Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Econometrics and Management Science

The Robustness of Heteroscedasticity Tests

Agnieszka Pechcińska (638592)

Supervisor:	dr. Mikhail Zhelonkin
Second assessor:	prof. dr. Richard Paap
Date final version:	31st July 2023

Abstract

Homoscedastic error terms are a central assumption for the Ordinary Least Squares (OLS) estimator to be the best linear unbiased estimator. Diagnostic tests are applied to verify this assumption. However, the classical heteroscedasticity tests can be easily misled by outliers, and the available alternatives apply the outlier removal strategy, which is not desirable. We, therefore, investigate the robustness properties of the three classical heteroscedasticity tests and propose a robust alternative with the score test framework of Heritier and Ronchetti (1994). The robustness properties are supplemented with a simulation study and empirical applications to real datasets. We show that the robust alternative outperforms the classical test, remains size-correct and retains its power in the presence of outliers.

Keywords: Heteroscedasticity; Robust test; Score test; Outlier

Contents

1 Introduction

The Ordinary Least Squares (OLS) estimator of the linear regression model is the best linear unbiased estimator if the Gauss-Markov assumptions are met (Greene, 2003). Among them, there are the expected value of error terms equal to zero and the spherical variance-covariance matrix of the error terms. The latter assumption translates into homoscedastic, that is, having constant variance across observations, and not serially correlated error terms. For reliable inference and predictions based on the OLS estimates, it is necessary to check whether the assumption of spherical errors is not violated.

Even if the OLS estimator remains unbiased when the assumption is not met, the standard errors may be underestimated. Consequently, the statistical tests based on them, such as t- and F-tests, may, for example, overestimate the significance of the regression coefficients (Verbeek, 2004). The OLS estimator is no longer the best one. To control for the presence of heteroscedastic error terms, a wide range of classical diagnostic tests has been proposed.

The classical parametric diagnostic tests are devised to work well when underlying assumptions are met, for example, distributional assumptions about error terms. If the sample contains outlying observations, the outliers may inflate the variance of some residuals to the extent that the classical heteroscedasticity test is no longer able to reject the null hypothesis of homoscedastic error terms for heteroscedastic data. Outliers may also mask the real homoscedastic error terms with falsely inflated variance and result in over-rejecting the null hypothesis. The methods of robust hypothesis testing from the field of robust statistics offer an appealing alternative to the classical heteroscedasticity tests, ensuring correct inference about the underlying process in the regression residuals.

Therefore, in this paper, we investigate the behaviour of classical heteroscedasticity tests in the presence of outliers of different types and verify whether the robust alternative to the Breusch-Pagan test offers both the robustness of validity and robustness of efficiency.

This research is both of scientific relevance and interest for practical applications. We apply the theoretical framework developed by Heritier and Ronchetti (1994) to a new group of tests, heteroscedasticity tests. The heteroscedastic error terms are present in economic and social data, thus appropriate testing methods are of interest to practitioners from different fields of applied science (Greene, 2003).

To investigate the robustness of classical heteroscedasticity tests, we select three tests easily available for practitioners in the statistical software, for example in the R package *lmtest* (Zeileis & Hothorn, 2002), the Breusch-Pagan test (1979), the Goldfeld-Quandt test (1965) and the Harrison-McCabe test (1979). We show their robustness properties and verify them against vertical and bad leverage outliers in sensitivity analysis and a simulation study of the power and the level of the test. As it turns out that none of the tests considered preserve robustness against both types of outliers for both homoscedastic and heteroscedastic data, we aim to find a robust counterpart.

Although there is some work on the robust alternatives to classical heteroscedasticity tests (Alih & Ong, 2015; Berenguer-Rico & Wilms, 2021; Rana et al., 2008), all apply outlier removal strategy which does not preclude the swamping effect and may result in the deletion of not outlying observations distorting the sample distribution. Consequently, such a constructed heteroscedasticity test may wrongly detect heteroscedasticity when data is homoscedastic. We address this gap with the development of the robust heteroscedasticity test based on the framework of a bounded-influence score test derived by Heritier and Ronchetti (1994) and we modify it to the specificities of heteroscedasticity testing.

We evaluate three classical heteroscedasticity tests, but the construction of the robust alternative is focused solely and entirely on the alternative to the Breusch-Pagan test. The robustness of the newly proposed test is evaluated first with the simulation study of the level and the power of the test in the presence of outliers. We perform the simulation in the setting of a large sample size and do not investigate the small-sample properties of the robust test. Next, the test is applied to real datasets of credit card data (Greene, 1992), and teacher ratings data (Hamermesh & Parker, 2005), and we verify whether the robustness of efficiency and the robustness of validity are preserved. Our results show that the heteroscedasticity score test constructed with Mallows type M-estimator (Huber, 1973; Mallows, 1975) remains size-correct and powerful in the simulated presence of outliers. Its robustness is also confirmed in real dataset applications.

In Section 2 we discuss relevant literature in the field of classical and robust heteroscedasticity tests. Section 3 presents the methodology of the classical heteroscedasticity tests and their robustness properties, which are verified with a simulation study. In Section 4 we propose the robust heteroscedasticity test and describe the design and results

of a simulation study that validate the robustness of the proposed test. In Section 5 we conduct an empirical application of selected classical and robust tests in real datasets. Section 6 concludes with the main findings, the limitations of our research and further research possibilities.

2 Literature

In this section, we provide an overview of the literature on heteroscedasticity testing. We start with a general motivation for why the heteroscedasticity tests are of great importance for reliable inference in regression models. Next, we concentrate on both classical and robust approaches to the heteroscedasticity tests, with an emphasis on identifying gaps in the development of robust alternatives.

The classical heteroscedasticity tests applied in the regression diagnostic in the estimation with OLS include, among others, the Breusch-Pagan test (1979), the Glejser test (1969), the Goldfeld-Quandt test (1965), the Harrison-McCabe test (1979), the Harvey test (1974), the jackknife tests (Sharma & Giaccotto, 1991), and the White test (1980). Despite the wide range of available homoscedasticity tests, Dufour et al. (2004) note that the practitioners tend to prefer the Breusch-Pagan test, the White test and the Goldfeld-Quandt test, with the clear dominance of the first one. All mentioned tests are devised to work well when the assumptions of the test are met, for example, distributional assumptions about error terms. However, the presence of outlying observations in the data may contribute to the failure to meet the basic assumptions underlying the tests and consequently lead to erroneous conclusions. For example, the data is heteroscedastic, but in the presence of outliers, the heteroscedasticity tests fail to detect the true nature of error terms, and they wrongly indicate homoscedasticity with no power to reject the null hypothesis.

Lyon and Tsai (1996) analyse the behaviour of different classical heteroscedasticity tests when there are either outlying errors or outlying explanatory variables. They find that the heteroscedasticity tests based on the likelihood ratio tests perform poorly, in contrast to the modified version of the Breusch-Pagan test (also known as the Koenker score test (1981), see Section 3.1.1) which remains robust in those settings. However, Lyon and Tsai (1996) do not verify the level and the power of those tests in the presence of bad leverage or vertical outliers. The earlier work by Ali and Giaccotto (1984) evaluates

the performance of both parametric and nonparametric heteroscedasticity tests when the distributional and independence assumptions about the errors are not met. They find that generally the power is substantially reduced if the errors do not follow the normal distribution, but in the presence of either long-tailed or skewed distributions, the Glejser test and the White test remain robust. The Breusch-Pagan test and the Goldfeld-Quandt test fail in such scenarios.

To address the lack of robustness of the classical heteroscedasticity test, Alih and Ong (2015) and Rana et al. (2008) develop robust versions of the Goldfeld-Quandt test. In both papers, the authors apply the outlier-removal strategy, Alih and Ong (2015) use the Mahalanobis distance to identify and exclude outliers from the next steps, while Rana et al. (2008) employ the Least Trimmed Squares (LTS) estimator for the same purposes. In the following steps, in both works, 'clean' samples are used to construct the test statistic in the form of a ratio of certain measures of squared residuals. However, neither Alih and Ong (2015) nor Rana et al. (2008) substantiate their work with theoretical derivations. Moreover, in both procedures, the final result of the test is based on the dataset in which the outliers are deleted, but with this approach, we cannot be sure whether the swamping effect does not contribute to deleting good points (see Appendix A.1 for an example of a simulation study of the modified Goldfeld-Quandt test (Rana et al., 2008) showing that the test is not size-correct). Berenguer-Rico and Wilms (2021) find that the outlier removal results in a low power or oversized (or undersized depending on the outlier type) White test, which we suspect may also be the case with the Goldfeld-Quandt test modified with the outlier-removal approach.

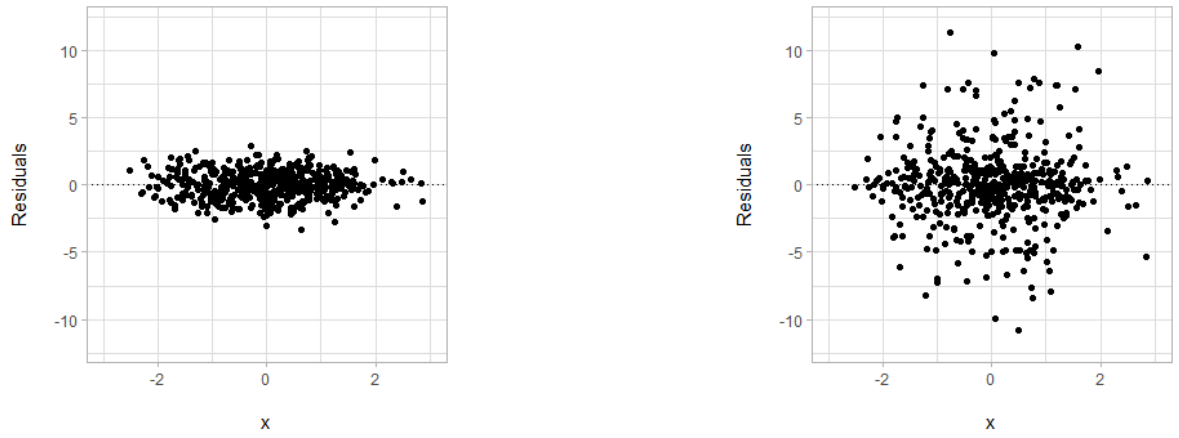
We extend the research on robust heteroscedasticity tests by applying methods from the field of robust statistics to the classical Breusch-Pagan test. Heritier and Ronchetti (1994) develop the robust score test based on M-estimator (Huber, 1973) which ensures a stable test level under small departures from the null hypothesis and a stable test power under small departures from alternative hypotheses. So far, this framework has not been adapted to the specificities of heteroscedasticity testing but has proven effective in the application to testing general restrictions in nonlinear and logistic regression in the presence of different types of outliers.

3 Classical heteroscedasticity tests

In this section, we focus on the classical heteroscedasticity tests. We start with a general overview of the available methods for testing heteroscedastic error terms, followed by a detailed explanation of the methodology of the three selected tests, together with a discussion of their robustness. We finalize with a simulation study investigating the robustness of the power and the level of selected tests.

3.1 General overview

The visual inspection of regression residuals is often the first step to control for heteroscedasticity. However, the conclusions drawn from such plots, see Figure 3.1, are subjective and prone to human error, thus they should be followed by the application of diagnostic tests. Greene (2003, Chapter 11) notes that most of the tests are applied to the OLS residuals since the OLS estimator remains consistent also in the presence of heteroscedastic error terms. Therefore, the assumption is made that these residuals mimic the heteroscedasticity of the true disturbances well enough.



(a) $e_i \sim \mathcal{N}(0, \sigma^2)$ *(b)* $e_i \sim \mathcal{N}(0, \sigma_i^2)$

Figure 3.1. Example of visual inspection of regression residuals with the plot of explanatory variable x against residuals. Residuals from the linear model, $y_i = x_i + e_i$ $(i = 1, \ldots, 500)$, with (a) homoscedastic error terms and (b) heteroscedastic error terms.

Among the classical heteroscedasticity tests, two main approaches to studying the variance of the error terms can be distinguished. These approaches vary in terms of how specific the researcher needs to be when stating the type of suspected heteroscedasticity in the data and thus formulating the alternative hypothesis. Tests, postulating that in the

data one can identify certain groups of observations between which the variance differs, but within those groups it remains constant, do not require specifying the function that drives the variance in each of those groups. The alternative hypothesis states that the variances are not equal between groups. The Goldfeld-Quandt test (1965) and the Harrison-McCabe test (1979) check whether variance differs between two groups, while with the likelihood ratio test (Fomby et al., 1984) or the jackknife test (Sharma & Giaccotto, 1991) the equal variance in more than two groups of observations can be tested. The likelihood ratio test may also constitute the second approach, namely the group of tests that postulate the variance of a certain form, for example, the Harvey test (1976) for multiplicative heteroscedasticity. Other tests assuming a certain form of variance function include, i.a. the Breusch-Pagan test (1979), the White test (1980) or the Glejser test (1969). All three require the estimation of the auxiliary regression, which estimates the parameters of the variance function. The null hypothesis of these tests postulates that the parameters associated with the variables driving the variance are equal to zero and the variance is constant.

In this study, we focus on three classical heteroscedasticity tests that are available in the R package *lmtest* (Zeileis & Hothorn, 2002), which is widely used for diagnostic checking in linear regression models. The tests include the Breusch-Pagan test (1979), the Goldfeld-Quandt test (1965), and the Harrison-McCabe test (1979). The methodology behind each of these tests is described in the subsequent sections.

3.1.1 Breusch-Pagan test

Breusch and Pagan (1979) propose a score test, which implies the OLS estimation of two regressions. We consider the linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + e_i, \text{ for } i = 1, ..., N, \tag{1}$$

where $e_i \sim \mathcal{N}(0, \sigma_i^2)$. The variance of e_i is expressed with a continuous function h

$$\sigma_i^2 = h(z_i^\top \boldsymbol{\alpha}), \tag{2}$$

where $z_{ij}^\top = (z_{i1}, \ldots, z_{im})$, $z_{i1} = 1$ for all i, and z_{ij}'s $(j = 2, \ldots, m)$ are known constants. $z_i^\top$ should be the suspected drivers of the errors' variance. They may be the regressors

from the model (1) or some known functions of them, for example, squares or interactions. The $\boldsymbol{\alpha}^\top = (\alpha_1, \ldots, \alpha_m)$ are unknown parameters. Assuming that e_i follows the normal distribution, the null hypothesis reads $H_0 : \alpha_2 = \ldots = \alpha_m = 0$, implying homoscedasticity. The test statistic is

$$\frac{1}{2}\left(\sum_{i=1}^{N} z_i u_i\right)^\top \left(\sum_{i=1}^{N} z_i z_i^\top\right)^{-1} \left(\sum_{i=1}^{N} z_i u_i\right), \tag{3}$$

where $u_i = \frac{\hat{e}_i^2}{\hat{\sigma}^2} - 1$, $\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{e}_i^2}{N}$ and $\hat{e}_i$ are the residuals from the regression model (1) estimated with OLS (Breusch & Pagan, 1979). The auxiliary regression in Equation (2) is also estimated with OLS. Under the null, the test statistic asymptotically follows the χ^2_{m-1} distribution. For the full derivation of the test statistic, see Breusch and Pagan (1979).

Koenker (1981) showed that the Breusch-Pagan test is sensitive to the assumption of normality. If the error terms do not follow a Gaussian distribution, the asymptotic size of the test is incorrect and, similarly, the asymptotic power is sensitive to the distributional assumptions. Therefore, he constructed a modified test statistic, which also follows χ^2 distribution, but it is based on a more robust estimator of the variance of squared residuals. When e_i does not follow the normal distribution, the modified statistic provides a more powerful test. It can also be noted that in the modified test, we can compute the test statistic as NR^2, where R^2 is a coefficient of determination from auxiliary regression in Equation (2) estimated with OLS.

Given the better properties of the modified test statistic at small deviations from the normal distribution, also confirmed with the simulation study carried out by Lyon and Tsai (1996), throughout this paper we consider the version of the test statistic with the correction proposed by Koenker (1981) when referring to the Breusch-Pagan test.

3.1.2 Goldfeld-Quandt test

Goldfeld and Quandt (1965) offer a different approach to testing for the presence of heteroscedastic error terms. They propose a variance ratio test, in which a test statistic is constructed as a ratio of the sum of squared residuals from two separate regressions.

We consider the linear regression model as defined in Equation (1). The researcher should be able to order the observations in the sample according to either the value of one

of the regressors x_k, suspected to drive the variance of error terms, or time if the variance is suspected to change over time. After ordering, the sample is split into two subsets, n_1 and n_2 where $n_1 + n_2 = N$, with one of the subsets containing observations associated with a smaller variance of error terms than in the other. Then for each of the subsets, n_1 and n_2, the regression is estimated with OLS and the sum of squared residuals from both regressions is computed. The test statistic is

$$R = \frac{S_2/n_2}{S_1/n_1}, \tag{4}$$

where S_2 an S_1 are the sums of squared residuals from the subsamples containing the observations with the larger and the smaller variances of error terms, respectively. It is assumed that error terms are independently and normally distributed. Under the null hypothesis of equal variances in both subsets, the test statistic follows the $\mathcal{F}$ distribution with n_1 and n_2 degrees of freedom.

The originally proposed test postulates to omit a certain number of central observations, c_N, and then estimate the regressions in both subsamples of smaller and larger variances. Goldfeld and Quandt (1965) underline that the power of the test depends on the choice of c_N, potentially leading to high power when c_N is small or close to zero. However, for such c_N the difference between variances in two groups might be also negligible, and as a result, it may decrease the power of the test. The authors do not draw a clear conclusion about which effect predominates and what is the optimal value of c_N. There exist several rules of thumb on how to determine c_N, for example, Goldfeld and Quandt (1965) suggest $c_N \approx 0.27N$, while Harvey and Phillips (1974) propose $c_N = \frac{N}{3}$, but no common approach is established. In general, if we expect that there exists a single breakpoint at which the variances change, all observations can be used and $c_N = 0$. In other cases, it depends on the researcher's decision.

3.1.3 Harrison-McCabe test

Harrison and McCabe (1979) propose an approach similar to the Goldfeld-Quandt heteroscedasticity test. The test statistic is also constructed as a ratio, but it requires the OLS estimation of only one regression.

