches, but power was slightly reduced for OLS in the presence of univariate outliers. This work will instead focus on continuous regressors in the group level, which has not yet been done for this model comparison.

In robust regression, a score function is used to differentially weight subjects according to the size of their residual, which will account for some forms of heteroscedasticity. Specifically, the ratio of the subject's residual and the overall standard deviation, σ, is passed into a function ρ and the minimization problem is defined by finding γ that minimizes

$$\sum_{i=1}^{N}\rho\left(\frac{\hat{\beta}_i - W_i\gamma}{\sigma}\right). \quad (3)$$

Effectively this turns into a weighted linear regression, where outlying subjects, as determined by their residual magnitude, contribute less to the parameter. The weight is a function of the score function, ρ. Common settings for ρ include Tukey's Bisquare or Huber's loss function, which are plotted in Figure 2 and must be chosen when running a robust regression. A second choice involves the estimator of σ and options include *M*, *S* and *MM*. Among other properties, these estimators differ in their computational ease and what is called the breakdown point, which indicates what proportion of the data can contain outliers before the estimator may fail. The *M* estimator uses a median absolute deviation (MAD) estimator of σ. Although it is computationally simple, it has a breakdown point of 0 (Huber, 1981). On the other hand, *S* estimators use a residual scale estimator of σ and are also simple to estimate, but have a breakdown point of 50%. The downside is the estimates tend to have low efficiency (i.e. are more variable). The *MM* estimator combines the *S* and *M* estimators, where the *S* estimation strategy is used as a starting point for the *M* estimator, which allows for a higher breakdown point (50%) while retaining efficiency. The specifics about these estimators can be found in Croux et al. (2003). Since the parameter, σ, appears within the minimization step, the two parameters, γ and σ are estimated iteratively using what is known as iteratively reweighted least squares. In this work, this model will be referred to as "singly robust" as only the estimate of γ is robust to outliers and not σ.

In Wager et al. (2005) both the Bisquare and Huber loss function were considered in singly robust models in comparison with OLS and some other estimation approaches not revisited here (e.g. dropping subjects with high Mahalanobis distance). Outliers were either univariate or multivariate and the regression setting was with a single continuous regressor and the sample sizes considered were between 20–40 subjects. The findings indicated that Type 1 error is controlled for univariate outliers in the explained variable for both robust methods and OLS, but OLS suffered from a loss in power. In the multivariate case, no method properly controlled Type I error rates.

A slightly different approach to robust regression was presented in Fritsch et al. (2015), where a Randomized Parcellation Based Inference (Da Mota et al., 2014) is combined with robust regression. Huber's loss function was used and the sample size was much larger than in Wager et al. (2005), typically 400–1000 subjects. Competing models included OLS, support vector regression (SVR) and least trimmed squares (LTS) and univariate outliers in models with continuous regressors were considered. As in Wager et al. (2005), OLS and robust regression were found to control Type I errors, but with a loss in power for OLS when univariate outliers were present. Interestingly, although Type I errors were found to be controlled, the real data analysis revealed a significant cluster with OLS that was not present with robust regression, in which a single influential outlier caused the false positive activation. This does not contradict the finding that Type I errors are preserved with OLS, but instead reflects that when using a p-value threshold of 0.05 there is still a 5% chance for false positives to occur with either method and the two methods will not necessarily both experience false positives with the same data.

The robust regression estimator considered up until now only provides an estimate of γ that is robust to outliers. This is the most commonly found robust regression model across software packages. Within the R software package, in some cases the p-values are output as part of robust regression output, but in many cases, such as the `rlm`, p-values are not supplied. This is because there is a debate over the proper way to derive p-values for robust regression (Croux et al., 2003). The R software package's `lmrob` function does output p-values and, therefore, may be thought to be a better robust model. This function uses an estimation strategy that additionally provides robust standard error estimates that are heteroscedasticity and autocorrelation consistent (HAC) in addition to being robust to outliers. The derivation of this estimator can be found in Croux et al. (2003), where, for large sample sizes (1000 subjects) it was found to have the strongest control of Type I error over singly robust models in the case of heteroscedasticity without outliers. Most notably, the singly robust approach and OLS were not able to control the Type I error rate due to an underestimate of the standard error. The weakness of the HAC estimator is when the errors are homoscedastic. Although the Type I error is preserved, at the large sample sizes they considered, the estimate of the standard error suffers from a loss in precision (Croux et al., 2003). Although this work provides promise for doubly robust regression, the sample sizes and heteroscedasticity settings were not realistic representations of a typical fMRI study, where sample sizes are likely less than 100 and if heteroscedasticity without outliers is present, it will not be as dramatic as in the simulations used in Croux et al. (2003).

Lastly, the outlier de-weighting algorithm that is part of FSL (Woolrich, 2008) models the between-subject variance as a mixture of two Gaussian distributions while also allowing for heteroscedastic within-subject variances using Flame 1. Inference is based on a Bayesian framework, using an expectation-maximization estimation. This allows subjects to have different between-subject variances and be weighted differentially in the model estimation. In Woolrich (2008), comparison models included OLS, singly robust with Tukey's Bisquare loss function and a permutation-based test that incorporated the within-subject variances. The permutation method used the lower level variances by permuting data in the usual way, but using the Flame 1 T-statistic instead of an OLS statistic. This permutation strategy was not considered here because in the situations where Flame 1 had issues, permutation tests would not be likely to offer any improvement due to exchangeability assumption violations. The simulations focused on univariate outliers in the between-subject variance with a 1-sample t-test as well as regression with a single continuous covariate. The Flame 1 with outlier de-weighting approach was found to generally perform better than OLS and robust regression with better control of Type I errors and higher power. The permutation test results fell between that of OLS and Flame 1 with outlier de-weighting. Importantly, Woolrich (2008) found that when the regression covariate was skewed, the Flame 1 with outlier de-weighting approach did not perform well. This is to be expected, since it would deviate from the assumption that the distribution of the errors follows a mixture of two Gaussian distributions.

This work combines all of these models to study and compare their performances across a wider set of heteroscedasticity settings than have been previously used and focus on a wide array of sample sizes, from 30–500. Also, instead of only focusing on a single magnitude of outlier, a continuum of outlier values is studied to provide a thorough comparison of the models. The primary question is whether there will be a model that can perform well in all situations and, if not, what can be done to help control the influence of outliers on results, while preserving the Type I error rate.