We consider the linear regression model as defined in Equation (1). We estimate this model with OLS using the whole sample of N observations and we compute the sum of

squared residuals. The test statistic is constructed as

$$b = \frac{S_1}{S}, \tag{5}$$

where S_1 and S are the sums of squared residuals from a certain subset of observations, n_1, and the whole sample, N, respectively. Assuming that the error terms are independent and follow a normal distribution, the null hypothesis postulates homoscedastic error terms. Under the null hypothesis, the test statistic b should be close to the ratio of n_1 and N (Krämer & Sonnberger, 1986). Harrison and McCabe (1979) show that a test criterion to reject the null hypothesis is constructed from the bounding distributions of two random variables that bound the test statistic b. The proposed bounds test has an inconclusive region if b is between the lower and upper bound. Therefore, the bound test might be supplemented with the exact test obtained with the Imhof method (Imhof, 1961) for the computation of the distribution of a ratio of quadratic forms in normal random vectors.

The size of n_1 depends on the suspected form of heteroscedasticity postulated in the alternative hypothesis. Similarly to the Goldfeld-Quandt test, either the variance of error terms varies across time or it is an increasing function of one of the regressors x_k. In the first case, S_1 should contain residuals associated with chronologically first n_1 observations, while in the second case, S_1 should correspond to the n_1 smallest values of the suspected regressor x_k. The choice of the exact size of n_1 is arbitrary, however, Harrison and McCabe (1979) suggest that if there is no a priori knowledge about the type of heteroscedasticity of the error terms in the model, the correct choice is n_1 equal to half of the sample.

3.2 Robustness properties

The Breusch-Pagan test, the Goldfeld-Quandt test and the Harrison-McCabe test require estimation of the linear regression with the Ordinary Least Squares estimator to construct a test statistic which is based on certain values obtained from the estimated regression model, for example, the sum of squared residuals. Heritier and Ronchetti (1994) note that the robustness of the test statistic is inherited from the robust estimator used. One of the tools suitable to assess the robustness of the estimator underlying each test is the influence function of an estimator (Hampel, 1974), which measures the effect of infinitesimal point mass contamination on the estimate. The influence function (Hampel, 1974)

at the distribution F is defined as

$$IF\left(\mathbf{w}; T, F\right) = \lim_{\epsilon \to 0+} \frac{T\left\{\left(1-\epsilon\right)F + \epsilon \Delta_{\mathbf{w}}\right\} - T\left(F\right)}{\epsilon}, \tag{6}$$

where T is a statistical functional, ϵ is a contamination level, and $\Delta_{\mathbf{w}}$ is a point mass at $\mathbf{w}$, where $\mathbf{w}$ is a vector of observations. In other words, the influence function provides information about the relative influence of a small proportion of outliers on the value of an estimate (Huber & Ronchetti, 2009). The robustness of the regression estimators requires the bounded influence function so that the estimates are stable under local perturbations (Hampel et al., 1986).

To evaluate the robustness of the classical heteroscedasticity tests, we consider the influence function of the OLS estimator entrenched in these procedures. For this purpose, we consider the regression model from Equation (1), which can be also written as

$$\mathbf{y}_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \boldsymbol{e}_i, \text{ for } i = 1, \ldots, N, \tag{7}$$

where $\mathbf{x}_i$ and $\boldsymbol{\beta}$ denote $k+1$ dimensional vectors $(1, x_{i1}, \ldots, x_{ik})$ and $(\beta_0, \ldots, \beta_k)$, respectively. Following Hampel et al. (1986), the influence function of the OLS estimator can be written as

$$IF\left(\left\{\mathbf{x}^\top, \mathbf{y}\right\}; T, F\right) = \boldsymbol{Q}^{-1}\left(\mathbf{y} - \mathbf{x}^\top \beta\right)\mathbf{x}, \tag{8}$$

where T is the functional defining the corresponding OLS estimator and $\boldsymbol{Q} = \int \mathbf{x}\mathbf{x}^\top dF$. From (8) we conclude that the OLS estimator is not robust to the contamination in both explanatory space and the dependent variable, since the influence function is unbounded in $\mathbf{x}$ and $\mathbf{y} - \mathbf{x}^\top \beta$. The considered heteroscedasticity test statistics inherit the robustness properties of the OLS estimator they are constructed upon. One single vertical outlier or a bad leverage point leads to a bias in the OLS estimates, resulting in the incorrect value of the test statistic in the classical heteroscedasticity test. The outliers can mask heteroscedasticity of error terms and, consequently, the test has no power to reject the null hypothesis. This issue is further investigated with the simulation study, in which we check the robustness of validity (i.e., the level of a test is stable under small, arbitrary departures from the null hypothesis) and the robustness of efficiency (i.e., the test has good power under small, arbitrary departures from the specified alternative hypotheses) of three classical heteroscedasticity tests.

3.3 Simulation design

In this section, we devise a simulation framework to investigate the robustness properties of three classical heteroscedasticity tests: the Breusch-Pagan test (1979) with the Koenker's correction (1981), the Goldfeld-Quandt test (1965) and the Harrison-McCabe test (1979). In the simulation study, the following linear model with two regressors and intercept is considered:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \text{ for } i = 1, \ldots, N, \tag{9}$$

where $x_{i1} \sim \mathcal{U}_{[1,10]}$, $x_{i2} \sim \mathcal{U}_{[-5,5]}$, $\beta_0 = \beta_1 = \beta_2 = 1$ and $e_i \sim \mathcal{N}(0, \sigma_i^2)$. The variance of e_i depends on the analysed scenario. For homoscedastic data $\sigma_i^2 = \sigma^2 = 1$. For heteroscedastic data, we investigate two types of heteroscedasticity: (I) $\sigma_i^2 = \lambda\sigma^2 x_{i1}^2$, (II) $\sigma_i^2 = \sigma^2$ for $i \leq N/2$ and $\sigma_i^2 = \lambda\sigma^2$ for $i > N/2$, where $\sigma^2 = 1$ and, λ is in both cases a parameter governing the degree of heteroscedasticity. The first type characterises the variance increasing with one of the regressors, while the second represents groupwise heteroscedasticity, namely the variance that differs between the first and second half of the sample. The analysed types of heteroscedasticity are inspired by the study of Ali and Giaccotto (1984). The choice of these two particular types of heteroscedasticity follows from the purpose for which each of the three heteroscedasticity tests considered was constructed. While the Breusch-Pagan test is more suitable for testing heteroscedasticity increasing with one of the regressors, the Goldfeld-Quandt test and the Harrison-McCabe test are mainly used when groupwise heteroscedasticity is suspected in the data. Therefore, we can expect tests to perform particularly well for the type of heteroscedasticity for which they were constructed and to perform poorly for other types of heteroscedasticity.

We assess the impact of outliers on the level and power of the test. The level α denotes the probability of incorrectly rejecting the null hypothesis, while the power of a test is the probability of correctly rejecting the null hypothesis. In the simulation, we consider a one-sample test of the nominal level of $\alpha = 0.05$, for a sample of $N = \{500, 1000\}$ observations. Since the asymptotic distribution is used to obtain the critical values for the Breusch-Pagan test (see Section 3.1.1) and the small-sample size properties of the classical and robust heteroscedasticity tests are not investigated in this paper, we do not consider the sample size smaller than 500 observations in the study.

Let $H_0 : \theta = \theta_0$ be the null hypothesis, and $\theta_N = \theta_0 + \frac{\Delta}{\sqrt{N}}$ a sequence of contiguous alternatives, where $\Delta > 0$ (Noether, 1955). Then, F_{θ_0} is a parametric model under the null hypothesis, and F_{θ_N} under the alternative hypothesis, respectively. G is an arbitrary contamination generating distribution and ϵ is a contamination level. Following Hampel et al. (1986) the contaminated distribution for the level is defined as

$$F^L_{\epsilon,N} = \left(1 - \frac{\epsilon}{\sqrt{N}}\right) F_{\theta_0} + \frac{\epsilon}{\sqrt{N}} G, \tag{10}$$

and

$$F^P_{\epsilon,N} = \left(1 - \frac{\epsilon}{\sqrt{N}}\right) F_{\theta_N} + \frac{\epsilon}{\sqrt{N}} G \tag{11}$$

for the power. We apply these particular types of contamination distribution to enable the shrinking neighbourhoods of the null hypothesis and the alternatives, and as a result, avoid overlapping between them (Huber & Ronchetti, 2009). In the simulation study, the general results from Equation (10) and Equation (11) are applied to the heteroscedasticity tests with the Tukey-Huber contamination model (Tukey, 1960, Huber, 1964) with a point mass contamination $G = \Delta_{(y^*,x_1^*,x_2^*)}$ at y^*, x_1^*, x_2^*. That is,

$$F^L_{\epsilon,N} = \left(1 - \frac{\epsilon}{\sqrt{N}}\right) F_{\theta_0} + \frac{\epsilon}{\sqrt{N}} \Delta_{(y^*,x_1^*,x_2^*)}, \tag{12}$$

and

$$F^P_{\epsilon,N} = \left(1 - \frac{\epsilon}{\sqrt{N}}\right) F_{\theta_N} + \frac{\epsilon}{\sqrt{N}} \Delta_{(y^*,x_1^*,x_2^*)}. \tag{13}$$

Thus, every observation in the initial clean dataset has a given probability of being an outlier, and no additional points are added to the dataset. The considered probability of being an outlier, i.e., contamination level, is $\epsilon = 0.01$. We generate the point mass contamination with two different settings, to evaluate the effect of a bad leverage point (i.e., a point characterised by a large distance in the explanatory space and a large standardised regression residual) and a vertical outlier (i.e, a point characterised by a large standardised regression residual).

For the evaluation of the level of the test, homoscedastic data is considered, produced with the data generating process based on Equation (9) and $e_i \sim \mathcal{N}(0, 1)$. Next to the non-contaminated dataset, the datasets including outliers are generated with the addition of vertical outliers (placed at $y^* = -100$) or bad leverage points (placed at

$y^* = x_1^* = x_2^* = -50$). The point mass contamination is added according to Equation (12), with a degree of contamination $\epsilon = 0.01$. In the simulation process, there are 100 runs, each of which generates 1000 test statistics and p-values. For each of the 100 runs, we calculate the percentage of tests in which the null hypothesis is rejected. This allows us to assess the level of the test and how much it differs from the prespecified level $\alpha = 0.05$. To thoroughly assess the spread of the test level, the boxplots are constructed. The specification of the heteroscedasticity tests considered is as follows. In the Breusch-Pagan test, both explanatory variables are used as potential drivers of the errors' variance in the function (2). In the Goldfeld-Quandt test, $c_N = 0$, $n_1 = n_2 = \frac{N}{2}$ and observations are assumed to be ordered according to the index i. In the Harrison-McCabe test, $n_1 = \frac{N}{2}$ and observations are also assumed to be ordered according to the index i.

For the evaluation of the power of the test, the heteroscedastic error terms are considered. The power is evaluated over the range of degrees of heteroscedasticity λ.

Thus, for a variance increasing with one of the regressors the evaluated range of variance starts with $\sigma_i^2 = 1$, homoscedastic variance, and then gradually the degree of heteroscedasticity increases as $\sigma_i^2 \in \{0.1x_{i1}^2, 0.2x_{i1}^2, \ldots, 4.9x_{i1}^2, 5x_{i1}^2\}$ for $i = 1, \ldots, N$, where N is a sample size. For groupwise heteroscedastic error terms, for half of the sample, the variance is equal to one, while for the other half, it ranges in $\{1, 1.1, \ldots, 4.9, 5\}$, once again starting from homoscedastic error terms when $\sigma_i^2 = 1$ in both groups.

Overall, for both types of heteroscedastic error terms, a wide range of the degree of heteroscedasticity is assessed. We analyse samples contaminated to the same extent, $\epsilon = 0.01$. The point mass contamination is added according to Equation (13), with the same vertical outliers and bad leverage outliers as in the evaluation of the level of the test. In the simulation, for each value of λ, 1000 replications providing test statistic and p-value are performed. Next, we calculate the percentage of tests in which the null hypothesis is rejected, which allows us to assess the power of the test. Calculated powers are plotted against the degree of heteroscedasticity λ providing the power curves for each combination of the heteroscedasticity test, the sample size, and the contamination scenario. The heteroscedasticity tests when applied to groupwise heteroscedasticity are specified in the same way as for the evaluation of the level of the test. While for the variance increasing with the regressor x_1 each test is specified in two ways. The first specification is the same as for groupwise heteroscedasticity, while the second differs as

follows. In the Breusch-Pagan test, the second specification postulates a more specific formula of the variance function (2), where $z_{i2} = x_{i1}$ and $z_{i3} = x_{i1}^2$. In the Goldfeld-Quandt test and the Harrison-McCabe test, the second specification differs in the assumed ordering of the observations, that is, the observations are ordered according to the values of x_1. The values of c_N, n_1, and n_2 remain the same. The comparison of the power curves for varying specifications of the same heteroscedasticity test verifies whether a more explicit researcher's assumption about the expected form of heteroscedasticity leads to a more powerful test, specifically in the presence of outliers. For the Breusch-Pagan test, we do not expect to observe a big difference in the test power between specifications with unspecified and specified variance functions (2), since this test should also be good enough when the form of heteroscedasticity in uncertain (Lyon & Tsai, 1996). However, the Goldfeld-Quandt test and the Harrison-McCabe test are expected to gain higher power when the observations are ordered according to x_1 as the variance increases proportionally to it.

Both parts of the simulation study, the evaluation of the level of the test under contamination and the power of the test under contamination, contribute to answering the main research question of this paper and demonstrating how robust the classical heteroscedasticity tests are. To acknowledge the robustness of the classical heteroscedasticity test, the performance of both size (i.e., the robustness of validity) and power (i.e., the robustness of efficiency) should be satisfactory (Heritier & Ronchetti, 1994). In this research, for a well-behaved test, the size should be close to 0.05 (the nominal level of significance α), while the power should tend to 1. Following Pearson and Please (1975), the test obtaining the actual level between 0.03 and 0.07 can be considered acceptable and robust. This range is less stringent than the approach allowing for sampling errors and considering the test to be robust if the actual level does not exceed the nominal level by two standard errors (for example, applied in the study of the heteroscedasticity tests by Ali and Giaccotto (1984)), which in this simulation design, $\alpha = 0.05$ and the number of replications 1000, means that the actual level should be no higher than 0.064 and no lower than 0.036.

3.3.1 Sensitivity analysis

The sensitivity analysis of the p-value obtained with the classical heteroscedasticity tests provides an illustrative but more superficial overview of the test stability than the detailed evaluation of the level and the power of these tests with simulation study as described in Section 3.3. The data is generated according to the data-generating process based on Equation (9) and the sample contains 500 observations. Three scenarios of error terms variance are investigated: homoscedastic, that is $\sigma_i^2 = \sigma^2 = 1$, groupwise heteroscedastic, that is $\sigma_i^2 = \sigma^2 = 1$ for $i \leq 250$ and $\sigma_i^2 = \sigma^2 = 4$ for $i > 250$, and heteroscedastic with variance increasing proportionally to x_1, that is $\sigma_i^2 = x_{i1}^2$, for $i = 1, \ldots, N$, where N is a sample size. The same classical heteroscedasticity tests are evaluated for each scenario - the Breusch-Pagan test (1979), the Goldfeld-Quandt test (1965) and the Harrison-McCabe test (1979). For homoscedastic and groupwise heteroscedastic error terms, the variance function of the Breusch-Pagan test contains both regressors. In the Goldfeld-Quandt test and the Harrison-McCabe test variances are ordered according to the observation index i, and while in the former $c_N = 0$, $n_1 = n_2 = \frac{N}{2}$, in the latter $n_1 = \frac{N}{2}$. For heteroscedastic error terms when variance increases proportionally to x_{i1}^2 two specifications of each test are considered (as in the power evaluation, see Section 3.3). For the Breusch-Pagan test, next to the default specification of the variance function, the function where $z_{i2} = x_{i1}$ and $z_{i3} = x_{i1}^2$ is analysed. For the Goldfeld-Quandt test and the Harrison-McCabe test, the alternative ordering of the observations is considered, according to the values of x_1. The influence of vertical outliers and bad leverage outliers is analysed by moving one single observation. For the contamination from vertical outliers, only the value of y is switched in the range $\{-100, -99, \ldots, 99, 100\}$, while x_1 and x_2 are held constant. For the case of bad leverage outliers, y, x_1, x_2 are moved in the range $\{-100, -99, \ldots, 99, 100\}$. The p-value of the robust test should be stable over the whole assessed range.

3.4 Simulation results

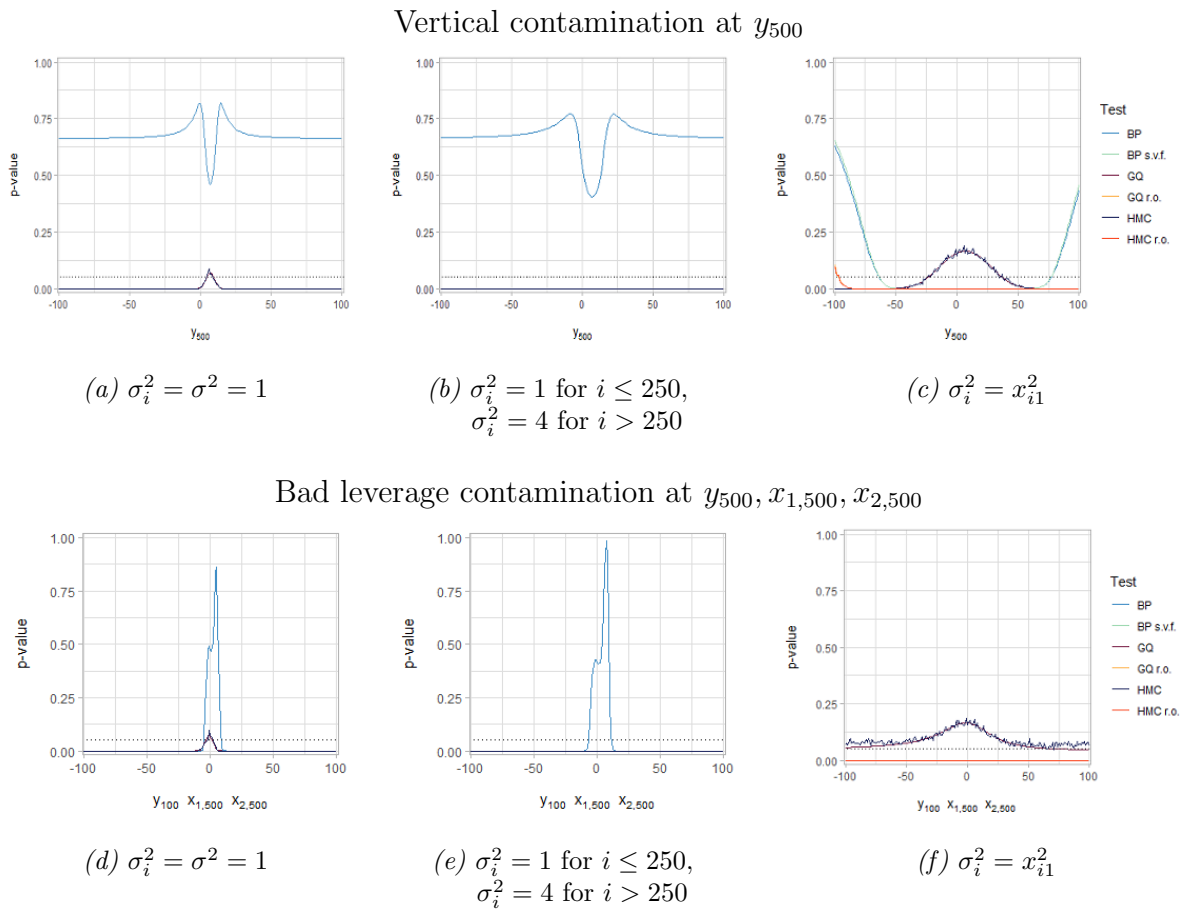
In this section, we present the results of the simulation study evaluating the sensitivity, level and power of three classical heteroscedasticity tests: the Breusch-Pagan test, the Goldfeld-Quandt test and the Harrison-McCabe test.

3.4.1 Sensitivity analysis

Figure 3.2 presents the results of the sensitivity analysis of p-values of the Breusch-Pagan test, the Goldfeld-Quandt test and the Harrison-McCabe test in the presence of contamination. Figure 3.2a and Figure 3.2d show that the level of all three classical heteroscedasticity tests is not stable for both vertical and bad leverage contamination when the homoscedastic error terms are present. Only for the Breusch-Pagan test in the case of vertical contamination, the p-values do not drop below the nominal level of the test, thus correctly indicating that there is no reason to reject the null hypothesis. The Goldfeld-Quandt and the Harrison-McCabe tests for the majority of the values of outlying observation considered achieve a p-value below the nominal level, providing false evidence to reject the null.

However, those two tests behave well when groupwise heteroscedasticity is analysed. Figure 3.2b and Figure 3.2e, demonstrate that over the whole assessed range of outlying observation, both tests achieve p-value close to zero and correctly detect heteroscedasticity of error terms. For groupwise heteroscedasticity with vertical contamination, the level of the Breusch-Pagan test remains considerably above the nominal level of the test irrespective of the placement of outlying observation, while for the contamination with bad leverage, the level is unstable and switches from p-values indicating homoscedastic error terms to p-values suggesting the presence of heteroscedasticity.

For variance of error terms increasing with x_{i1}^2, the p-values of the Breusch-Pagan test with both specifications of the variance function behave similarly, see Figure 3.2c and Figure 3.2f. In the case of bad leverage contamination, the p-values are not influenced by the outlying observation and lay below 0.05, while for vertical outliers the p-values drop from high values above 0.6 to values below 0.05 when y_{500} is between -63 and 76. The Goldfeld-Quandt test and the Harrison-McCabe test with the same ordering of the observations achieve similar p-values. For bad leverage contamination, the specifications with ordering by x_1 are stable and have p-value close to zero. While the same tests, but with observations ordered according to the index i, characterise higher p-values. When y_{500} lays between -24 and 37, p-value ranges from 0.05 to 0.19. For more extreme values of a vertical outlier, the level stays approximately constant between $0.05 - 0.06$. Thus, both specifications with the index ordering incorrectly suggest no evidence to reject the null hypothesis about homoscedastic error terms. In the case of vertical outlier contamination,



(a) $\sigma_i^2 = \sigma^2 = 1$ *(b)* $\sigma_i^2 = 1$ for $i \leq 250$, $\sigma_i^2 = 4$ for $i > 250$ *(c)* $\sigma_i^2 = x_{i1}^2$

(d) $\sigma_i^2 = \sigma^2 = 1$ *(e)* $\sigma_i^2 = 1$ for $i \leq 250$, $\sigma_i^2 = 4$ for $i > 250$ *(f)* $\sigma_i^2 = x_{i1}^2$

Figure 3.2. Sensitivity analysis of the p-value of the Breusch-Pagan test, the Goldfeld-Quandt test and the Harrison-McCabe test performed on the residuals from the linear regression with homoscedastic error terms, (a) and (d), and heteroscedastic error terms, (b), (c), (e), and (f), for sample size $N = 500$ in the presence of either vertical or bad leverage outlier. The value of y_{500} (vertical outlier) or the whole observation $i = 500$ (bad leverage outlier) is moved in range $\{-100, -99, \ldots, 99, 100\}$. BP denotes the Breusch-Pagan test with unspecified variance function (2), BP s.v.f. denotes the Breusch-Pagan test with specified variance function (2), where $z_{i2} = x_{i1}$ and $z_{i3} = x_{i1}^2$. GQ denotes the Goldfeld-Quandt test with index ordering, GQ r.o. denotes the Goldfeld-Quandt test with ordering by regressor x_1, HM denotes the Harrison-McCabe test with index ordering and HM r.o. denotes the Harrison-McCabe test with ordering by regressor x_1. The level of the test $\alpha = 0.05$ is shown with the black dotted line.

the Goldfeld-Quandt test and the Harrison-McCabe test with the index ordering are characterised by p-value curve shapes similar to those obtained by the same tests in a scenario with bad leverage contamination. However, for the most extreme placements of outlying vertical outliers, both tests achieve a p-value equal to zero, considerably lower than the p-values in the scenario with bad leverage outliers for the corresponding cases.

This illustrative example of sensitivity analysis of p-value shows that none of the classical heteroscedasticity tests investigated is stable and not influenced by the outlying observation in all scenarios considered, thus pointing out the need to develop the robust

heteroscedasticity test.

3.4.2 Evaluation of the level of the test

The level of the classical heteroscedasticity tests is evaluated for homoscedastic error terms, for sample sizes of 500 and 1000 observations under three contamination scenarios.

Figure 3.3 shows the boxplots for a sample size of 500 observations. In the case of a non-contaminated sample, all three tests remain size-correct.

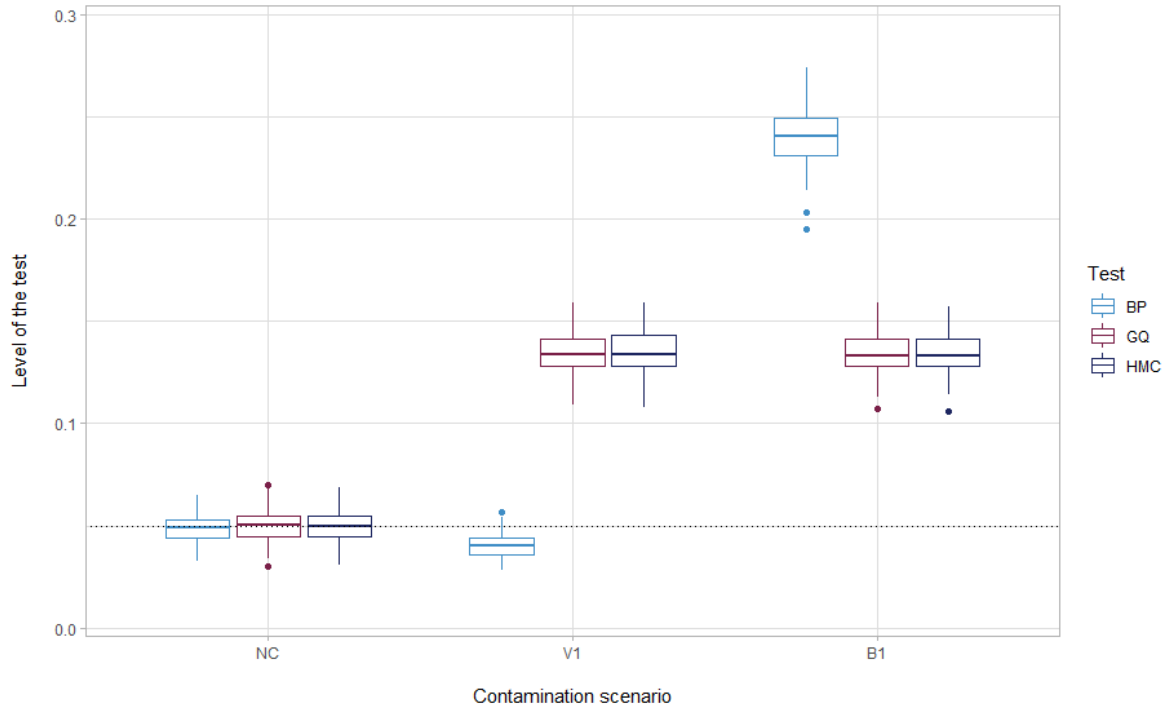


Figure 3.3. The level of the Breusch-Pagan test (BP), the Goldfeld-Quandt test (GQ) and the Harrison-McCabe test (HMC) for sample size $N = 500$ under three contamination scenarios. NC denotes the scenario without contamination, B denotes the model contaminated with bad leverage points, and V with vertical outliers. $\epsilon = 0.01$ is denoted with 1. The level of the test $\alpha = 0.05$ is shown with the black dotted line.

When bad leverage points are present, the Breusch-Pagan test is the worst-performing out of the three tests considered. The median of its actual level is 0.24, and therefore in 24 out of 100 cases it incorrectly rejects the null hypothesis that error terms are homoscedastic. Although the Breusch-Pagan test is the least robust to bad leverage points, it performs well for vertical outliers. The actual level of the test stays below the nominal level of 0.05 with the median value of the actual level of 0.04. The Goldfeld-Quandt test and the Harrison-McCabe test are not robust to any type of investigated contamination

(median level is approximately 0.13) and their boxplots achieve similar spread. None of the tests considered is robust to bad leverage points.

Figure 3.4 demonstrates the boxplots for a sample size of 1000 observations. In the scenario without any contamination, all three tests remain size-correct. Similarly, as in the case of $N = 500$, the Breusch-Pagan test is the worst-performing test for bad leverage points, with a considerably inflated median level of 0.306. While for vertical outliers, the actual test level lies below the nominal test level, with a median value of 0.039, thus the test remains robust to this type of contamination. The Harrison-McCabe test and the Goldfeld-Quandt test are not robust to any sample contamination, their median level is approximately 0.16. None of the tests considered is robust to bad leverage points.

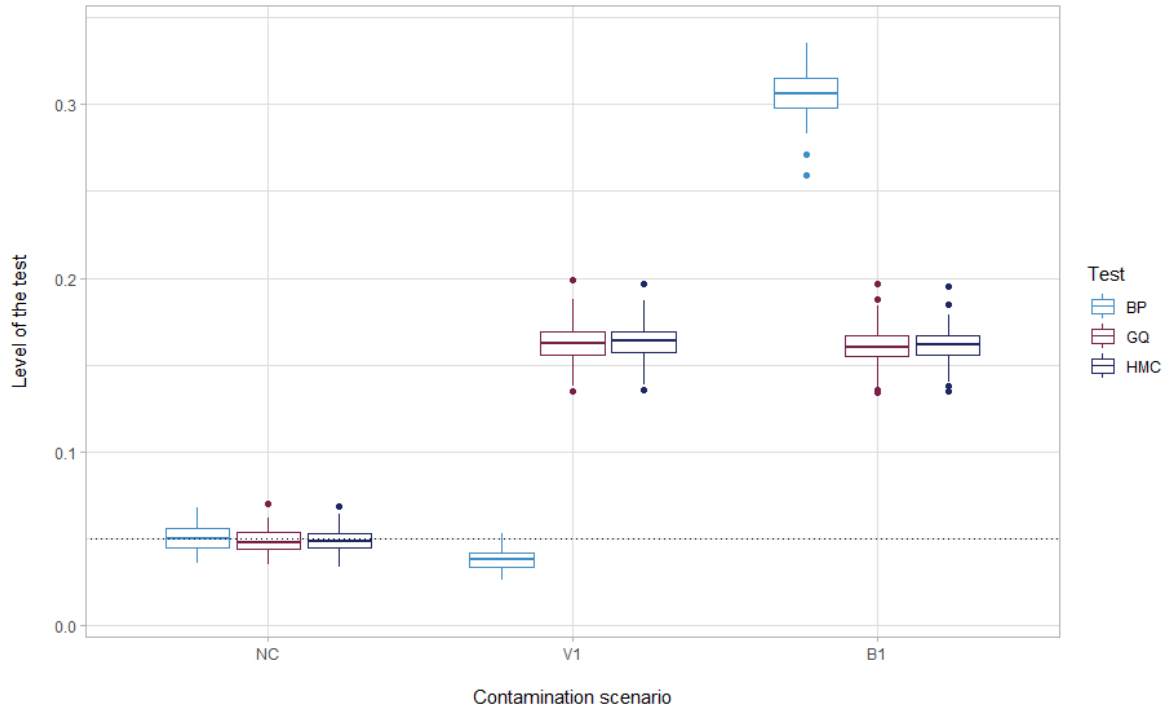


Figure 3.4. The level of the Breusch-Pagan test (BP), the Goldfeld-Quandt test (GQ) and the Harrison-McCabe test (HMC) for sample size $N = 1000$ under three contamination scenarios. NC denotes the scenario without contamination, B denotes the model contaminated with bad leverage points, and V with vertical outliers. $\epsilon = 0.01$ is denoted with 1. The level of the test $\alpha = 0.05$ is shown with the black dotted line.

Overall, when the robustness of the classical heteroscedasticity tests is evaluated in terms of the level of the test, there is no single test out of the three considered that preserves the nominal level of the test and performs well in all contamination scenarios. The Breusch-Pagan test appears to be robust against vertical outliers in both sample sizes considered. The Harrison-McCabe test and the Goldfeld-Quandt test are oversized and

nonrobust regardless of the considered sample size and contamination scenario. These results point to the need to develop a heteroscedasticity test that is robust to contamination regardless of the type of outliers present in the data.

3.4.3 Evaluation of the power of the test

Evaluation of the power of the tests is done using the power curves. Two types of heteroscedastic error terms, groupwise heteroscedasticity and the heteroscedastic error terms increasing with the value of one of the regressors, are considered for sample sizes $N = \{500, 1000\}$.

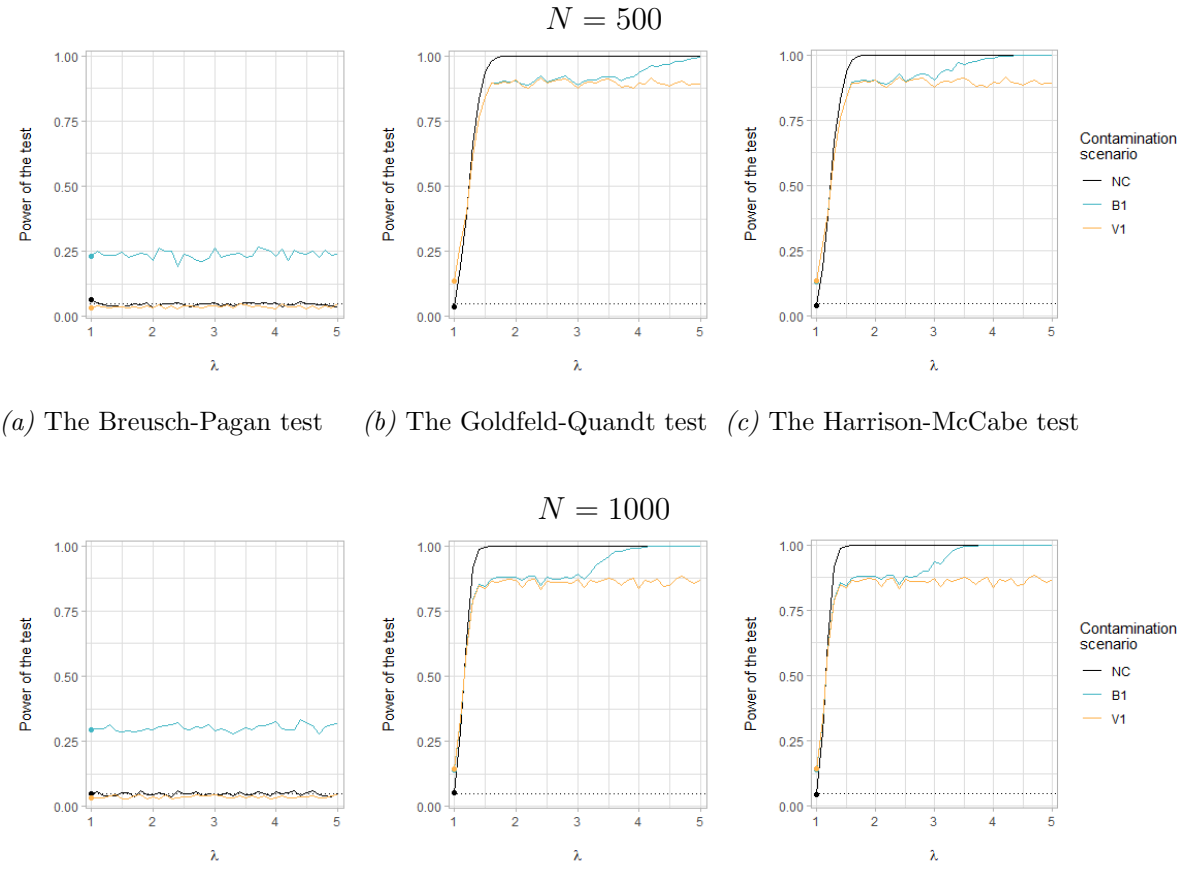
Groupwise heteroscedestacity

First, we analyse the groupwise heteroscedasticity, that is the variance of error terms varying between two groups of observations. The evaluation starts with equal variances in both groups (homoscedastic error terms, $\lambda = 1$), thus the curve should start at approximately 0.05. The higher λ, the more pronounced the difference between variances in both groups is. Figure 3.5 presents the power curves obtained for three heteroscedasticity tests when groupwise heteroscedasticity is considered.