Methods

Models considered

The robust regression routines used were from the R software package (www.r-project.org). For singly robust regression, the `rlm` function within the `MASS` library was used and the doubly robust regression, using the HAC estimator for the variance, was implemented using the `lmrob` function within the `robustbase` library. In both cases MM estimation using Tukey's bisquare score function was used. For OLS regression, the lm function was used and when using a Cook's D criterion to select and omit outliers, the `cooks.distance` function was used and thresholds of 1 and $4/N$, where N is the number of subjects, were considered. Kendall's rank correlation p-values were computed using the `cor.test` function. The Flame 1 and Flame 1 with outlier de-weighting algorithms were implemented using the `flameo` function, which is part of FSL. Permutation tests were coded in R. Although it is standard to use a cluster-based permutation, the focus here is on voxelwise statistics and so uncorrected, voxelwise permutation tests were conducted. In the

permutation tests, subject labels were permuted and 5000 permutations were used to derive the nonparametric null distribution. All p-values correspond to a 2-tailed hypothesis.

Real data

The Balloon Analog Risk task (Lejuez et al., 2002) of the "Generality of Self Control" data set in the open fmri data base (set 00008 in openfmri.org) was used to guide the parameter value choices in the simulations. In the task, participants saw a balloon and could either choose to inflate it further or stop inflating it with left and right button presses, respectively. The value of the balloon started at 50 cents and increased by 25 cents with each pump. When a participant decided to stop pumping the balloon, the current value was saved as winnings and if the balloon was pumped until explosion, which occurred after a variable number of pumps, no winnings were added. The first level contrast of interest was the BOLD activation difference for inflating a balloon versus choosing to stop inflating (accept-reject) and the group level covariate of interest was the within-subject number of balloons that exploded. There were a total of 24 subjects with a mean age of 20.8 (range 18–33, 10 females). Further details about the paradigm can be found in the study description posted on openfmri.org (openfmri.org/media/ds000009/ds009 methods 0 CchSZHn.pdf). The data were analyzed using the FSL software package, which supplies both within-and between-subject variance estimates that were used drive reasonable parameter settings in the simulations. In other words, variance magnitudes that would be found in real data.

FMRI data processing was carried out using FEAT (FMRI Expert Analysis Tool) Version 6.00, part of FSL. The following pre-statistics processing was applied; motion correction using MCFLIRT (Jenkinson et al., 2002); non-brain removal using BET (Smith, 2002); spatial smoothing using a Gaussian kernel of FWHM 5mm; grand-mean intensity normalization of the entire 4D dataset by a single multiplicative factor; highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with sigma=50.0s). Registration to high resolution structural and/or standard space images was carried out using FLIRT (Jenkinson and Smith, 2001; Jenkinson et al., 2002). Time-series statistical analysis was carried out using FILM with local autocorrelation correction (Woolrich et al., 2001). In addition to the task related regressors (accept, explode and reject trials) nuisance regressors were added to the model to address motion artifact. The 6 standard motion parameters plus the extended motion parameters (derivative, square and derivative of square) were added as well as indicator regressors for high motion time points, indicated by a framewise displacement (FD) larger than 0.9. Higher-level analysis was carried out using FLAME (FMRIB's Local Analysis of Mixed Effects) stage 1 (Beckmann et al., 2003; Woolrich et al., 2004; Woolrich, 2008). The group level model only included a regressor for the number of balloons exploded, for each subject and an intercept.

General simulation setup

Data were simulated for a single voxel and Type I error and power were calculated based on 10,000 single voxel analyses. Time series data were simulated based on the two stage model:

$$\begin{aligned} Y_i &= X\beta_i + \varepsilon_i \\ \beta_i &= W_i\gamma + u_i. \end{aligned} \quad (4)$$

In the first level model, Y_i is the BOLD time series of length T for subject i, X is the first level design matrix, β_i is the within-subject BOLD activation magnitude for subject i, and the first level error term follows a Gaussian distribution, $\varepsilon_i \sim N(0, \sigma_i^2 I_T)$, where I_T is a $T \times T$ identity matrix. It is assumed that the design matrix, X, was the same across all subjects and contained a single regressor. The variance estimates are based on the real data, so the effective regressor for the accept-reject contrast for the first subject in the real data analysis was used as the regressor in the simulations (Smith et al., 2007). The within-subject variance of $\hat{\beta}_i$ is then given by

$$\sigma_{w,i}^2 = (X'X)^{-1}\sigma_i^2. \quad (5)$$

In the second level, W_i is the group level covariate value for subject i and u_i is the second level error term where $u_i \sim N(0, \sigma_b^2)$. In simulating the data it was assumed that the group level model did not include an intercept, but an intercept was always included when modeling the data.

The different types of heteroscedasticity were simulated either through $\sigma_{w,i}^2$ or σ_b^2. To generate the simulated time series the subject-specific estimate, β_i, is sampled from a Gaussian,

$$\beta_i \sim N(W_i\gamma, \sigma_b^2), \quad (6)$$

and the time series for a single voxel for subject i is then given by drawing a sample from

$$Y_i \sim N(X\beta_i, \sigma_i^2 I_T), \quad (7)$$

Time points were assumed to be independent, as this should not impact the performance of the group-level models.

After simulating the BOLD time series data, OLS regression was used to estimate $\hat{\beta}_i$ and $\hat{\sigma}_{w,i}^2$, which were then entered into the competing second level models.

Univariate outliers in explanatory variable

The real data had an average estimated within-subject variance around 5000 and between subject variance around 2500, so these values were used for $\sigma_{w,i}^2$ and σ_b^2, respectively. In this

simulation the outlier was in the explanatory variable, W_i, which was simulated by drawing the nonoutlying values of the explanatory variable from a standard normal distribution and then 10% of the data were assigned as outliers and their value for W_i was sampled from a normal distribution with variance ranging between 1–9.

Univariate outliers in explained variable

Again the non-outlying within- and between-subject variances are 5000 and 2500, but 10% of the subjects were denoted outliers by inflating either $\sigma^2_{w,i}$ or σ^2_b, but not both. Outliers in $\sigma^2_{w,i}$ ranged between 10,000 to 70,000, while outliers in σ^2_b ranged between 5000 and 70,000. This means, that in each scenario the outlying mixed effects *standard deviation* was 1–3 times larger than the non-outlying group. For the power calculations, γ was set to 50, for sample sizes of 30, and 3 when studying power over a range of sample sizes and W_i was sampled from a standard normal distribution.

Bivariate outlier

Just as before 10% of the data were assigned as outliers and the outlying variances followed the same setup as the univariate outliers in the explained variable and we focused on outliers in W_i sampled from a normal distribution with a variance of 9, while the nonoutlying covariate values were sampled from the standard normal.