For the sample size $N = 500$, the Breusch-Pagan test, see Figure 3.5a, is not able to detect heteroscedastic error terms for both vertical and bad leverage outliers, but also in the non-contaminated dataset. The power of the test in any scenario does not exceed 0.3. For the Goldfeld-Quandt test, see Figure 3.5b, when vertical outliers are present the power stabilises around a variance of 1.5 times greater in the second half of the sample (the mean power is approximately 0.9). For bad leverage points, the power curve increases over the whole assessed range of λ and tends to 1. The power curve of the Harrison-McCabe test for vertical outliers, see Figure 3.5c, characterises an almost identical shape to the corresponding power curve of the Goldfeld-Quandt test. Both tests have close mean power after flattening out around $\lambda = 1.5$. For bad leverage points, the power curve of the Harrison-McCabe test tends to 1 and stays at 1 for $\lambda > 4$. The power of 1 is achieved for a smaller value of λ than in the case of the Goldfeld-Quandt test.

For a larger sample size $N = 1000$, the power curves of individual tests do not differ considerably from those for $N = 500$. The Breusch-Pagan test does not achieve power higher than 0.35 in any of the contamination scenarios analysed, see Figure 3.5d. In the

scenario with vertical outliers, the power curves of the Harrison-McCabe test and the Goldfeld-Quandt test flatten out around a variance of 1.5 times greater in the second half of the sample and stay at the mean power of 0.86, which is a slightly lower value than in the case of $N = 500$, see Figures 3.5f and 3.5e. For bad leverage points, the power curves of the Harrison-McCabe test and the Goldfeld-Quandt test tend to 1, the former achieves maximum power when $\lambda > 3.7$, while the latter when $\lambda > 4.1$. For both cases, the power of 1 is achieved for a smaller value of λ than in the respective scenario when $N = 500$.



(a) The Breusch-Pagan test *(b)* The Goldfeld-Quandt test *(c)* The Harrison-McCabe test

(d) The Breusch-Pagan test *(e)* The Goldfeld-Quandt test *(f)* The Harrison-McCabe test

Figure 3.5. Power curves for the Breusch-Pagan, the Goldfeld-Quandt and the Harrison-McCabe tests, performed on the residuals from the linear regression (see Equation (9)) with heteroscedastic error terms, $\sigma_i^2 = \sigma^2$ for $i \leq N/2$ and $\sigma_i^2 = \lambda\sigma^2$ for $i > N/2$, where $\sigma^2 = 1$ and $\lambda \in \{1, 1.1, \ldots, 4.9, 5\}$, for the sample sizes of $N \in \{500, 1000\}$ under three contamination scenarios. NC denotes the scenario without contamination, B denotes the model contaminated with bad leverage points, and V with vertical outliers. $\epsilon = 0.01$ is denoted with 1. The level of the test $\alpha = 0.05$ is shown with the black dotted line.

Generally, in terms of the test power when groupwise heteroscedasticity of error terms is present, the Harrison-McCabe test appears to be the most robust to bad leverage points. When it comes to the robustness against the vertical outliers, both the Goldfeld-Quandt test and the Harrison-McCabe test perform similarly. The achieved powers stay constant

once λ exceeds a certain value, but the power curves do not tend to 1 as λ increases. The Breusch-Pagan test performs poorly for all considered contamination scenarios. It has no power to reject the null when vertical outliers are present, even for larger deviations from the null hypothesis. Both tests developed specifically for testing groupwise heteroscedasticity confirmed their better performance in this simulation exercise, however, none of them performs well enough in all cases to be considered robust.

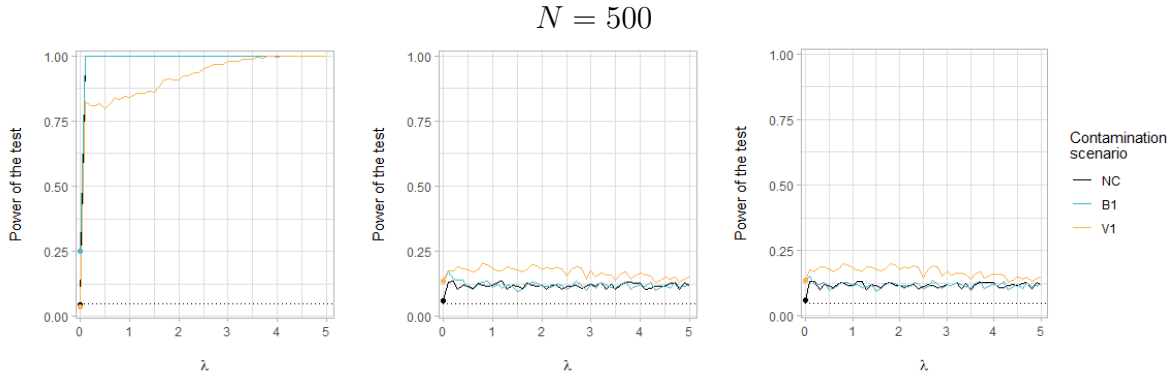
Heteroscedestacity increasing with one of the regressors

The evaluation of the power curves for error terms characterised by heteroscedasticity increasing with one of the regressors is performed with two different specifications of each test, see details in Section 3.3. The evaluation starts with $\sigma_i^2 = 1$ and then heteroscedastic variance, λx_{i1}^2, starts to increase with the degree of heteroscedasticity λ.

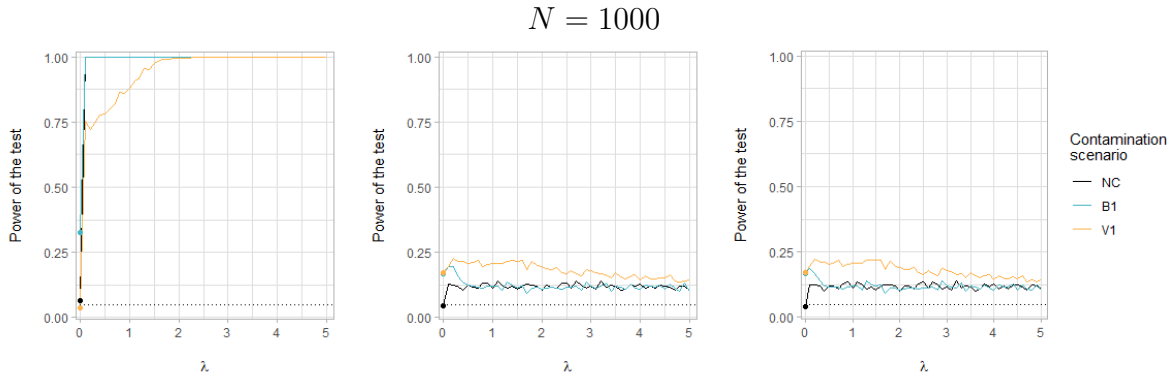
Comparison of the power curves for both sample sizes for the Goldfeld-Quandt test and the Harrison-McCabe test with different observation orderings shows that in all cases considered, the proper choice of observations ordering is crucial to obtain the high power of the test, that is the ordering according to x_1 - the regressor with which the variance of error terms increases. For the non-contaminated datasets with index ordering, both tests have an average power of approximately 0.12, with a maximum of approximately 0.14, see Figures 3.6b, 3.6c, 3.6e and 3.6f. Therefore, in the following section, the interpretation of the power curves obtained with the Goldfeld-Quandt test and the Harrison-McCabe test focuses only on the specifications with the ordering by x_1.

For the sample size $N = 500$, the power curves of the Breusch-Pagan test, for both unspecified, see Figure 3.6a, and specified function, see Figure 3.7a, stay constant at 1 for all values of $\lambda > 0$ both in the absence of contamination and in the case of bad leverage contamination. For vertical outliers, the power curves tend to 1 and stabilize around 1 for $\lambda > 3$. The differences in the individual power values between test specifications with the specified and unspecified version of the variance function are minor, but in the majority of cases, the specified version achieves slightly higher power. For the Goldfeld-Quandt test and scenario with vertical outliers, the power curve tends to 1, see Figure 3.7e. However, the value of λ at which it stabilizes at 1 is lower than for the analogous scenario with the Breusch-Pagan test. It achieves the power of one when $\lambda > 1.5$. The behaviour of the power curves of the Harrison-McCabe test for the scenario with vertical

outliers is almost identical, see Figure 3.7f. For contamination with bad leverage points and no contamination scenario, the power curves of both the Goldfeld-Quandt test and the Harrison-McCabe test are constant at 1 once error terms become heteroscedastic.



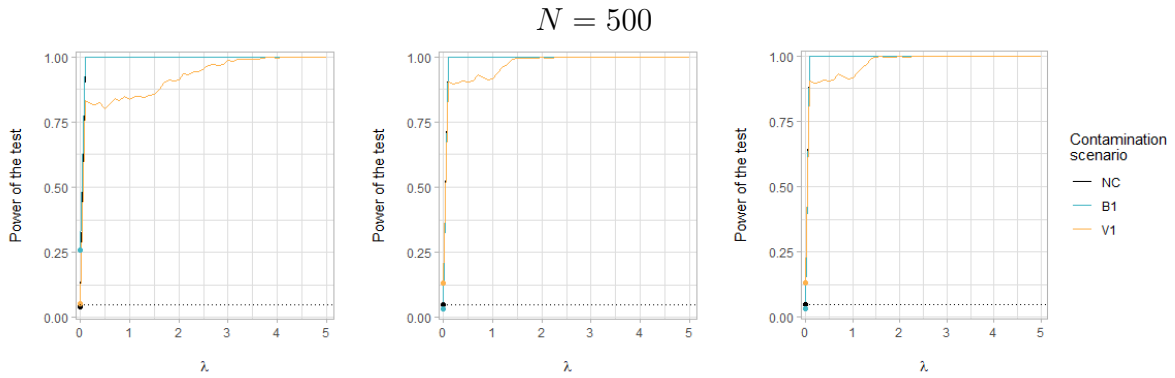
(a) The Breusch-Pagan test *(b)* The Goldfeld-Quandt test *(c)* The Harrison-McCabe test



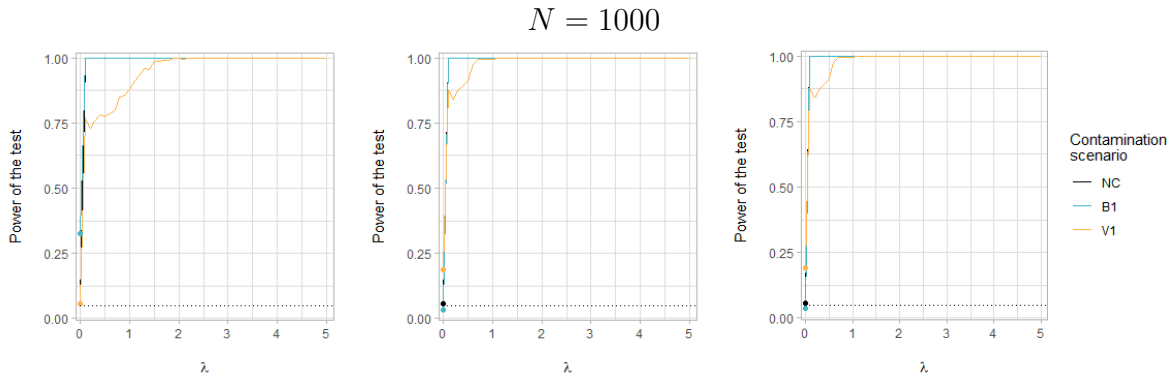
(d) The Breusch-Pagan test *(e)* The Goldfeld-Quandt test *(f)* The Harrison-McCabe test

Figure 3.6. Power curves for the Breusch-Pagan test with unspecified variance function, the Goldfeld-Quandt test with index ordering and the Harrison-McCabe test with index ordering, performed on the residuals from the linear regression (see Equation (9)) with heteroscedastic error terms, $\sigma_i^2 = \lambda\sigma^2 x_{i1}^2$, where $\sigma^2 = 1$ and $\lambda \in \{0.1, 0.2, \ldots, 4.9, 5\}$, for the sample sizes of $N \in \{500, 1000\}$ under three contamination scenarios. NC denotes the scenario without contamination, B denotes the model contaminated with bad leverage points, and V with vertical outliers. $\epsilon = 0.01$ is denoted with 1. The level of the test $\alpha = 0.05$ is shown with the black dotted line.

For the larger sample size, $N = 1000$, the power curves of individual tests do not differ considerably from those obtained when $N = 500$. For scenarios without any contamination and with bad leverage outliers, both specifications of the Breusch-Pagan test, see Figure 3.6d and 3.7d, the Harrison-McCabe test, see Figure 3.7f, and the Goldfeld-Quandt test, see Figure 3.7e, achieve a maximum power of 1 once error terms become heteroscedastic. In the case of contamination with vertical outliers, all three tests tend



(a) The Breusch-Pagan test *(b)* The Goldfeld-Quandt test *(c)* The Harrison-McCabe test



(d) The Breusch-Pagan test *(e)* The Goldfeld-Quandt test *(f)* The Harrison-McCabe test

Figure 3.7. Power curves for the Breusch-Pagan test with specified variance function, the Goldfeld-Quandt test with ordering by x_1 and the Harrison-McCabe test with ordering by x_1, performed on the residuals from the linear regression (see Equation (9)) with heteroscedastic error terms, $\sigma_i^2 = \lambda\sigma^2 x_{i1}^2$, where $\sigma^2 = 1$ and $\lambda \in \{0.1, 0.2, \ldots, 4.9, 5\}$, for the sample sizes of $N \in \{500, 1000\}$ under three contamination scenarios. NC denotes the scenario without contamination, B denotes the model contaminated with bad leverage points, and V with vertical outliers. $\epsilon = 0.01$ is denoted with 1. The level of the test $\alpha = 0.05$ is shown with the black dotted line.

to the power of 1 and eventually achieve it, however, it happens for different values of λ. For both the Harrison-McCabe and the Goldfeld-Quandt test, it is $\lambda = 0.7$, while for the Breusch-Pagan test the maximum power is achieved when $\lambda > 1.8$.

The analysis of the power curves, when the variance of error terms increasing with one of the regressors is present, shows a lack of notable differences between the Goldfeld-Quandt test and the Harrison-McCabe test when observations are ordered by x_1 (the specifications of the same tests but with index ordering result in unsatisfactorily low power below 0.15). Likewise, the differences between the power curves of the Breusch-Pagan test with specified and unspecified variance functions are rather minor in all cases. For both sample sizes $N = 500$ and $N = 1000$, all three tests are robust against bad

leverage points and as could be expected all tests become more powerful when the sample size increases. When vertical outliers are present, all three tests are robust if the deviation from the null hypothesis is large enough (λ exceeds a certain value), but no test is robust for all values of λ.

The overall results of the simulation study for the evaluation of the level and the power of the classical heteroscedasticity tests indicate that none of the three tests considered, i.e., the Breusch-Pagan test, the Goldfeld-Quandt test and the Harrison-McCabe test, preserves both the robustness of validity and the robustness of efficiency when small contamination is present in the data.

4 Robust heteroscedasticity score test

In this section, we propose a robust alternative to the classical Breusch-Pagan test, which we expect to be more robust compared to the classical heteroscedasticity tests. We require a proposed test statistic to have a bounded influence function, and we show why this boundedness holds. Next, we investigate the robustness of the proposed test with the simulation study of the level and the power of the test.

4.1 Theoretical framework

The construction of the robust heteroscedasticity test starts from a classical heteroscedasticity Breusch-Pagan test (see Section 3.1.1), which belongs to the class of score tests. We consider the same linear regression model as in the Breusch-Pagan test, see Equation (1), with the continuous function of the error terms variance, see Equation (2). The null hypothesis of the classical test, $H_0 : \alpha_2 = \ldots = \alpha_m = 0$, implies homoscedasticity, that is $\sigma_i^2 = h(\alpha_1) = \sigma^2$.

The robust heteroscedasticity score test is derived following the approach of Heritier and Ronchetti (1994), that is, constructing the robust test statistic based on the M-estimator (Huber, 1973). To construct a robust test statistic, we need to define a parametric model of interest. We consider the function of the error terms variance from Equation (2) as a parametric model F_θ, where θ, defined as $\{\alpha_1^\top, \alpha_j^\top\}^\top$ and $j = 2, \ldots, m$, lies in $\mathbf{\Omega}$ an open convex subset of $\mathbb{R}^p$, and a sample $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ of n iid random vectors. Following the notation from Heritier and Ronchetti (1994), $\mathbf{a} = \{a_{(1)}^\top, a_{(2)}^\top\}^\top$ denotes the partition of a vector $\mathbf{a}$ into $p - q$ and q components and $\mathbf{A}_{(ij)}, i, j = 1, 2$, denotes the

corresponding partition of $p \times p$ matrices. Using this notation the null hypothesis states $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_{0(2)} = 0$, $\boldsymbol{\theta}_{0(1)}$ unspecified, implying homoscedasticity. In this case, $\boldsymbol{\theta}_{0(2)}$ corresponds to $\alpha_j^\top$ and $\boldsymbol{\theta}_{0(1)}$ to $\alpha_1^\top$. The alternative hypothesis reads $H_1 : \boldsymbol{\theta}_{0(2)} \neq 0$, $\boldsymbol{\theta}_{0(1)}$ unspecified and implies heteroscedasticity.