Heteroscedasticity without outliers

Through inspection of multiple fMRI data sets, we did not find clear cases where there was a strict linear increase in the variance as a function of the group level covariate, as illustrated in Figure 1, but instead found that when heteroscedasticity was paired with a skewed group-level covariate, results between the methods differed. The simulation setting was heteroscedasticity driven by two levels of variability corresponding to a median split of the group-level covariate. Using the real data we split the subjects into low and high explosion groups, depending upon whether their number of balloon explosions was below or above the median (12 explosions). After estimating the between-subject variance separately for each group using Flame 1, we split the mixed effects variance ratio (high/low explosion) into 20 percentiles and the separate within- and between-subject variance estimates for each percentile are shown in Table A.1. Although both the within- and between-subject variances vary across the percentiles, the between-subject variance differences drive the heteroscedasticity and so the simulations used the average within-subject variance (4788) while varying the between-subject variance according to the numbers in the Table.

To simulate the data and assign the variance values, we first randomly created the group-level regressor, number of balloon explosions, by randomly drawing a value from the Gaussian kernel density estimate, with a bandwidth of 1.56, of the real data distribution. This was done by first randomly sampling a value, with replacement, from the number of explosions data, call this μ_{sub}, and then simulating a subject's number of explosions by drawing a single sample from $N(\mu_{sub}, (1.56)^2)$. Figure 3 shows the distribution of this value over subjects as well as the kernel density estimate. Once the group level covariate, W_i, was determined, the simulated subject was assigned to the low or high explosion group according

to the median value of the original number of explosions regressor (12 explosions) and variances were assigned according to the values in Table A.1 for the between subject variance while 4788 was used for the within-subject variance. For Type I error estimation, γ was set to 0 and for power it was set to 8.

Results

The following will present the Type I error and power results for all simulations. The OLS model using Cook's $D < 4/N$ never appears in plots, as this approach did not control the Type I error well in any of the simulations. Also, note that throughout the results the Permutation-based results were so similar to OLS that a separate line was not plotted. This may seem unexpected, but is a result of heteroscedasticity violating the exchangeability assumption of the permutation test. More about the history of the permutation test and heteroscedasticity is included in the Discussion section. In some cases Cook's $D < 1$ also matched OLS, so is either included with OLS result or will have a separate line when it differed from OLS as specified in the figure legends.

Lastly, when interpreting the results for Power, keep in mind that the level of Power can only be considered if a test is valid, or when the Type I error is controlled.

Univariate outliers in explanatory variable

Figure 4 shows the Type I error rate (top) and Power (bottom) for the case where there was a univariate outlier in the explanatory variable for 10% of the data, with a total of 30 subjects. The x-axis indicates the degree of the outlier in the explanatory variable. Note that the ratio starts at 1, so this represents the case of no outliers. The doubly robust method does not control the Type I error rate, while all other models do well. Importantly, when the sample size is increased, the doubly robust regression does produce valid test results for large sample sizes (Supplemental Figure S1). Flame 1, Flame 1 with outlier de-weighting and OLS have the highest power, followed by singly robust and Kendall. Note the power for the doubly robust method cannot be considered, since the Type I error is not preserved. It may seem counterintuitive that power is reduced with the singly robust model, since the univariate outlier case was found to have higher power with singly robust than OLS in Wager et al. (2005) and Fritsch et al. (2015), but those univariate outliers were in the explained variable, which is covered in the next section. Since the outlier is in the explanatory variable, the β_i may be outlying, but will follow the true regression slope without having an inflated residual error. Therefore, the complexity of the robust regression model is not needed and the standard error estimates are slightly inflated, causing the decrease in power. Note that this is the case when there are no outliers (ratio = 1) as well.

Univariate outliers in explained variable

Figure 5 shows the Type 1 error rates (top panels) and power estimates (bottom panels) when the outlier is driven by the between-subject variance (left) and within-subject variance (right) for a sample size of 30 when 10% of the data are outliers. The doubly robust model again fails to control the Type I error rate and the singly robust has a slightly inflated Type I error rate in both settings. Although the doubly robust method does not control the Type I

error rate at a sample size of 30, as the sample size increases, the Type I error rate and power improve for this approach although it requires a sample size of at least 500 subjects. Supplementary Figure S2 shows this result when the outlier is driven by the between-subject variance. At a sample size of 30, for outliers driven by the between-subject variance, the power gain depends on the severity of the outlier, where for less severe outliers, Flame 1 and OLS outperform the singly robust approach. For more severe outliers, the singly robust model has the highest power, followed by Flame 1 with outlier de-weighting, although Flame 1 with outlier de-weighting has inflated Type I errors for extreme outliers.

For the setting where the outlier is driven by the within-subject variance, the results are similar except for the Flame-based models, which is to be expected. Both Flame 1 and Flame 1 with outlier de-weighting have superior power, compared to the singly robust approach since they specifically use the within-subject variances to downweight noisy subjects. Flame 1 is known to *overestimate* the between subject variance when the between-subject variance is small, leading to very conservative tests (Eklund et al., 2015) and so a simulation where the between subject variance was set to 100 for all subjects and the within-subject variance controlled outliers as before was performed. Supplementary Figure S3 shows that, as expected, the Type I error rate is conservative for the Flame 1 algorithms in this setting, but the power is still highest across all other models.

Bivariate Outlier

Figure 6 shows the results for the bivariate outlier when the outlying standard deviation in the explanatory variable was 3 times that of the non-outlying group. When the outlier is in the between-subject variance, all models, but Kendall's Tau, fail to control the Type I error rate. This resembles results previously found in Wager et al. (2005). As is to be expected, when the heteroscedasticity is in the within-subject variance, the Flame 1-based approaches do control Type I errors, better than Kendall's Tau and with much higher power. As in the previous cases, increasing the sample size does improve the performance of the doubly robust model, but the added sample size does not improve the Type I error rate for the singly robust modeling approach (Figure 7). Thus, with a large enough sample size, the doubly robust regression is the only regression-based approach with controlled Type I errors for this type of heteroscedasticity. Although Kendall's Tau has controlled Type I errors as well, the model cannot include multiple covariates and the interpretation is limited.