In the robust heteroscedasticity score test, M-estimator $\hat{\theta} := \{\hat{\alpha}_1^\top, \hat{\alpha}_j^\top\}^\top$ is defined by $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\Psi}(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_n) = \mathbf{0}$, where $\boldsymbol{\Psi}$ denotes the score function. The test statistic is built upon the restricted version of the M-estimator under the null hypothesis, that is the M-estimator, $\hat{\boldsymbol{\theta}}_{res}$, which solves the equation

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}\left(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_{res}\right)_{(1)} = \mathbf{0} \text{ with } \left(\hat{\boldsymbol{\theta}}_{res}\right)_{(2)} = \mathbf{0}. \tag{14}$$

For the conditions for the existence of M-estimators, see Heritier and Ronchetti (1994, p. 902). The test statistic itself is computed as follows:

$$R_{het} := \mathbf{Z}_n^\top \mathbf{C}^{-1} \mathbf{Z}_n, \tag{15}$$

where $\mathbf{Z}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Psi}\left(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_{res}\right)_{(2)}$, and the matrix $\mathbf{C} = \mathbf{C}(\boldsymbol{\Psi}, F_\theta)$ is the asymptotic covariance matrix of $\mathbf{Z}_n$. Matrix $\mathbf{C}$ is constructed as

$$\mathbf{C}(\boldsymbol{\Psi}, F_\theta) = \mathbf{M}(\boldsymbol{\Psi}, F_\theta)_{(22.1)} \mathbf{V}(\boldsymbol{\Psi}, F_\theta)_{(22)} \mathbf{M}(\boldsymbol{\Psi}, F_\theta)_{(22.1)}^\top, \tag{16}$$

where $\mathbf{M}(\boldsymbol{\Psi}, F_\theta) = -\int (\partial \boldsymbol{\Psi}/\partial\theta)(\mathbf{z}, \theta) dF_\theta(\mathbf{z})$ and $\mathbf{V}(\boldsymbol{\Psi}, F_\theta) = \mathbf{M}(\boldsymbol{\Psi}, F_\theta)^{-1} \mathbf{Q}(\boldsymbol{\Psi}, F_\theta) \mathbf{M}(\boldsymbol{\Psi}, F_\theta)^{-1}$, where $\mathbf{Q}(\boldsymbol{\Psi}, F_\theta) = \int \boldsymbol{\Psi}(\mathbf{z}, \theta) \boldsymbol{\Psi}(\mathbf{z}, \theta)^\top dF_\theta(\mathbf{z})$. The null asymptotic distribution of the statistic nR_{het} is the χ^2_{m-1} distribution (Heritier & Ronchetti, 1994).

4.2 Influence function

We evaluate the robustness of the test statistic (15) using the influence function introduced by Hampel (1974), see Equation (6). In general, robust methods necessitate the bounded influence of outliers. The robustness of any estimator requires the bounded influence function so that the estimate is stable under local perturbations. Otherwise, if the influence function is unbounded, the possible bias in the neighbourhood of the considered parametric model F_θ can be infinite (Hampel et al., 1986). Using the influence function, one can also derive the first order approximations of the impact of contamination on the

size and power of the test. Heritier and Ronchetti (1994) show that bounding the self-standardised influence function of the test statistic ensures the stability of and the limit in the bias in the level and the power of the test with respect to small deviation from the assumed model. The robustness of validity and the robustness of efficiency are preserved. Heritier and Ronchetti (1994) also note that the robustness of the test statistic is inherited from the robust estimator used. Therefore, the heteroscedasticity score test constructed upon the M-estimator is (locally) robust if the influence function of the M-estimator is bounded. The influence function of the M-estimator is computed as

$$\mathrm{IF}\left(\mathbf{z}, \boldsymbol{\Psi}, F_\theta\right) = \mathbf{M}\left(\boldsymbol{\Psi}, F_\theta\right)^{-1} \boldsymbol{\Psi}\left(\mathbf{z}, \theta\right), \tag{17}$$

and we see that the influence function is bounded if the function $\boldsymbol{\Psi}$ is bounded. The boundedness is achieved with, for example, the Mallows type score function (Mallows, 1975).

4.3 Construction

The Mallows type score function (Mallows, 1975) ensures that the test statistic (15) is robust to both bad leverage and vertical outliers. The function is defined as

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\hat{\boldsymbol{\theta}})}{\hat{\sigma}}\right) \omega(\mathbf{x}_i)\mathbf{x}_i = 0, \tag{18}$$

where $r_i(\hat{\boldsymbol{\theta}})$ are the residuals from the restricted model (14), ψ is a downweighting function, ω is a weight function, and $\hat{\sigma}$ is the estimated residual variance.

A downweighting function, ψ, applied to the standardised residuals, for simplicity denoted as r_i, is chosen from Tukey's biweight function or the Huber function. Tukey's biweight function is defined as

$$\psi(r_i; c) = \begin{cases} r_i\left(1 - \left(\frac{r_i}{c}\right)^2\right)^2, & \text{if } |r_i| \leq c, \\ 0, & \text{if } |r_i| > c. \end{cases} \tag{19}$$

The Huber function is defined as

$$\psi(r_i; c) = \begin{cases} r_i, & \text{if } |r_i| \leq c, \\ c\ \mathrm{sign}(r_i), & \text{if } |r_i| > c. \end{cases} \tag{20}$$

Variable c is a tuning constant determining the asymptotic efficiency. For the Huber function $c = 1.345$ and for Tukey's biweight function $c = 4.685$ ensure 95% asymptotic efficiency compared to the Least Squares estimator at a normal distribution of the error terms (Maronna et al., 2019).

A weight function, ω, to downweight outliers in the covariate space is based either on the hat matrix, where $\omega(\mathbf{x}_i) = \sqrt{1 - H_{ii}}$ and H_{ii} is the i'th diagonal element of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, or the robust Mahalanobis distance with the Minimum Covariance Determinant (MCD) (Rousseeuw, 1985). A weight function using the robust Mahalanobis distance, $d(\mathbf{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})}$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated robustly using the MCD, is defined as

$$\omega(\mathbf{x}_i) = \begin{cases} 1, & \text{if } d(\mathbf{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq \tilde{c}, \\ \tilde{c}/d(\mathbf{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}), & \text{if } d(\mathbf{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) > \tilde{c}, \end{cases} \tag{21}$$

where $\tilde{c}$ is the 0.95-quantile of the χ^2 distribution with the number of degrees of freedom equal to the dimension of $\mathbf{x}_i$. Both weight functions guarantee bounded influence function, however, weights defined through the robust Mahalanobis distance have higher breakdown properties than weights based on the hat matrix (Cantoni & Ronchetti, 2001). In real data applications, the implementation with the covariates weights based on the robust Mahalanobis distance may encounter computational issues. The issue may occur especially with datasets including several binary predictors in which appear the rows of 0's and 1's, thus causing the singular covariance matrix and precluding the execution of the MCD algorithm. In such cases, we recommend using the weights based on the hat matrix.

In the robust test, we can use two specifications of the error terms variance function, see Equation (2). In the first specification, the independent variables in the function (2) include all explanatory variables from Equation (1), $x_{i1}, \ldots, x_{ik}$, while in the second case, they include also the function of the explanatory variables, $x_{i1}^2, \ldots, x_{ik}^2$. The choice of function specification impacts the number of covariates that are downweighted with a weight function ω. When the hat matrix is applied, all explanatory variables from the variance function are included, while with the robust Mahalanobis distance with the MCD only $x_{i1}, \ldots, x_{ik}$. The variables $x_{i1}^2, \ldots, x_{ik}^2$ are not considered when computing the MCD, we include only unique variables.

The application of such defined test statistic requires the computation of the regression residuals to be supplied in the equation of the error terms variance. For this purpose, the Mallows type M-estimator is first applied to the linear regression model, Equation (1), and then the estimated residuals from the robust regression are supplied as a dependent variable to the error terms variance function. The estimated residual variance which is used to standardise the squared residuals in Equation (2) is also estimated with a robust scale estimator - the median absolute deviation (MAD). The median absolute deviation, defined as the median of the absolute deviations from the median, has a breakdown point of 50% and its influence function is bounded (Huber & Ronchetti, 2009), which makes it a good estimator to replace the nonrobust sample variance.

In practice, the algorithm of the robust heteroscedasticity test contains the following steps:

1. Estimate Equation (1) with Mallows type M-estimator and save the residuals.

2. Estimate Equation (2) with Mallows type M-estimator, where the dependent variable is constructed as in the classical Breusch-Pagan test, that is $\frac{\hat{e}_i^2}{\hat{\sigma}^2} - 1$, where $\hat{e}_i^2$ are squared estimated residuals from Equation (1), but instead of the estimated residual variance computed as $\hat{\sigma}^2 = \frac{1}{N}\sum \hat{e}_i^2$, we use the robust estimate, that is the median absolute deviation of residuals from Equation (1).

3. Compute R_{het} and compare obtained test statistic with the critical value from χ^2_{m-1} distribution.

The first step of the proposed robust test, that is the additional estimation with Mallows type M-estimator, is not included in the framework of the robust score test of Heritier and Ronchetti (1994), thus we perform the numerical assessment to check whether the asymptotic distribution χ^2_{m-1} is preserved. The details of this assessment follow in Section 4.4.

4.4 Simulation design

In this section, we propose a simulation design to verify the robustness properties of the proposed heteroscedasticity score test. The considered data-generating process follows the linear model with two regressors and intercept, as in the simulation study for the classical heteroscedasticity tests, see Equation (9). The evaluation of the level and the

power of the test is performed with different specifications of downweighting function ψ - Tukey's biweight or the Huber function, weight function ω - the hat matrix or the robust Mahalanobis distance with the MCD, and explanatory variables in the variance function - either x_{i1}, x_{i2} or $x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2$. Consequently, we assess eight different specifications of the robust test. Next to the robust test, the classical Breusch-Pagan is also included in the evaluation (x_{i1}, x_{i2} in the variance function and the test statistic with applied Koenker's correction (1981)).

The evaluation of the test level is performed with the homoscedastic data, $\sigma_i^2 = \sigma^2 = 1$, for sample sizes $N \in \{500, 1000\}$. Since we do not focus on the small-sample properties of the test, we do not consider a sample size smaller than $N = 500$. Three contamination scenarios are considered: a sample without any contamination and two datasets including outliers, either vertical outliers (placed at $y^* = -100$) or bad leverage points (placed at $y^* = x_1^* = x_2^* = -50$). The point mass contamination is added according to Equation (12) with a degree of contamination $\epsilon = 0.01$. We consider only a small level of contamination following the argument of Heritier and Ronchetti (1994) who underline that such constructed test statistic (15) ensures the stable level and power of the test in the presence of small deviations from the assumed model, but does not guarantee it in the presence of large deviations.

In the simulation, two nominal levels are considered $\alpha = 0.05$ and $\alpha = 0.01$. In the first case, the simulation generates 1000 test statistics and p-values, next, we calculate the percentage of tests in which the null hypothesis about homoscedastic error terms is rejected and thus the actual test level is obtained. In the case of $\alpha = 0.01$ the number of generated test statistics is 10000, the remaining steps are performed in the same way. Following the approach allowing for the sampling errors, the test is considered robust if the actual level does not exceed the nominal one by more than two standard deviations, for $\alpha = 0.05$ with 1000 replications, it would mean the actual test level in the range of $0.036 - 0.064$, while for $\alpha = 0.01$ with 10000 replications, the range $0.008 - 0.012$.

To numerically assess whether the test statistic of the robust test constructed with the Heritier and Ronchetti (1994) framework preceded with the additional step of Mallows type M-estimation of Equation (1) preserves the χ^2 distribution, the two series of boxplots are produced within the simulation study. In the first case, to verify the actual test level when the nominal test level is $\alpha = 0.05$, the simulation generates 25 runs with 5000

replications each. In each replication, the classical Breusch-Pagan test and the robust test are conducted, of which eight specifications are considered. In the second case, we consider the nominal level of $\alpha = 0.01$. The number of replications in each run increases to 10000. The evaluated sample size in both cases is $N = 1000$ and the considered data-generating process follows Equation (9) with homoscedastic error terms and no contamination in the sample.

For the evaluation of the power of the test, the error terms characterised by the variance increasing with x_{i1} are considered in the sample sizes $N \in \{500, 1000\}$. Since the classical Breusch-Pagan test is constructed specifically for heteroscedasticity increasing with one of the regressors, and not for groupwise heteroscedasticity, we evaluate only one type of heteroscedasticity. The power evaluation starts with homoscedastic variance, that is $\sigma_i^2 = 1$ for $i = 1, \ldots, N$, where N is a sample size. Then, we introduce heteroscedastic error terms of different magnitude, that is error terms with variance $\sigma_i^2 = \lambda\sigma^2 x_{i1}^2$, where $\sigma^2 = 1$. The evaluated range of degrees of heteroscedasticity contains $\lambda \in \{0.01, 0.02, \ldots, 0.09, 0.1\} \cup \{0.2, 0.3, \ldots, 1.9, 2\}$, where λ is a heteroscedasticity degree. This results in the evaluation of the following range of heteroscedastic variances $\sigma_i^2 \in \{0.01x_{i1}^2, 0.02x_{i1}^2, \ldots, 0.1x_{i1}^2, 0.2x_{i1}^2, \ldots, 1.9x_{i1}^2, 2x_{i1}^2\}$. The point mass contamination is added according to Equation (13) with a contamination level $\epsilon = 0.01$. The contamination scenarios include the same vertical outliers and bad leverage outliers as in the evaluation of the test level. In the simulation, for each degree of heteroscedasticity, 1000 replications providing test statistic and p-value are performed. Next, we calculate the percentage of the tests in which the null hypothesis is rejected, which allows for the assessment of the power of the test. The calculated powers are plotted against the degree of heteroscedasticity, λ, providing the power curves for each combination of the heteroscedasticity test specification, the sample size and the contamination scenario.

Next to the general assessment of the level and the power of the test, the particular type of vertical contamination for heteroscedastic error terms is evaluated. The considered data-generating process is a linear model with only one regressor and intercept, that is:

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i, \text{ for } i = 1, \ldots, N, \tag{22}$$

where $x_{i1} \sim \mathcal{U}_{[1,10]}$, $\beta_0 = \beta_1 = 1$, $e_i \sim \mathcal{N}(0, \sigma_i^2)$, where $\sigma_i^2 = \lambda\sigma^2 x_{i1}^2$ and $\sigma^2 = 1$. The evaluated sample size is $N = 500$ and the considered range of degree of heteroscedasticity

is $\lambda \in \{0.01, 0.02, \ldots, 0.09, 0.1\} \cup \{0.2, 0.3, \ldots, 1.9, 2\}$. In the simulation, four observations are switched to be vertical outliers. These outliers are placed in a way that a value of variable y corresponds to either maximum (two outliers) or minimum (two outliers) values of the dependent variable in the sample, while $x_1 = 1$ for all four outliers. The forced placement of exactly four outliers renders this simulation not following the contaminated distribution for the power as defined in Equation (13). However, with this example, we check whether the test preserves the robustness of efficiency when the high residuals occur also for small values of x_1 under the variance of error terms increasing with this regressor (see Appendix A.2 Figure A.3 for illustrative example how the residuals from such specified contaminated model may look like). The power curves are constructed in the same way as for the general assessment of the test power described before (1000 replications for each value of λ).

4.5 Simulation results

In this section, we present the results of the simulation study evaluating the level and the power of the proposed heteroscedasticity score test. For comparison purposes, the classical Breusch-Pagan test (1979) with the Koenker's correction (1981) is also evaluated.

4.5.1 Evaluation of the level of the test

The level of the proposed heteroscedasticity score test is evaluated for homoscedastic error terms, for sample sizes of 500 and 1000 observations under three contamination scenarios. We consider no contamination scenario, vertical contamination $\epsilon = 0.01$ and bad leverage contamination $\epsilon = 0.01$. Table 4.1 presents the actual test level for both sample sizes when eight different specifications of the test are considered. Specifications differ in the choice of the explanatory variables in the error terms variance function (x_{i1}, x_{i2} or $x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2$), the downweighting function ψ (the Huber function or Tukey's biweight function) and the weight function ω for covariates (using the hat matrix or the robust Mahalanobis distance with the MCD). Additionally, we include the actual level of the classical Breusch-Pagan test.

Table 4.1. *The level of the Breusch-Pagan test and eight specifications of the robust test for sample sizes $N \in \{500, 1000\}$. In the robust test, the explanatory variables in the variance function include either x_1, x_2, or x_1, x_2, x_1^2, x_2^2 as defined in Section 4.4.* **H** *stands for the hat matrix, and MCD stands for the robust Mahalanobis distance with the MCD. The nominal level of the test $\alpha = 0.05$. 1000 replications. In every row, the level values of a robust test with a level closest to the nominal level of 0.05 are underlined.*

Explanatory variables in the variance function x_{i1}, x_{i2}						
N	Contamination	Classical BP	Huber, **H**	Tukey, **H**	Huber, MCD	Tukey, MCD
500	None	0.047	0.046	0.055	0.045	<u>0.053</u>
	Vertical, 1%	0.039	0.031	0.037	0.034	<u>0.039</u>
	Bad leverage, 1%	0.241	<u>0.044</u>	0.034	0.037	0.039
1000	None	0.055	<u>0.051</u>	0.04	<u>0.051</u>	0.041
	Vertical, 1%	0.033	0.044	0.051	0.046	<u>0.05</u>
	Bad leverage, 1%	0.298	<u>0.056</u>	0.039	0.041	<u>0.044</u>
Explanatory variables in the variance function $x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2$						
N	Contamination	Classical BP	Huber, **H**	Tukey, **H**	Huber, MCD	Tukey, MCD
500	None	0.047	0.039	<u>0.045</u>	0.04	0.044
	Vertical, 1%	0.039	0.042	0.054	0.042	<u>0.051</u>
	Bad leverage, 1%	0.241	<u>0.048</u>	0.039	0.041	0.043
1000	None	0.055	<u>0.053</u>	0.062	<u>0.053</u>	0.056
	Vertical, 1%	0.033	<u>0.044</u>	0.066	0.042	0.059
	Bad leverage, 1%	0.298	0.059	0.03	<u>0.043</u>	0.036

In the scenario without contamination, for a sample size $N = 500$, the level of the robust tests with $x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2$ as explanatory variables in the variance function is in all four cases systematically lower than the test level observed when the variance function has simpler specification of explanatory variables, that is x_{i1}, x_{i2}. For a larger sample size, $N = 1000$, we observe the opposite pattern. However, in all the cases considered, the actual test level remains in the range of $0.036 - 0.064$. For a smaller sample size under no contamination scenario, the selection of ψ - Tukey's biweight function with simple variance function results in the actual test level above the nominal level, 0.055 and 0.053, while the choice of the same ψ function but with alternative specification of the variance function yields the actual test level below the nominal one, 0.045 and 0.044. We observe larger differences in the actual test level between the tests with different variance function specifications if ψ is Tukey's biweight function. Comparing the specifications with the same ω, the actual test level is always higher when ψ is Tukey's biweight function. While

controlling for the same ψ, the choice of the ω does not contribute to high differences in the actual test level between different specifications, differences are of 0.01 or 0.02. For a larger sample size, specifications with the same variance function and ψ - the Huber function result in the same actual test level irrespective of the choice of ω. In both cases, those specifications achieve also an actual level closest to the nominal level, 0.051 and 0.053. The selection of ψ - Tukey's biweight function results in an undersized test level when the simple variance function is considered, and an oversized test level when the complex variance function is used.

The contamination scenario with vertical outliers characterises the varying actual test levels depending on the test specification and sample size. For the majority of cases, the actual test level is below the nominal one. The exception is a few cases with ψ - Tukey's biweight function for which the test level is inflated above 0.05. Controlling for the sample size, the variance function and ω function, robust tests with ψ - Tukey's biweight function have higher actual levels. For a smaller sample size, two out of eight robust test specifications achieve considerably low test levels below 0.036 (two specifications with ψ - the Huber function and the simple variance function). For $N = 500$, the level of the classical Breusch-Pagan test stays in an accepted range of $0.036 - 0.064$, while for a larger sample size, it drops to 0.033. For both sample sizes, the specification with ψ - Tukey's biweight function and ω - the robust Mahalanobis distance with the MCD achieves an actual level closest to the nominal, 0.05 for $N = 1000$ with the simple variance function and 0.051 for $N = 500$ with the complex variance function.

For the contamination scenario with bad leverage $\epsilon = 0.01$, the classical Breusch-Pagan test is no longer robust, regardless of the sample size. For a smaller sample size, all eight specifications of the proposed score test yield the actual test level below the nominal one. For a larger sample size, only the specification with ψ - the Huber function and ω - the hat matrix achieves the actual level higher than 0.05, 0.056 for a specification with the simple variance function and 0.059 for a specification with the complex variance function. Except for the specification of ψ - Tukey's biweight function, ω - the hat matrix and the complex variance function when $N = 1000$, the actual test level is in the accepted range $0.036 - 0.064$. The specifications with ω - the hat matrix achieve considerably higher actual level when ψ is the Huber function, regardless of the sample size and the variance function. For both sample sizes, the actual level closest to the nominal level of the test

is achieved with ψ - the Huber function and ω - the hat matrix, 0.048 for $N = 500$ and 0.056 for $N = 1000$, however for $N = 1000$ also the specification with ψ - Tukey's biweight function and ω - the robust Mahalanobis distance with the MCD achieves the actual level of 0.046 that is as far away from the nominal level as for the previously mentioned specification.

Overall, ignoring the specification of the variance function and considering only the combination of sample size and contamination scenario, that is six cases, the two specifications of the robust test are equally often closest to the nominal test level. That is, especially in a contamination scenario with vertical outliers ψ - the Tukey's biweight function, ω - the robust Mahalanobis distance with the MCD, and in a contamination scenario with bad leverage ψ - the Huber function with ω - the hat matrix. Taking back into consideration the specification of the variance function, for a larger sample size, the test with a simple variance function is preferred, while for smaller sample size, it is a complex specification.

We obtain the additional results for the test level 0.01, see Appendix A.3 Table A.1. For a larger sample size, there is a clear preference for the specification with ψ - the Huber function and ω - the robust Mahalanobis distance, which ensures the correct test level under all contamination scenarios. For a smaller sample size, there is a preference for ω - the hat matrix, while ψ - the Tukey's biweight function performs better in vertical contamination scenario and ψ - the Huber function when the contamination comes from the bad leverage.

Next to the evaluation of the test level under different contamination scenarios, we analyse the boxplots of the level of the test in the clean sample setting to check whether systematic oversizing or undersizing is present. Figure 4.1 demonstrates the boxplots for a sample size of 1000 observations with homoscedastic error terms in a scenario without contamination when the nominal level considered is 0.05. The classical Breusch-Pagan test is compared with eight specifications of the proposed heteroscedasticity score test. Specifications differ in the choice of the explanatory variables in the error terms variance function (x_{i1}, x_{i2} or $x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2$), the downweighting function ψ (the Huber function or Tukey's biweight function) and the weight function ω for covariates (using the hat matrix or the robust Mahalanobis distance with the MCD). The median level of all eight specifications lies below the nominal test level of 0.05. The specifications with ψ - the

Huber function are slightly closer (values between $0.049 - 0.0496$) than the specifications with ψ - Tukey's biweight function (values between $0.048 - 0.0488$). In all compared pairs with the same variance function and the same weight function ω, the test level of the test version with the Huber function is higher than that with Tukey's biweight function. The smallest difference of 0.006 between specifications with the Huber function and Tukey's biweight function occurs for the case with simple variance function and ω - the hat matrix. When comparing pairs of different variance specifications (the downweighting function ψ and the weight function ω are the same), specifications with a simple variance function characterise the median test level systematically closer to 0.05 than the specifications with a complex variance function.

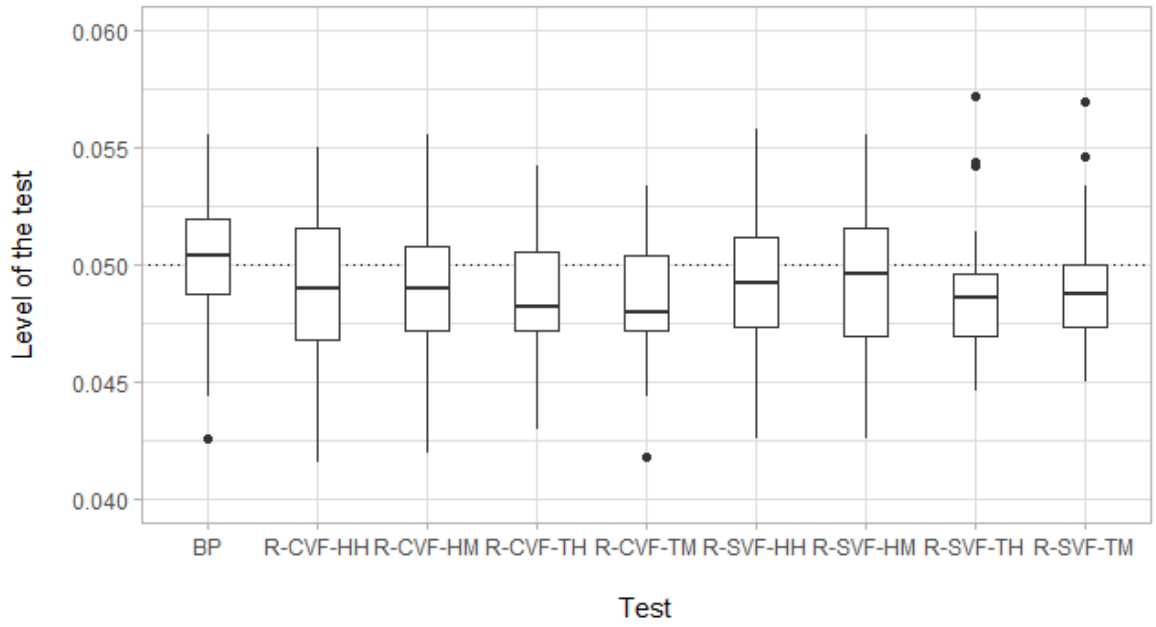


Figure 4.1. The level of the Breusch-Pagan test (BP) and the robust heteroscedasticity tests (R) with a different selection of a downweighting function, ψ, weight function, ω, and explanatory variables in the variance function for a sample size $N = 1000$ without contamination. Homoscedastic error terms. SVF denotes a simple specification of a variance function with x_{i1}, x_{i2} as explanatory variables, while CVF denotes a complex variance function with $x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2$ as explanatory variables. HH denotes ψ - the Huber function and ω - the hat matrix, TH denotes ψ - Tukey's biweight function and ω - the hat matrix, HM denotes ψ - the Huber function and ω - the robust Mahalanobis distance with the MCD, TM denotes ψ - Tukey's biweight function and ω - the robust Mahalanobis distance with the MCD. The level of the test $\alpha = 0.05$ is shown with the black dotted line.

There is no clear pattern observed when we consider the weight function ω. The specification with the simple variance function, ψ - the Huber function and ω - the hat matrix achieves the actual level, 0.0496, closest to the nominal level of 0.05. The actual

test level differs from the prespecified level by 0.0004 and the difference is the same as in the case of the test level of the classical Breusch-Pagan test but in the opposite direction.

The dispersion of the actual test levels varies between test specifications with different downweighting functions ψ. The specifications with Tukey's biweight function have smaller dispersion than the classical Breusch-Pagan test, while for ones with the Huber function the dispersion is equal or larger. However, in all cases, the minimum and the maximum values do not exceed the range of $0.036 - 0.064$. We do not observe a clear pattern of dispersion linked with either a variance function specification or weight function ω. The only case when the dispersion of the robust test is less than that of a classical one, and both a minimum and a maximum level of the robust test are closer to 0.05 is the specification with a complex variance function, ψ - Tukey's biweight function and ω - the hat matrix.

Overall, the evaluation of the boxplots points out the specification with a simple variance function and the Huber function as the one achieving the actual test level closest to the nominal test level. The choice of ω function does not result in high differences, thus both options are possible. Other specifications result in a slightly undersized test level.

Figure 4.2 also demonstrates the boxplots for a sample size of 1000 observations with homoscedastic error terms in a scenario without contamination, but the nominal level considered is 0.01. As in the case of Figure 4.1, the classical Breusch-Pagan test is compared with eight specifications of the proposed heteroscedasticity score test. Specifications differ in the choice of the explanatory variables in the error terms variance function (x_{i1}, x_{i2} or $x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2$), the downweighting function ψ (the Huber function or Tukey's biweight function) and the weight function ω for covariates (using the hat matrix or the robust Mahalanobis distance with the MCD).

The median level of only one out of four robust test specifications with a complex variance function lies exactly at the nominal test level of 0.01, that is specification with ψ - the Tukey's biweight function and ω - the robust Mahalanobis distance, other specifications achieve a lower median test level. For the specifications with a simple variance function three out of four specifications achieve the actual median test level equal to the nominal one. The fourth specification, that is with ψ - the Huber function and ω - the hat matrix, has a slightly higher level of 0.0101. These results point out that specifications

with the simple variance function yield more size-correct results.

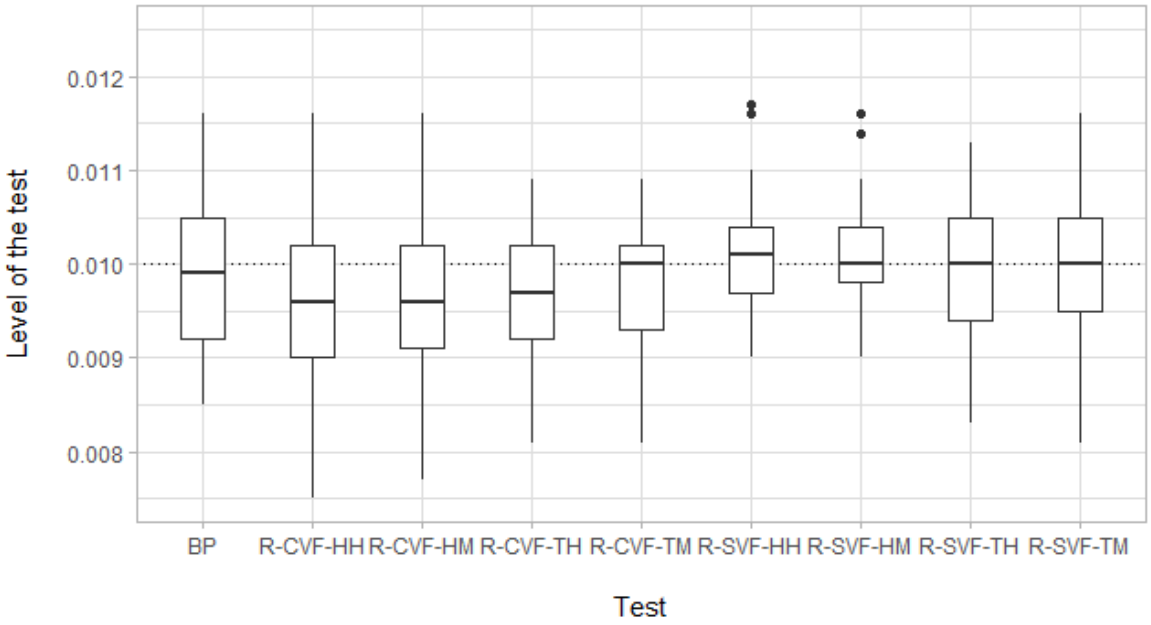


Figure 4.2. The level of the Breusch-Pagan test (BP) and the robust heteroscedasticity tests (R) with a different selection of a downweighting function, ψ, weight function, ω, and explanatory variables in the variance function for a sample size $N = 1000$ without contamination. Homoscedastic error terms. SVF denotes a simple specification of a variance function with x_{i1}, x_{i2} as explanatory variables, while CVF denotes a complex variance function with $x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2$ as explanatory variables. HH denotes ψ - the Huber function and ω - the hat matrix, TH denotes ψ - Tukey's biweight function and ω - the hat matrix, HM denotes ψ - the Huber function and ω - the robust Mahalanobis distance with the MCD, TM denotes ψ - Tukey's biweight function and ω - the robust Mahalanobis distance with the MCD. The level of the test $\alpha = 0.01$ is shown with the black dotted line.

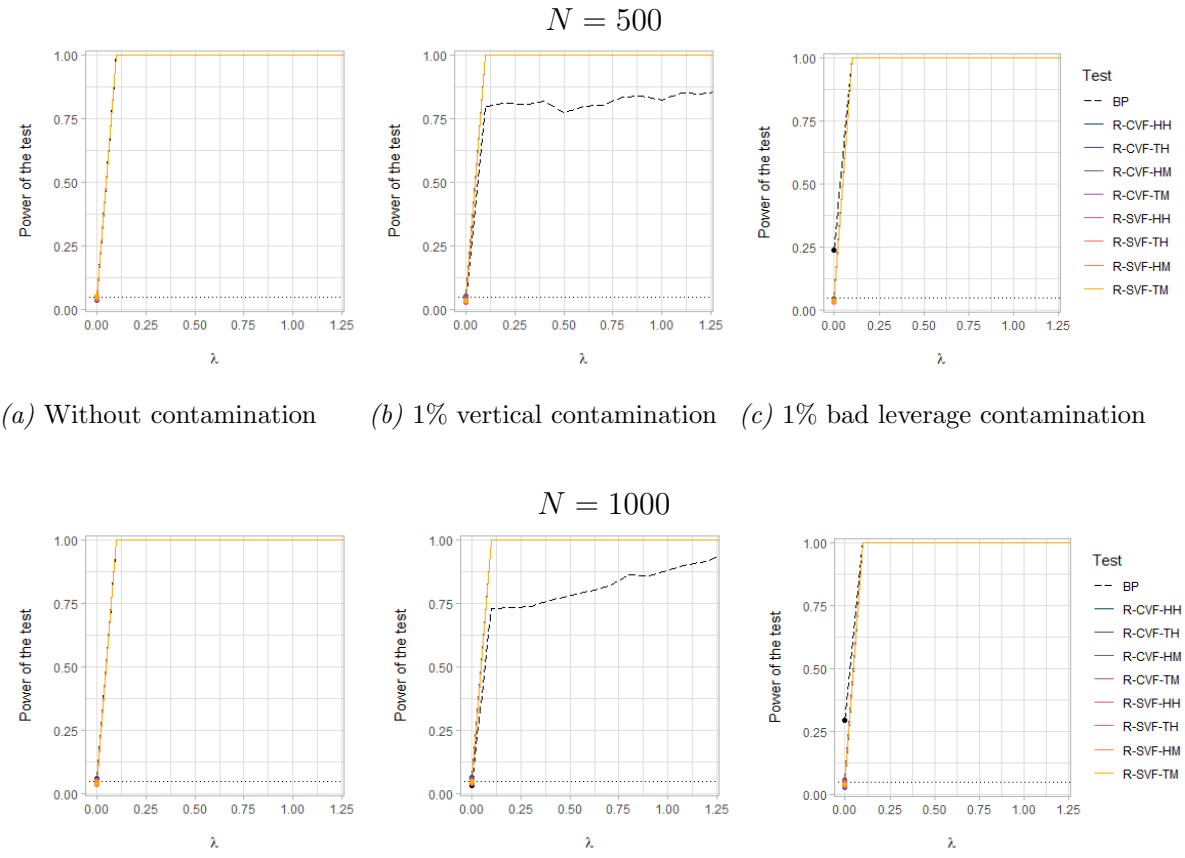
The dispersion of the actual test levels varies between test specifications with different downweighting functions ψ and the variance function specification. For the complex variance function, the specifications with ψ - the Huber function are more dispersed than the specifications with ψ - the Tukey's biweight function, while for the simple variance function, we observe the opposite pattern. For only two out of eight test specifications, the minimum values observed are below the accepted value of 0.008, those are the specifications with complex variance function and ψ - the Huber function. For the remaining six specifications, the minimum and maximum values of the test level lie in the range of $0.08 - 0.012$.

The evaluation of the boxplots, when we consider the nominal level of 0.01, indicates that the specifications with the simple variance function achieve the actual median test level closest to the nominal one. The choice of ω and ψ does not result in high differ-

ences, thus all options are possible. The complex variance specifications yield a slightly undersized test level.

4.5.2 Evaluation of the power of the test

Evaluation of the power of the test is done using the power curves. Error terms with heteroscedasticity increasing with one of the regressors under three contamination scenarios are considered for sample sizes $N = \{500, 1000\}$, see Figure 4.3.