Heteroscedasticity without Outliers

Figure 8 shows the Type 1 error and power across all levels of heteroscedasticity shown in Table A.1 for a sample size of 30. In both plots the x-axis represents the ratio of the true mixed effects variance for the below median explosion group compared to the above median explosion group. When the ratio is less than 1, the variance is larger for the low explosion group and there is a stronger impact on the inflation of Type 1 errors across the models compared to when the ratio is larger than 1, where the variance is larger for the high explosion group. This is because the distribution is left skewed (Figure 3) and so inflating the variance in the low explosion group, where the explosions fall into the lower tail of the distribution is somewhat like a more gentle version of the multivariate outlier. The Kendall's Tau approach is the only one with controlled Type I errors across all settings, but has

reduced power compared to the OLS- and Flame-based methods when their error rates are controlled. Also, notable, is that OLS has better Type I error control than singly robust regression in this setting.

Supplemental Figure S5 shows the impact of increasing the sample size when the heteroscedasticity was the worst (ratio = 0.364) and shows that the doubly robust model, which is specifically designed to handle this type of heteroscedasticity, does not have controlled Type I errors until the sample size is around 500 subjects, whereas the singly robust model does not have improvement in the error rate as the sample size increases. In fact, the singly robust regression model has a higher error rate than all other models. Kendall's Tau has a controlled error rate, while the OLS-based, Flame-based and permutation approaches have similar performances, with a slight decrease in Type I error as the sample size increases.

Discussion

This work has presented an extended review of models typically used in fMRI analysis, or considered when outliers may be present. Type I error rates and power were compared across a wider set of heteroscedasticity settings, degree of heteroscedasticity and sample sizes than has been done in previous work. The findings support that there is not a single model that can control Type I error properly for all types of heteroscedasticity, although some approaches should be avoided regardless of degree of heteroscedasticity and sample size. Specifically, blindly applying Cook's D to omit outliers either shows no improvement over OLS or has inflated Type I error rates and, thus, should never be implemented. Although permutation tests are more effective at controlling the Type I error rate for multiple comparison correction (Eklund et al., 2015), in terms of handling outliers in the settings considered here, it performs similarly to OLS with parametric thresholding. This is discussed in more detail below. The singly robust regression model loses Type I error control in the bivariate outlier case as well as the heteroscedasticity with outlier case, with a slight increase when the outlier is univariate (in the explained variable). In some cases a boost in power was given by the singly robust model, but when no heteroscedasticity is present, the inflated variance estimates slightly decrease the power performance of this model. The doubly robust model has poor performance for small sample sizes, but has type I error control for very large sample sizes. OLS was found to have slightly better control of Type I error in the heteroscedasticity without outliers case. Kendall's Tau is discussed more below, but generally has good control of Type I error rates and poor power, although the Type I errors can increase with strong, bivariate outliers. The Flame 1 approach offers an obvious benefit when the outliers are present in the within-subject variance, with higher power than robust regression in the univariate outlier cases, if the outlier is driven by the within-subject variability. Since Flame 1 performs so well when the outlier is driven by the within-subject variance, one approach might be to use Flame 1 and carefully inspect the model residuals to see if there are any outliers driven by the between-subject variance. If so, Flame 1 with outlier de-weighting could be implemented, although it does not always have strong control over Type I error rate and is not as highly powered as robust regression, in some settings where the outlier is driven by the between-subject variance. Notably the outlier de-weighting algorithm is computationally intensive to implement. It took two hours for the model to

complete using the 24 subject real data (BART study described above) across 226,000 voxels on a Linux 7 computer with 8 × 3.5GHz Intel Xeon processor cores and 16GB of RAM.

Why does the permutation test fail?

The primary assumption of the permutation test is that exchangeability holds under the null, meaning the underlying distribution of the data are not perturbed by the permutations. The most common example where exchangeability does not hold is when data are correlated. Permuting in this case involves swapping time points and violates exchangeability due to disruptions in the correlation structure. Outliers in the case of the 1-sample t-test do not violate this assumption since the permutation involves multiplying randomly chosen subject's estimates by -1. This preserves the outliers' behavior in the data, in that they remain outliers. In this work the most egregious problems occurred with the bivariate outlier, where no method controls the false positive rate well. In this case the "feature" of the distribution that must be preserved is the behavior of the influential outlier. For most shuffes of subject labels, the permutation strategy in this case, the outlier will fall into the univariate outlier case, which has less influence. Results similar to what was found here were found in Hayes (1996), Rasmussen (1989) and Hahn et al. (2013).

What can be done about outliers?

It is a bit discouraging that there isn't a single model that can handle all types of heteroscedasticity, reliably. This seems like a larger issue for imaging analyses where hundreds of thousands of regressions are simultaneously estimated and viewing the data is a daunting task. What this work should encourage is a more careful inspection of data. For example, the worst offenders in elevated Type I error occurred when there were issues with the explanatory variable, which is easy for the researcher to investigate prior to the analysis, through plotting histograms or boxplots. If the explanatory variable shows skew, which could be alleviated by a transformation, that will automatically remove the analysis from falling into the heteroscedasticity without outliers setting studied here, although severe heteroscedasticity can still inflate Type I error rates without skew in the covariate as shown in Croux et al. (2003). If the explanatory variable shows outliers, using a priori cutoffs, subjects that appear to be outliers should be considered for removal. Models with and without that subject should be compared. In situations like this when sample sizes are small, omitting one outlier usually reveals more outliers, which is an unfortunate issue with small studies and nothing can be done to alleviate this problem and the limitations of the study must be realized. Likely the best strategy, which yields limited modeling options and interpretation, is to use the Kendall's Tau for very small sample sizes.

Although it may seem impossible to study the explained variable in the same amount of detail, if the subject-specific contrast images are concatenated into a single 4D image and a "movie" is played in a viewer, such as fslview, outlying subjects may be observed. Special attention should be given to voxels in regions that showed significance. This can also be done with the within-subject variance estimates. When outliers occur, it could be the case that this subject was flagged earlier in the analysis as a potential outlier, possibly due to poor task performance or high motion. It is not recommended that a subject be discarded, simply

because the data appear to be outlying, but this could be a good indication that a robust regression, Flame 1 or Flame 1 with outlier de-weighting should be considered.

Residuals from the model can also be viewed as movies to try and identify potential outliers. In the case of a regression model with a single regresor, if the subjects are ordered according to their regressor value, the residual plots as shown in Figure 1 can easily be created if the image viewer used also displays the time series plots.

Avoid p-hacking

Unfortunately, the advice given here may lead to comparing the results from various models to determine whether an outlier was influencing the results. It is heavily encouraged to not simply choose the model with the lowest p-value as "best", but to carefully consider the results. If it is a weak effect that slightly tips from $p < 0.05$ to $p > 0.05$, proceed with caution and present any result with honesty and clarity. It is not encouraged that all of these models be applied to the same data set. Whatever modeling approaches are used with the data, it should be reported in the resulting manuscript, both the model the reported results were based on as well as any other models used.

Weighted regression models

Both the Robust and Flame-based approaches operate by differentially weighting subjects. An important consideration when using these types of approaches is looking at who was downweighted in the analysis. In some cases, a subpopulation of the study may have been greatly downweighted, in which case the interpretation of the results may change. For example, in a behavioral analysis with a depression index covariate, which was skewed regardless of transformation, when robust regression was applied, all of the highly depressed subjects were heavily downweighted in the analysis, implying that results were likely more appropriate within a subpopulation with lower depression and should not be interpreted for the high depression group.

Limitation of simulations

The simulations created here focused on single voxels, whereas typically whole brain analyses using cluster-wise inference procedures are used. Due to the spatial smoothness of fMRI data, it is unlikely that an outlier will exist only in isolated voxels, but in sets of voxels. For example, in Fritsch et al. (2015), an influential outlier that caused a false positive linear relationship when using OLS was consistent across many voxels.

Kendall's Tau

In many small sample studies, it is often tempting to use Kendall's Tau and these results show that in most cases the Type I error rate is fairly well controlled, but often with a large decrease in power. Unfortunately, in some cases the Type I errors are a bit inflated and, more importantly, the interpretation can be difficult as the shape of the relationship between the explanatory and explained variable can not be concluded without looking at the data. The only conclusion that can be made is the relationship is monotonic. Lastly, this approach limits to ability to control for other possible confounding variables.

Conclusion

There is not a single model that can handle all types of heteroscedasticity for all sample sizes, although for large sample sizes (>500), the doubly robust model is a good alternative. The worst scenarios, in terms of controlling Type I error, involved issues with the explanatory variable, which is simple to visually inspect. If the explanatory variable is skewed, a transformation should be considered. If there are outliers in the explanatory variable, they should be considered from removal, since even the robust regression approaches fail to control Type I error in the presence of bivariate outliers. By avoiding skew and outliers in the explanatory variable, it is more likely that inferences will be valid. Although difficult, it is not impossible to visually inspect imaging data to look for potential outliers and it is highly encouraged. If outliers are driven by the within-subject variance, the FSL's Flame 1 algorithm can handle any type of heteroscedasticity and so one approach would be to use Flame 1 and then study the residuals from the model to see whether there are any outliers driven by the between-subject variance. If outliers seem to be present, without any reason to simply omit the subjects, using an approach such as robust regression or Flame 1 with outlier de-weighting can be useful. To alleviate "p-hacking", be clear about all models used in the study when reporting results and do not simply choose the result with the smallest p-value as this could be a Type I error.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grant P30HD003352.

References

Beckmann CF, Jenkinson M, Smith SM. General multilevel linear modeling for group analysis in FMRI. Neuroimage. 2003; 20(2):1052–1063. [PubMed: 14568475]

Chen G, Saad ZS, Nath AR, Beauchamp MS, Cox RW. FMRI group analysis combining effect estimates and their variances. Neuroimage. 2012; 60(1):747–765. [PubMed: 22245637]

Croux, C., Dhaene, G., Hoorelbeke, S. Discussion Paper Series 03.16. Center for Economic Studies, Catholic University of Leuven; 2003. Robust standard errors for robust estimators.

Da Mota B, Fritsch V, Varoquaux G, Banaschewski T, Barker GJ, Bokde AL, Bromberg U, Conrod P, Gallinat J, Garavan H, Martinot JL, Nees F, Paus T, Pausova Z, Rietschel M, Smolka MN, Strohle A, Frouin V, Poline JB, Thirion B. Randomized parcellation based inference. Neuroimage. 2014; 89:203–215. [PubMed: 24262376]

Eklund, A., Nichols, T., Knutsson, H. Can parametric statistical methods be trusted for fMRI based group studies?. 2015. arXiv:1511.01863

Fritsch V, Da Mota B, Loth E, Varoquaux G, Banaschewski T, Barker GJ, Bokde AL, Bruhl R, Butzek B, Conrod P, Flor H, Garavan H, Lemaitre H, Mann K, Nees F, Paus T, Schad DJ, Schumann G, Frouin V, Poline JB, Thirion B. Robust regression for large-scale neuroimaging studies. Neuroimage. 2015; 111:431–441. [PubMed: 25731989]

Hahn, S., Konietschke, F., Salmaso, L. A comparison of efficient permutation tests for unbalanced ANOVA in two by two designs–and their behavior under heteroscedasticity; 2013. p. 1-20.arXiv: 1309.7781http://arxiv.org/abs/1309.7781

Hayes A. Permutation test is not distribution free: testing H0:ρ = 0. Psychological Methods. 1996; 1:184–198.

Huber, PJ. Robust Statistics. Wiley; New York: 1981.

Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage. 2002; 17(2):825–841. [PubMed: 12377157]

Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Med Image Anal. 2001; 5(2):143–156. [PubMed: 11516708]

Lejuez CW, Read JP, Kahler CW, Richards JB, Ramsey SE, Stuart GL, Strong DR, Brown RA. Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). J Exp Psychol Appl. 2002; 8(2):75–84. [PubMed: 12075692]

Mumford JA, Nichols T. Modeling and inference of multisubject fMRI data. IEEE Eng Med Biol Mag. 2006; 25(2):42–51. [PubMed: 16568936]

Mumford JA, Nichols T. Simple group fMRI modeling and inference. Neuroimage. 2009; 47(4):1469–1475. [PubMed: 19463958]

Rasmussen J. Computer-intensive correlational analysis: Bootstrap and approximate ranodmization techniques. British Journal of Mathematical and Statistical Psychology. 1989; 42:103–111.