(a) Without contamination *(b)* 1% vertical contamination *(c)* 1% bad leverage contamination

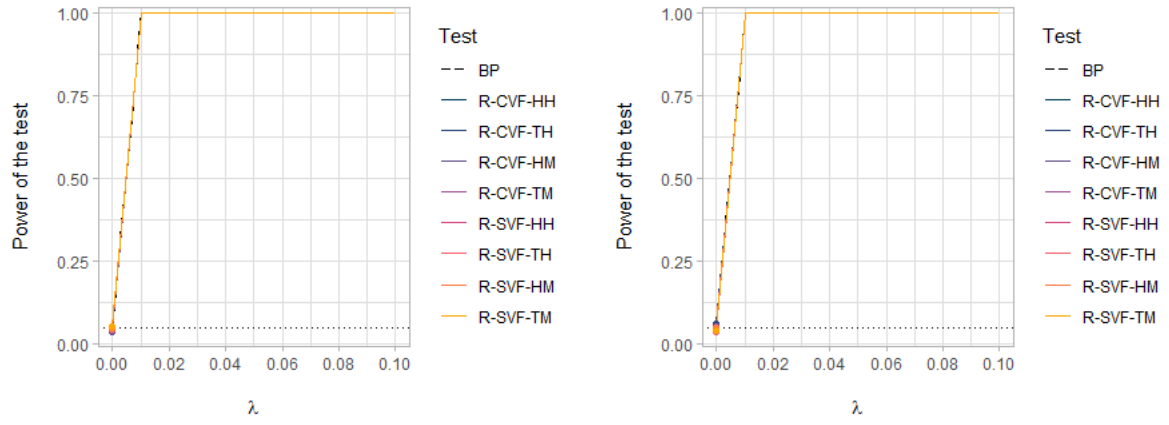
(d) Without contamination *(e)* 1% vertical contamination *(f)* 1% bad leverage contamination

Figure 4.3. Power curves for the classical Breusch-Pagan (BP) test and the robust heteroscedasticity tests (R) with a different selection of a downweighting function, ψ, weight function, ω, and explanatory variables in the variance function. Performed for the model with heteroscedastic error terms, $\sigma_i^2 = \lambda\sigma^2 x_{i1}^2$, where $\sigma^2 = 1$ and $\lambda \in \{0.1, 0.2 \ldots, 1.2\}$, for the sample sizes of $N \in \{500, 1000\}$ under different contamination scenario with either vertical outliers or bad leverage when $\epsilon = 0.01$. SVF denotes a simple specification of a variance function with x_{i1}, x_{i2} as explanatory variables, while CVF denotes a complex variance function with $x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2$ as explanatory variables. HH denotes ψ - the Huber function and ω - the hat matrix, TH denotes ψ - Tukey's biweight function and ω - the hat matrix, HM denotes ψ - the Huber function and ω - the robust Mahalanobis distance with the MCD, TM denotes ψ - Tukey's biweight function and ω - the robust Mahalanobis distance with the MCD. The level of the test $\alpha = 0.05$ is shown with the black dotted line.

Although the originally assessed degree of heteroscedasticity covered λ up to 2, the power curves shown in Figure 4.3 end at $\lambda = 1.2$ as from this degree of heteroscedasticity, the test power remains constant in all cases.

For both sample sizes, there is no difference between robust test specifications under different contamination scenarios. Once error terms are heteroscedastic, the power is equal to 1. All specifications are equally powerful and always correctly reject the null hypothesis of homoscedastic error terms. The performance of the robust test is particularly superior to the performance of the classical Breusch-Pagan test when the vertical contamination scenario is considered, see Figures 4.3b and 4.3e.

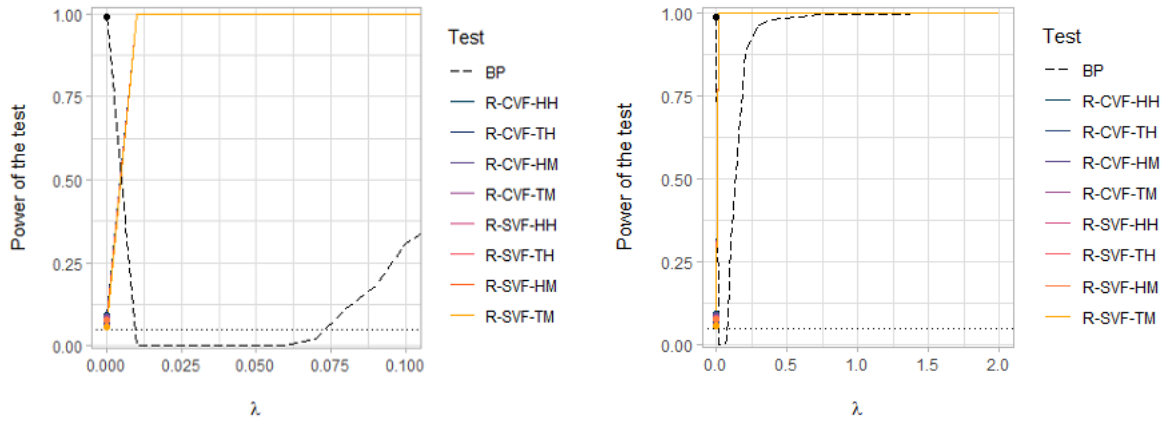
To confirm that the high power of the robust tests is preserved when $\lambda < 0.01$, we perform an additional check for the scenario without contamination for values of λ ranging between 0.01 and 0.1 (with a step of 0.01, smaller than in Figure 4.3 where the step is 0.1). The results once again confirm the equally powerful behaviour of all the specifications of the robust test. All of them achieve the power of 1, once error terms are heteroscedastic, see Figure 4.4. Overall, all specifications of the robust test are equally powerful and there is no preference for any specification.



(a) $N = 500$ *(b)* $N = 1000$

Figure 4.4. Power curves for the classical Breusch-Pagan (BP) test and the robust heteroscedasticity tests (R) with a different selection of a downweighting function, ψ, weight function, ω, and explanatory variables in the variance function. Performed for the model with heteroscedastic error terms, $\sigma_i^2 = \lambda\sigma^2 x_{i1}^2$, where $\sigma^2 = 1$ and $\lambda \in \{0.01, 0.02, \ldots, 0.1\}$, for sample sizes $N \in \{500, 1000\}$ without contamination. SVF denotes a simple specification of a variance function with x_{i1}, x_{i2} as explanatory variables, while CVF denotes a complex variance function with $x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2$ as explanatory variables. HH denotes ψ - the Huber function and ω - the hat matrix, TH denotes ψ - Tukey's biweight function and ω - the hat matrix, HM denotes ψ - the Huber function and ω - the robust Mahalanobis distance with the MCD, TM denotes ψ - Tukey's biweight function and ω - the robust Mahalanobis distance with the MCD. The level of the test $\alpha = 0.05$ is shown with the black dotted line.

The additional results obtained for a particular case of vertical contamination with quasi-symmetrical vertical outliers also confirm the preference for the robust test, see Figure 4.5.



(a) $\lambda \in \{0.01, 0.02, \ldots, 0.1\}$ *(b)* $\lambda \in \{0.1, 0.2 \ldots, 2\}$

Figure 4.5. Power curves for the classical Breusch-Pagan (BP) test and the robust heteroscedasticity tests (R) with a different selection of a downweighting function, ψ, weight function, ω, and explanatory variables in the variance function. Performed for the model with heteroscedastic error terms, $\sigma_i^2 = \lambda\sigma^2 x_{i1}^2$, where $\sigma^2 = 1$ and $\lambda \in \{0.01, 0.02, \ldots, 0.09, 0.1, 0.2. \ldots, 2\}$, for a sample size of $N = 500$ under contamination scenario with four quasi-symmetrical vertical outliers. SVF denotes a simple specification of a variance function with x_{i1}, x_{i2} as explanatory variables, while CVF denotes a complex variance function with $x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2$ as explanatory variables. HH denotes ψ - the Huber function and ω - the hat matrix, TH denotes ψ - Tukey's biweight function and ω - the hat matrix, HM denotes ψ - the Huber function and ω - the robust Mahalanobis distance with the MCD, TM denotes ψ - Tukey's biweight function and ω - the robust Mahalanobis distance with the MCD. The level of the test $\alpha = 0.05$ is shown with the black dotted line.

In that particular case, the classical Breusch-Pagan test turns out to be not robust, while the proposed heteroscedasticity score test achieves high power once the error terms are heteroscedastic. Figure 4.5a demonstrates the power curves for a smaller step of 0.01, we observe the lack of power of the classical Breusch-Pagan test to reject the null hypothesis when heteroscedasticity of a small degree, $0.01 \leq \lambda < 0.07$, is present in the data. Only when λ reaches the value of 0.5 the classical test achieves a power similar to the power of the robust test, see Figure 4.5b. For homoscedastic error terms, see points in Figures 4.5a and 4.5b when $\lambda = 0$ and $\sigma^2 = 1$ for all observations, the robust tests have an actual level close to the nominal level of 0.05, while the classical heteroscedasticity test provides a wrong indication of the heteroscedastic error terms with the p-value of 1.

5 Empirical Application

In this section, we compare the performance of the classical Breusch-Pagan test with the robust alternatives in two real-data applications. The datasets include both heteroscedastic and homoscedastic data. Next to analysis in economic papers, both datasets are also used as datasets accompanying econometric handbooks, publicly available through the R package *AER* (Kleiber & Zeileis, 2008). In original applications, the datasets are treated as if no contamination was present. We verify this and if no outliers are detected, we apply the artificial contamination that should mimic a real case scenario.

5.1 Credit card data

The first real-world data is the dataset used in a paper by Greene (1992) to model credit card expenditure provided by an anonymous credit card company, covering credit card applications in one month in 1988. A sample of 1319 observations from this dataset was later used for illustrative purposes of heteroscedastic error terms in an econometric manual by Greene (2003). Greene (2003) takes the first 100 observations out of 1319 and estimates a linear regression with OLS to model credit card expenditures. Only the subset of applicants with non-zero expenditure is analysed, resulting in a sample of 72 observations. The dataset includes several explanatory variables, from which the author selects the *age* of an applicant, the yearly *income* of the applicant and the yearly *income squared*, and one dummy variable indicating whether the individual *owns a home*. The dependent variable is the average monthly credit card *expenditure*. Heteroscedasticity in the error terms is driven by *income*.

Because of the sample size and lack of random selection of a sample in an illustrative application (Greene, 2003), we prefer to use the whole subset of applicants with non-zero expenditure, thus resulting in a sample size of 1002 observations. The variables *age* and *income* characterise right-skewed distributions. In the *AER* package notes, we can find information that the value of age in certain observations was manually corrected. In fact, in a sample of 1002 observations, there are six individuals whose age is below one year. Those are probably examples of incorrect data coding, and depending on the standardised residual, they could appear in regression as bad or good leverage points with large distances in the explanatory space. We verify this with a diagnostic plot. Rousseeuw and van Zomeren (1990) proposed a robust alternative to a classical regression

diagnostic plot (the Mahalanobis distance versus standardised least squares residuals) where standardised residuals from high-breakdown regression estimator, originally the least median squares, are plotted against robust Mahalanobis distance, based on the Minimum Volume Ellipsoid. In the classical case, the Mahalanobis distance should tell us how far away from the cloud of points the single observation is. However, Rousseeuw and van Zomeren (1990) point out that the classical Mahalanobis distance may suffer from the masking effect and the multivariate outliers do not necessarily have large Mahalanobis distance. Thus, the application of robust methods in the diagnostic plot should help us to correctly identify outlying observations.

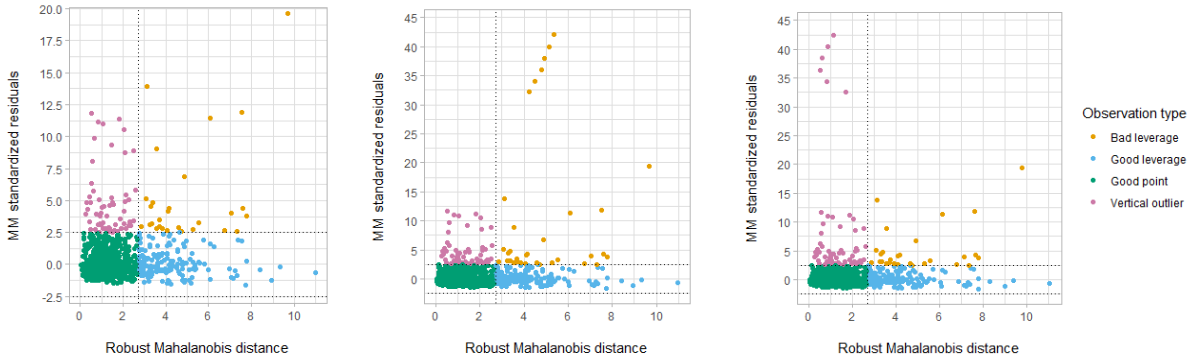
We need to define cutoff values for a diagnostic plot. In the classical regression diagnostic plot, the cutoff value for the Mahalanobis distance is obtained from the χ^2 distribution, namely $\sqrt{\chi^2_{p,0.975}}$, where p is the dimension of data. For the standardised residuals $r_i/\hat{\sigma}$, where r_i is the OLS regression residual and $\hat{\sigma}$ is the scale estimate obtained with the standard deviation, the cutoff values are -2.5 and 2.5. All boundaries are inspired from Rousseeuw and van Zomeren (1990). With these cutoff values, the points with the Mahalanobis distance higher than $\sqrt{\chi^2_{p,0.975}}$ are leverage points. The points with standardised residuals outside $[-2.5, 2.5]$ are regression outliers, labelled vertical outliers if the Mahalanobis distance is smaller than the respective cutoff value, otherwise labelled as bad leverages.

In the robust regression diagnostic plot, we use the robust Mahalanobis distance with mean and covariance robustly estimated based on the Minimum Covariance Determinant, and the residuals from the robust regression estimator, the MM-estimator (Yohai, 1987), standardised with the robust scale estimate, that is the median absolute deviation. Rousseeuw and van Zomeren (1990) determine the cutoff values in the same way as in the classical case. However, Hardin and Rocke (2005) note that the robust Mahalanobis distances have an exact χ^2 distribution when data follows the normal distribution. If data deviates from the Gaussian distribution, finding distributional cutoff values using χ^2 distribution may fail and too many points are declared outliers. Therefore, they propose asymptotic formulas based on $\mathcal{F}$ distribution to calculate cutoff values for outlying distances computed with the MCD. Hardin and Rocke (2005) argue that the $\mathcal{F}$ distribution provides better distributional information about outliers than the χ^2 distribution. Thus, in one of the diagnostic plots, we use $\mathcal{F}$ distribution with degrees of freedom calculated

from the adjusted asymptotic formulas and scaling constant from the asymptotic formula to determine the cutoff values, for the details see Hardin and Rocke (2005, p. 940).

Overall, to identify outliers, we construct several diagnostic plots to cross-check what type and how many outliers they detect. The plots include the classical diagnostic plot with all three explanatory variables, the classical diagnostic plot with only *age* and *income* in the Mahalanobis distance, the robust diagnostic plot with only *age* and *income* in the Mahalanobis distance and cutoff values according to the χ^2 distribution, the robust diagnostic plot with only *age* and *income* in the Mahalanobis distance and cutoff values according to the $\mathcal{F}$ distribution. It is not feasible to construct a robust diagnostic plot with all three explanatory variables, because of the dummy variable that renders computation of the MCD numerically impossible.

Figure 5.1a demonstrates the robust regression diagnostic plot with the cutoff values from the χ^2 distribution when the original data is considered. Although in the original application, the dataset is not treated as a contaminated sample, the diagnostic plot finds outliers in a sample.



(a) Original sample *(b)* Added bad leverages *(c)* Added vertical outliers

Figure 5.1. Robust regression diagnostic plots for three contamination scenarios of the credit card dataset (Greene, 1992). In the plots, the robust Mahalanobis distance is plotted against standardised residuals of the MM regression estimator. The cutoff values for the robust Mahalanobis distance are found according to $\sqrt{\chi^2_{2,0.975}}$. Sample size $N = 1002$.

Both classical diagnostic plots identify 25 outliers (bad leverages or vertical outliers) constituting approximately 2.5% of the sample, while both robust label 9% of the observations (91 observations) as outlying, see Table 5.1 and Appendix A.4 Table A.3. As expected, the robust diagnostic plots detect more outliers than the classical ones. Both robust diagnostic plots identify the same number of outliers, but the number of bad leverages and vertical outliers differs between different specifications of cutoff values. We

expected the specification with χ^2 distribution to label too many points as outliers, however, it does not happen. Thus, we can infer that the data distribution does not deviate considerably from the Gaussian distribution and the χ^2 distribution-based cutoff values can be trusted. The classical diagnostic plots and another robust diagnostic plot can be found in Appendix A.4 Figure A.4.

Table 5.1. *Count and types of points in the regression for three contamination scenarios of the credit card dataset (Greene, 1992). Points identified with the robust regression diagnostic plot (standardised residuals of the MM regression estimator vs the robust Mahalanobis distance with the cutoff values* $\sqrt{\chi^2_{2,0.975}}$*). Sample size* $N = 1002$*.*

Contamination scenario	Good point	Good leverage	Bad leverage	Vertical outlier
Original data	752	159	28	63
Added vertical outliers	755	152	28	67
Added bad leverages	754	153	34	61

To introduce more extreme outliers, we modify the values of the dependent variable and *age* variable of six wrongly coded observations. As a result, we obtain datasets with a higher number of either vertical outliers or bad leverage points, see Table 5.1. In the case of additional bad leverage points, we modify the value of *credit card expenditure* of those six observations (for the descriptive statistics of the original variables, see Appendix A.4 Table A.2). We take the maximum value from the sample and multiply it by 1.5, and successively increase by 0.1 the number by which we multiply the maximum value, in the next step, six inflated values of the dependent variable are assigned to six observations with wrongly coded data. In the sample with artificially added vertical outliers, we use the same modification of *credit card expenditure* value as in the case of bad leverage, but we correct the *age* of each applicant and assign them the value of median age, successively increasing the value of the assigned age by 1 for each successive observation. The original sample and modified datasets are subject to the heteroscedasticity tests with the classical Breusch-Pagan and the robust heteroscedasticity score test. The original application of data provides information about the variance function ($income + income^2$), thus we use this function specification in all tests. We evaluate four specifications of the robust test with a different selection of the downweighting function ψ - Tukey's biweight or the Huber function, and the weight function ω - the hat matrix or the robust Mahalanobis distance with the MCD.