Smith S, Jenkinson M, Beckmann C, Miller K, Woolrich M. Meaningful design and contrast estimability in FMRI. Neuroimage. 2007; 34(1):127–136. [PubMed: 17070706]

Smith SM. Fast robust automated brain extraction. Hum Brain Mapp. 2002; 17(3):143–155. [PubMed: 12391568]

Wager TD, Keller MC, Lacey SC, Jonides J. Increased sensitivity in neuroimaging analyses using robust regression. Neuroimage. 2005; 26(1):99–113. [PubMed: 15862210]

Woolrich M. Robust group analysis using outlier inference. Neuroimage. 2008; 41(2):286–301. [PubMed: 18407525]

Woolrich MW, Behrens TE, Beckmann CF, Jenkinson M, Smith SM. Multilevel linear modelling for FMRI group analysis using Bayesian inference. Neuroimage. 2004; 21(4):1732–1747. [PubMed: 15050594]

Woolrich MW, Ripley BD, Brady M, Smith SM. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage. 2001; 14(6):1370–1386. [PubMed: 11707093]

Appendix A. Simulation settings for heteroscedasticity without outliers

Table A.1

Settings for simulation of heteroscedasticity without outliers. Using the real data, the ratio of the mixed effects variance for subjects whose group level covariate was below to above the median was split into 20 percentiles (first column). Within each percentile the average within and between subject variance was estimated for the low and high explosion groups (columns 2–4). The simulations were based on the between-subject variances in this table and the average within-subject variance (4788).

Percentile range of mfx variance ratio	$\sigma^2_{b,below}$	$\sigma^2_{b,above}$	$\sigma^2_{w,below}$	$\sigma^2_{w,above}$
[0.0362, 0.581)	15468	2580	7321	6558
[0.581, 0.707)	6370	1873	5093	5527
[0.707, 0.799)	4570	1709	4634	5223
[0.799, 0.879)	3627	1628	4340	5056
[0.879, 0.948)	2906	1549	4205	4948
[0.948, 1.01)	2415	1476	4041	4854

Percentile range of mfx variance ratio	$\sigma^2_{b,below}$	$\sigma^2_{b,above}$	$\sigma^2_{w,below}$	$\sigma^2_{w,above}$
[1.01, 1.07)	2132	1513	3840	4706
[1.07, 1.13)	1657	1300	3680	4563
[1.13, 1.18)	1483	1345	3496	4384
[1.18, 1.22)	1289	1256	3360	4324
[1.22, 1.28)	1179	1259	3295	4334
[1.28, 1.33)	1241	1495	3377	4520
[1.33, 1.4)	1256	1665	3486	4795
[1.4, 1.48)	1354	2023	3586	5069
[1.48, 1.58)	1503	2523	3766	5520
[1.58, 1.71)	1615	3215	3847	5765
[1.71, 1.9)	1741	4073	4055	6368
[1.9, 2.19)	1852	5313	4168	6922
[2.19, 2.77)	1896	7344	4374	7939
[2.77, 40.5)	2018	16245	4525	10080

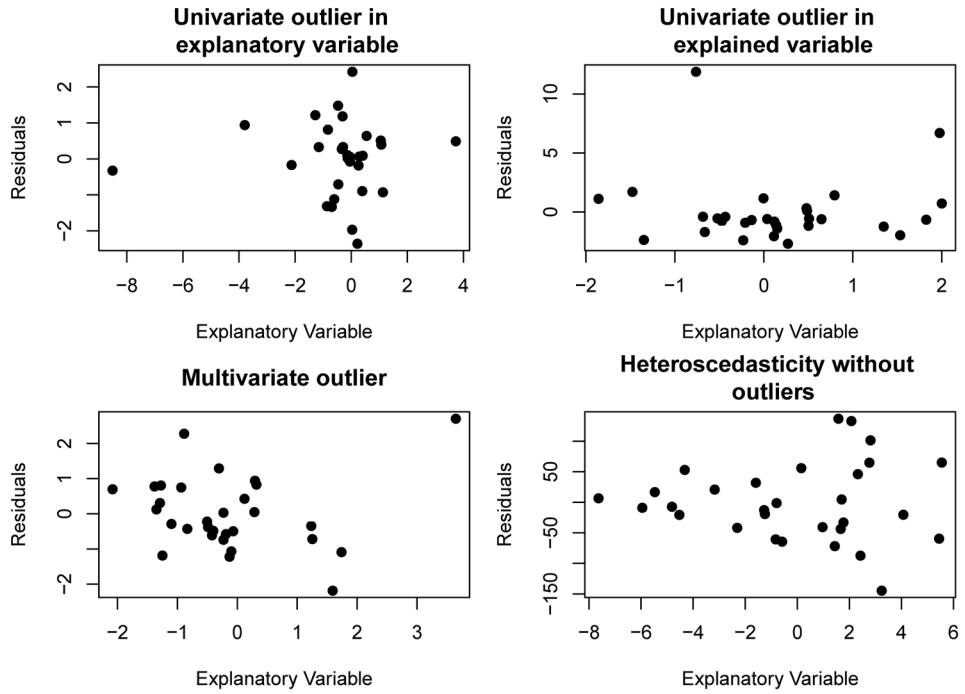

Figure 1.

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.

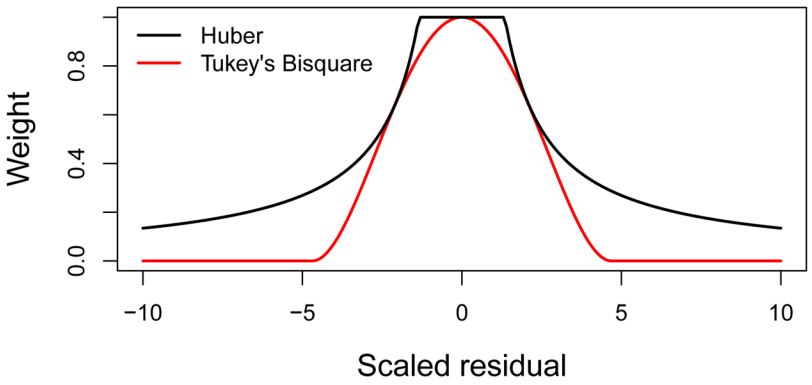

Figure 2.
Plots of the Huber's (black) and Tukey's Bisquare (red) loss functions.

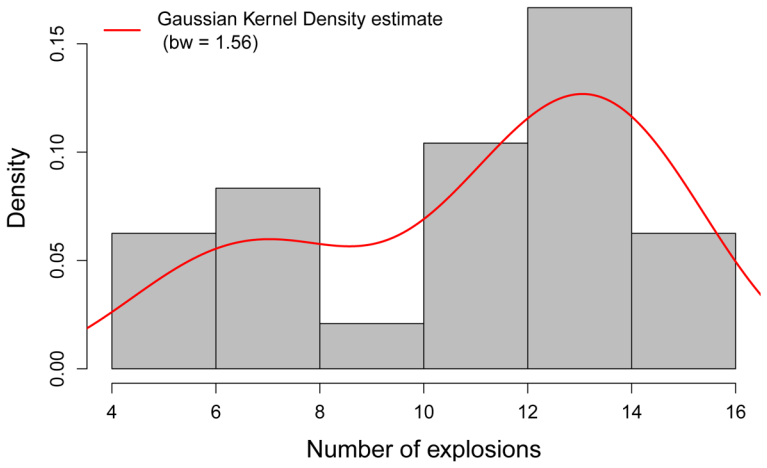

Figure 3.

Distribution of number of balloon explosions from real data analysis with Gaussian kernel density estimate. Samples from the kernel density estimate were used in simulation study.

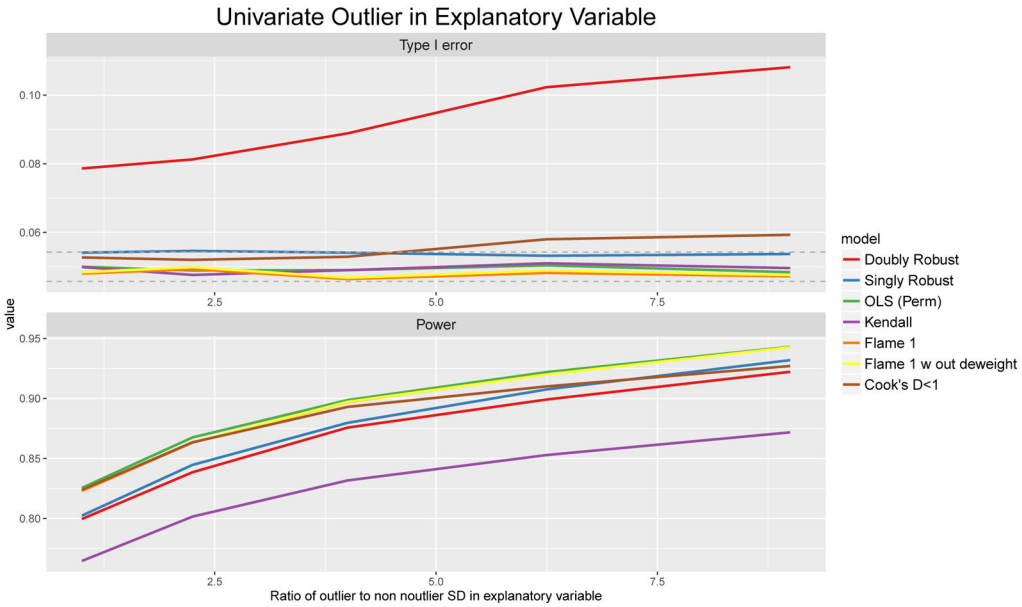

Figure 4.

Univariate outlier in explanatory variable for a sample size of 30 with 10% outliers. The top panel shows Type I error rates, while the bottom shows power. The horizontal dotted lines indicate 95% confidence intervals around the Type I error of 0.05. The x-axis is the ratio of the standard deviation of the explanatory variable for the outliers compared to non-outliers. Note, Power levels can only be considered if there is Type I error control.

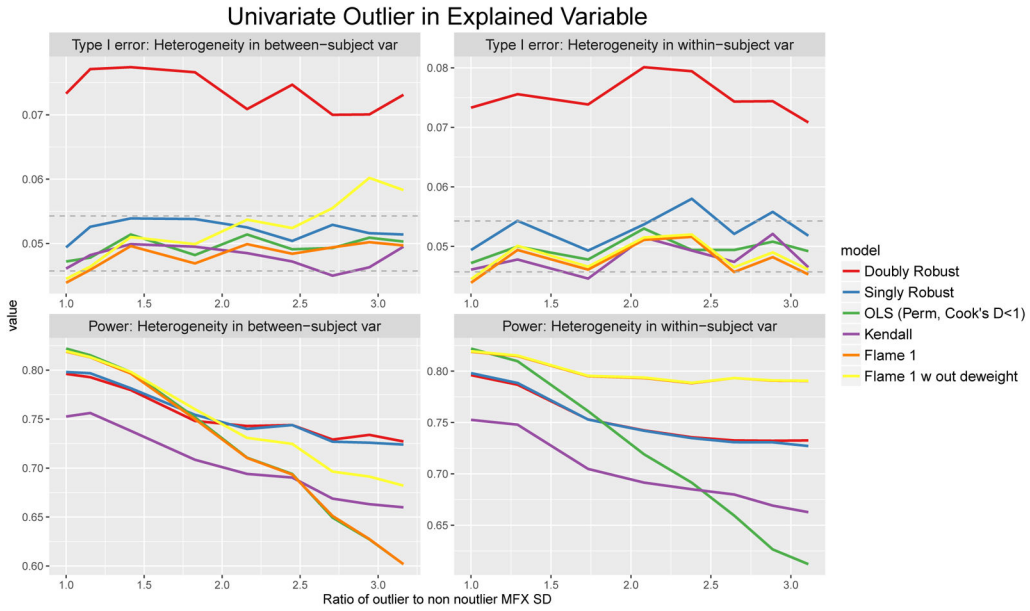

Figure 5.

Univariate outlier in explained variable for a sample size of 30 with 10% outliers. The top panels show Type I error rates, while the bottom show power. The x-axis is the ratio of the mixed effects variance, $\sigma^2_{w,i}+\sigma^2_b$, for outliers, compared to non-outliers. The horizontal dotted lines indicate 95% confidence intervals around the Type I error of 0.05. Note that permutation and OLS often have very similar results, so the permutation result isn't plotted.

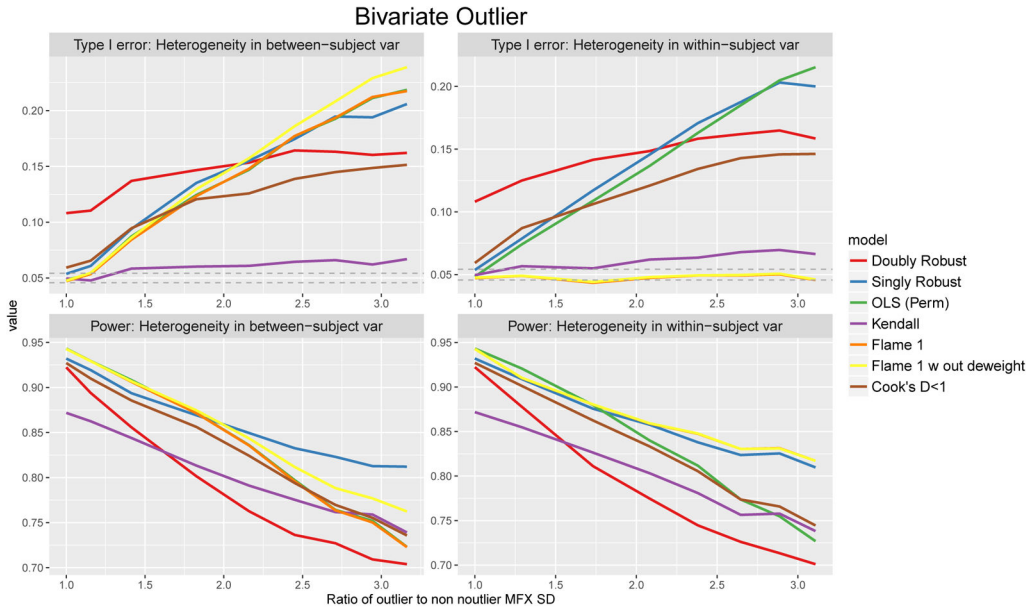

Figure 6.

Bivariate outlier for a sample size of 30 with 10% outliers. The standard deviation was three times higher for the outlier in the explanatory variable, compared to the non-outliers. The top panels show Type I error rates, while the bottom show power. The x-axis is the ratio of the mixed effects variance, $\sigma^2_{w,i}+\sigma^2_b$, for outliers, compared to non-outliers. The horizontal dotted lines indicate 95% confidence intervals around the Type I error of 0.05.

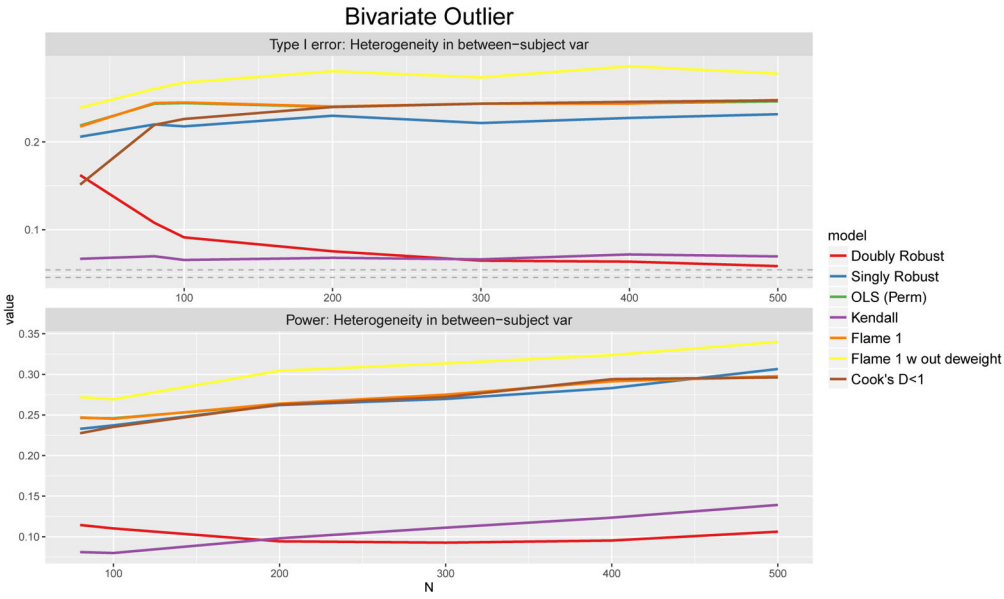

Figure 7.

Bivariate outlier over a range of sample sizes when the ratio of the outlier to non-outlier mixed effects standard deviation is 3 for both the explanatory and explained variables in 10% of the subjects. The horizontal dotted lines indicate 95% confidence intervals around the Type I error of 0.05. The doubly robust approach has improved Type I error control as the sample size increases.

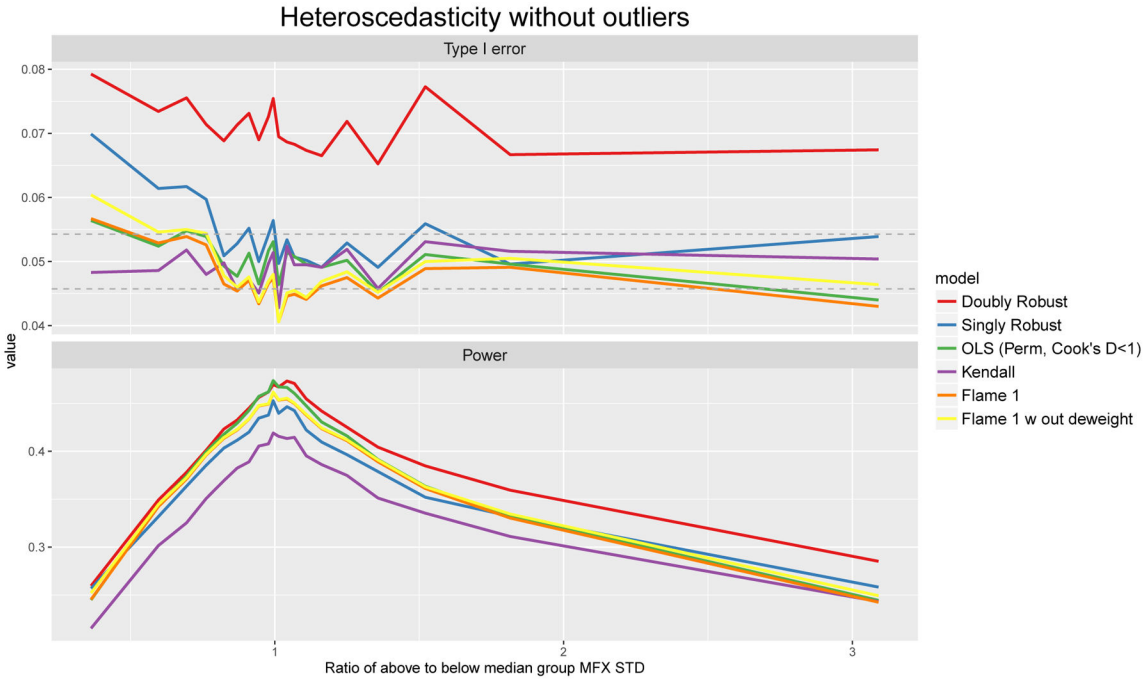

Figure 8.

Type I error and Power for the case of heteroscedasticity without outliers for 30 subjects. The x-axis reflects the ratio of the mixed effects variance for the above median group, compared to the below median group. The horizontal dotted lines indicate 95% confidence intervals around the Type I error of 0.05.