Table 5.2. *The level of the Breusch-Pagan test (Classical BP) and four specifications of the robust heteroscedasticity test: Huber,* $\boldsymbol{H}$ *(ψ - the Huber function, and ω - the hat matrix), Tukey,* $\boldsymbol{H}$ *(ψ - Tukey's biweight function, and ω - the hat matrix), Huber, MCD (ψ - the Huber function, and ω - the robust Mahalanobis distance with the MCD), and Tukey, MCD (ψ - Tukey's biweight function, and ω - the robust Mahalanobis distance with the MCD) for three contamination scenarios of the credit card dataset (Greene, 1992). The variance function with income and income^2 as explanatory variables. Sample size $N = 1002$.*

Contamination scenario	Classical BP	Huber, **H**	Tukey, **H**	Huber, MCD	Tukey, MCD
Original data	0.00	0.00	0.00	0.00	0.00
Added bad leverage	0.67	0.00	0.00	0.00	0.00
Added vertical outliers	0.85	0.00	0.00	0.00	0.00

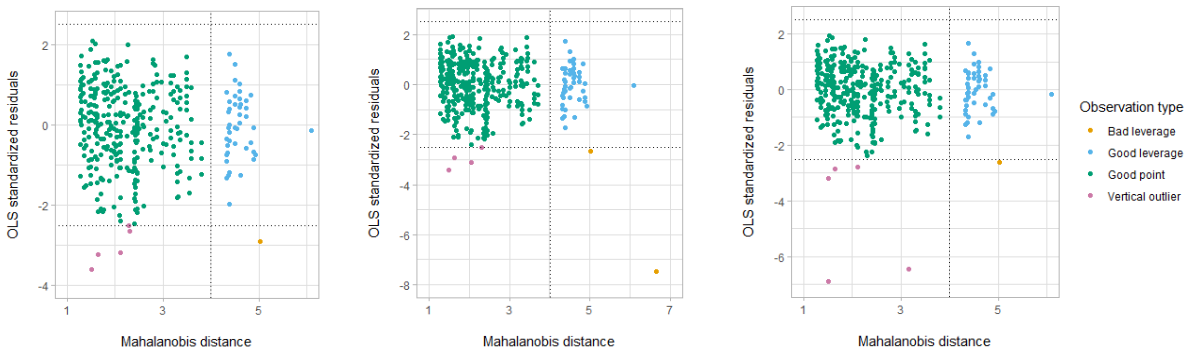
Table 5.2 demonstrates the results of the heteroscedasticity tests. The robust tests reject the null hypothesis of homoscedastic error terms in all the cases considered. Even though the overall number of outliers in the artificial scenarios does not differ considerably from the original scenario, the Breusch-Pagan test fails to detect heteroscedastic error terms in the scenarios with the increased number of both vertical outliers and bad leverage points. The imputed outliers are more extreme than the original data points, thus their impact on the classical Breusch-Pagan test is visible in the inflated level of the test. Overall, we can see that using the classical heteroscedasticity test for the illustrative purposes of heteroscedasticity in Greene (2003) does not result in the wrong inference about the variance of error terms despite the outliers in the sample, however, the small increase in the number of outlying observations leads to incorrect results.

5.2 Teacher ratings data

In the second empirical application we use the dataset analysed in a paper by Hamermesh and Parker (2005), with data on course evaluations, professor and course characteristics collected for 463 courses over three academic years at the University of Texas in Austin. The dataset is also used for general illustrative purposes in an econometric handbook by Stock and Watson (2007). We choose this dataset for two reasons. First, in the original application, the error terms are assumed to be homoscedastic, unlike in Section 5.1. Secondly, this dataset allows us to demonstrate what problems can arise when we apply outliers diagnostic and robust tests to the datasets including mainly dummy variables.

In the originally analysed regression, the dependent variable is *teaching evaluation score*, while the set of explanatory variables includes one continuous variable *the in-*

structor's physical appearance rating and six dummy variables (some information about the teacher: *the instructor's gender*, *whether the instructor belongs to a non-Caucasian minority*, *whether the instructor is a native English speaker*, *whether the instructor is on tenure track*, and two characteristics of the course: *upper/lower course division*, *single-credit elective*). The presence of several binary predictors precludes computing the MCD since the algorithm encounters numerical problems. As a result, we cannot use the robust Mahalanobis distance and therefore also the robust diagnostic plot. An alternative to the robust Mahalanobis distance would be to use the diagonal elements of the hat matrix to identify leverage points, however, Rousseeuw and van Zomeren (1990) point out that the hat matrix suffers from the masking effect and does not detect leverage points correctly. Therefore, we construct only the classical version of the regression diagnostic plot following the same steps as in Section 5.1.



(a) Original sample *(b)* Added bad leverages *(c)* Added vertical outliers

Figure 5.2. Classical regression diagnostic plots for three contamination scenarios of the teacher ratings dataset (Hamermesh & Parker, 2005). In the plots, the Mahalanobis distance is plotted against standardised residuals of the OLS regression estimator. The cutoff values for the Mahalanobis distance are found according to $\sqrt{\chi^2_{7,0.975}}$. Sample size $N = 463$

The diagnostic plot finds one bad leverage and five vertical outliers in the original sample, see Figure 5.2a and Table 5.3. From the simulation study, see Section 3.4, we can suspect that the small number of vertical outliers should not deflate the level of the classical Breusch-Pagan test and the hypothesis about homoscedastic error terms should not be rejected. However, the presence of a bad leverage outlier may incorrectly suggest heteroscedasticity in the data.

Table 5.3. *Count and types of points in the regression for three contamination scenarios of the teacher ratings dataset (Hamermesh & Parker, 2005). Points identified with the classical regression diagnostic plot (standardised residuals of the OLS regression estimator vs the Mahalanobis distance with the cutoff values* $\sqrt{\chi^2_{7,0.975}}$*). Sample size* $N = 463$*.*

Contamination scenario	Good point	Good leverage	Bad leverage	Vertical outlier
Original data	404	53	1	5
Added vertical outliers	404	53	1	5
Added bad leverages	404	53	2	4

We introduce two additional scenarios to check how the classical and the robust tests behave in the original sample and when the number of outliers is increased. In the first scenario, the number of vertical outliers is held constant, but for two of them, the value of *teaching evaluation score* is changed to 0, making these outliers more extreme. In the second scenario, one vertical outlier is replaced by the additional bad leverage (for one observation *teaching evaluation score* is changed to 0, and the value of *instructor's physical appearance rating* increases to 5), see Figure 5.2 and Table 5.3. For the descriptive statistics of the original variables, see Appendix A.4 Table A.4.

We test the heteroscedasticity of error terms in all three contamination scenarios with the classical Breusch-Pagan test and two specifications of the robust test. The presence of several dummy variables in the explanatory space precludes the application of the robust test with ω - the robust Mahalanobis distance with the MCD, thus, we evaluate only the specifications with ω - the hat matrix with the downweighting functions ψ - Tukey's biweight or the Huber function. The variable suspected to drive the variance is *the instructor's physical appearance rating.*

Table 5.4. *The level of the Breusch-Pagan test (Classical BP) and two specifications of the robust heteroscedasticity tests: Huber,* $\mathbf{H}$ *(*ψ *- the Huber function, and* ω *- the hat matrix), and Tukey,* $\mathbf{H}$ *(*ψ *- Tukey's biweight function, and* ω *- the hat matrix), for three contamination scenarios of the teacher ratings dataset (Hamermesh & Parker, 2005). The variance function with the instructor's physical appearance rating as an explanatory variable. Sample size* $N = 463$*.*

Contamination scenario	Classical BP	Huber, $\mathbf{H}$	Tukey, $\mathbf{H}$
Original data	0.45	0.80	0.41
Added bad leverages	0.00	0.49	0.46
Added vertical outliers	0.46	0.93	0.43

Table 5.4 shows the results of the heteroscedasticity tests under three contamination scenarios. The results of both robust tests in all three cases of data contamination do not suggest rejecting the null hypothesis of homoscedasticity. The classical Breusch-Pagan remains robust in the scenario with more extreme vertical outliers, but with the increased number of bad leverage (2 observations) it incorrectly indicates heteroscedasticity in the data. Thus, we can infer that in this data configuration, the classical test breaks with only two bad leverage outliers, while the robust counterparts preserve robustness. Based on the simulation study evaluating the Breusch-Pagan test robustness, see Section 3.4.2, we expected that under the scenario with the limited number of vertical outliers (only 1 % of the sample), the test would remain robust and indeed, this empirical application confirms our expectations.

However, the classical Breusch-Pagan test and its robust alternatives are not able to detect groupwise heteroscedasticity, which is possible to occur in a dataset with several dummy variables that can easily divide the sample into two subsets, for example, based on *the instructor's gender*. Therefore, with the available robust score test, we cannot be completely sure that the data is homoscedastic in this aspect as well.

6 Conclusion

In this paper, we investigated the robustness properties of the heteroscedasticity tests. In particular, we examined three classical tests: the Breusch-Pagan test (1979), the Goldfeld-Quandt test (1965) and the Harrison-McCabe test (1979), and we proposed a robust alternative to the Breusch-Pagan test based on the robust score test framework of Heritier and Ronchetti (1994). We showed that the influence function of the OLS estimator is unbounded, and consequently, the classical heteroscedasticity tests which are constructed upon this estimator inherit its robustness properties and are nonrobust to outliers. The results of the simulation study confirmed that none of the classical heteroscedasticity tests preserves robustness against outliers for both homoscedastic and heteroscedastic data. With these findings, we proposed a robust alternative constructed with the framework of the robust bounded-influence score test (Heritier & Ronchetti, 1994). The application of the Mallows type score function ensures the bounded influence function of the M-estimator based on which the test statistic is constructed. In the simulation study, the proposed test remains size-correct and powerful in the presence of outliers. The empirical application

also showed that the proposed test is more robust than the classical Breusch-Pagan test.

The main limitation of our paper considers the type of heteroscedasticity the robust test can detect and the simulation study. The application of the robust heteroscedasticity score test is limited only to the data-generating processes where heteroscedasticity increasing with one of the regressors is present. There is still a gap in the research considering the robust tests suitable for groupwise heteroscedasticity. In the simulation study, we investigate the behaviour of the tests only for the nominal levels of $\alpha = 0.05$ and $\alpha = 0.01$. The more extensive study can also cover smaller levels $\alpha = 0.001$ and $\alpha = 0.0001$ with the increased number of simulation runs to obtain more accurate results. The simulation can also be supplemented with the empirical breakdown analysis to verify how many outliers the robust test can withstand before the test statistic is distorted and does not provide reliable results anymore. Besides, we did not examine moderate vertical and bad leverage outliers and focused solely on extreme outliers. Additionally, further research can consider small-sample properties of the constructed robust test and different choices in its construction, such as a selection of other estimators of scale instead of MAD to ensure higher efficiency, for example, M-estimator of dispersion (Maronna et al., 2019).

References

Ali, M. M., & Giaccotto, C. (1984). A study of several new and existing tests for heteroscedasticity in the general linear model. *Journal of Econometrics*, *26*(3), 355–373. https://doi.org/10.1016/0304-4076(84)90026-5

Alih, E., & Ong, H. C. (2015). An outlier-resistant test for heteroscedasticity in linear models. *Journal of Applied Statistics*, *42*(8), 1617–1634. https://doi.org/10.1080/02664763.2015.1004623

Berenguer-Rico, V., & Wilms, I. (2021). Heteroscedasticity testing after outlier removal. *Econometric Reviews*, *40*(1), 51–85. https://doi.org/10.1080/07474938.2020.1735749

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, *47*(5), 1287–1294. https://doi.org/10.2307/1911963

Cantoni, E., & Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, *96*(455), 1022–1030. https://doi.org/10.1198/016214501753209004

Dufour, J.-M., Khalaf, L., Bernard, J.-T., & Genest, I. (2004). Simulation-based finite-sample tests for heteroskedasticity and ARCH effects. *Journal of Econometrics*, *122*(2), 317–347.

Fomby, T. B., Johnson, S. R., & Hill, R. C. (1984). Heteroscedasticity. In *Advanced econometric methods* (1st ed., pp. 170–204). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-8746-4_9

Glejser, H. (1969). A new test for heteroskedasticity. *Journal of the American Statistical Association*, *64*(325), 316–323. https://doi.org/10.1080/01621459.1969.10500976

Goldfeld, S. M., & Quandt, R. E. (1965). Some tests for homoscedasticity. *Journal of the American Statistical Association*, *60*(310), 539–547. https://doi.org/10.1080/01621459.1965.10480811

Greene, W. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Greene, W. H. (1992). *A statistical model for credit scoring* (NYU Working Paper No. EC-92-29).

Hamermesh, D. S., & Parker, A. (2005). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review*, *24*(4), 369–376. https://doi.org/10.1016/j.econedurev.2004.07.013

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*(346), 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions* (1st ed.). New YorkJohn Wiley & Sons. https://doi.org/10.1002/9781118186435

Hardin, J., & Rocke, D. M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, *14*(4), 928–946. https://doi.org/10.1198/106186005X77685

Harrison, M. J., & McCabe, B. P. M. (1979). A test for heteroscedasticity based on ordinary least squares residuals. *Journal of the American Statistical Association*, *74*(366a), 494–499. https://doi.org/10.1080/01621459.1979.10482544

Harvey, A., & Phillips, G. (1974). A comparison of the power of some tests for heteroskedasticity in the general linear model. *Journal of Econometrics*, *2*(4), 307–316. https://doi.org/10.1016/0304-4076(74)90016-5

Harvey, A. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, *44*, 461–65. https://doi.org/10.2307/1913974

Heritier, S., & Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association*, *89*(427), 897–904. https://doi.org/10.1080/01621459.1994.10476822

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, *1*(5), 799–821. https://doi.org/10.1214/aos/1176342503

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9780470434697

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, *48*(3/4), 419–426. https://doi.org/10.1093/biomet/48.3-4.419

Kleiber, C., & Zeileis, A. (2008). *Applied econometrics with R* (1st ed.). New York: Springer-Verlag.

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, *17*(1), 107–112. https://doi.org/10.1016/0304-4076(81)90062-2

Krämer, W., & Sonnberger, H. (1986). *The linear regression model under test* (1st ed.). Heidelberg: Physica-Verlag HD. https://doi.org/10.1007/978-3-642-95876-2_1

Lyon, J. D., & Tsai, C.-L. (1996). A comparison of tests for heteroscedasticity. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *45*(3), 337–349. https://doi.org/10.2307/2988471

Mallows, C. (1975). *On some topics in robustness* (Unpublished memorandum). Bell Telephone Laboratories, Murray Hill, NJ.

Maronna, R., Martin, R., Yohai, V., & Salibián-Barrera, M. (2019). *Robust statistics: Theory and methods (with R)* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Noether, G. E. (1955). On a theorem of Pitman. *The Annals of Mathematical Statistics*, *26*(1), 64–68. https://doi.org/10.1214/aoms/1177728593

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika*, *62*(2), 223–241. https://doi.org/10.1093/biomet/62.2.223

Rana, S., Midi, H., & Imon, A. (2008). A robust modification of the Goldfeld-Quandt test for the detection of heteroscedasticity in the presence of outliers. *Journal of Mathematics and Statistics*, *4*.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*(388), 871–880. https://doi.org/10.1080/01621459.1984.10477105

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, *85*(411), 633–639. https://doi.org/10.1080/01621459.1990.10474920

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze & W. Wertz (Eds.), *Mathematical statistics and applications* (1st ed., pp. 283–297). Dordrecht: Reidel.

Sharma, S. C., & Giaccotto, C. (1991). Power and robustness of jackknife and likelihood-ratio tests for grouped heteroscedasticity. *Journal of Econometrics*, *49*(3), 343–372. https://doi.org/10.1016/0304-4076(91)90002-U

Stock, J., & Watson, M. W. (2007). *Introduction to econometrics* (2nd ed.). Boston, MA: Addison Wesley.

Tukey, J. (1960). A survey of sampling from contaminated distributions. In I. Olkin (Ed.), *Contributions to probability and statistics. Essays in honor of Harold Hotelling* (1st ed.). Stanford, CA: Stanford University Press.

Verbeek, M. (2004). *A guide to modern econometrics* (2nd ed.). Chichester, West Sussex: John Wiley & Sons.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. https://doi.org/10.2307/1912934

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, *15*(2), 642–656. https://doi.org/10.1214/aos/1176350366

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10.

A Appendix

A.1 Evaluation of the level and the power of the modified Goldfeld-Quandt test with outlier-removal strategy (Rana et al., 2008)

We construct the modified Goldfeld-Quandt test with an outlier-removal strategy following the procedure of Rana et al. (2008). The test is expected to detect groupwise heteroscedastic error terms, thus we start with ordering observations according to the regressor or index which is suspected to drive the variance of error terms and divide the whole sample into two subsets like in the classical Goldfeld-Quandt test. In the next steps, we first detect outliers with the Least Trimmed Squares estimator (Rousseeuw, 1984), and next we compute the deletion residuals for the entire sample based on regression coefficients estimated with a clear set. Finally, the median of the squared deletion residuals is computed for two groups of observations. The test statistic is a ratio of those medians. Rana et al. (2008) state that under normality, the test statistic follows the $\mathcal{F}$ distribution with the degrees of freedom, each of $(N - c_N - 2k)/2$, where N is a sample size, k is the number of all regressors in the estimation, and c_N is the number of central observations omitted before the outliers detection starts. However, Rana et al. (2008) do not specify the value of c_N.

To evaluate the level and power of the modified Goldfeld-Quandt test, we conduct a simulation study similar to the evaluation performed for the classical tests (see Section 3.3) and the robust score test (see Section 4.4). We consider the data-generating process that follows the linear model with two regressors and intercept, see Equation (9), for a sample size $N = 500$. In the evaluation of the level of the test, we analyse the homoscedastic error terms $\sigma_i^2 = \sigma^2 = 1$, and we consider two different observation ordering, that is according to either regressor x_1 or observation index. While in the analysis of the power of the test, we consider two types of heteroscedastic error terms: groupwise heteroscedasticity and the variance increasing with one of the regressors (see Section 3.3 for more details about types of heteroscedasticity). We start with homoscedastic error terms, and then the evaluated range of degrees of heteroscedasticity λ covers $\lambda \in \{1.01, 1.02, \ldots, 1.09, 1.1\} \cup \{1.2, 1.3, \ldots, 2.9, 3\}$ for groupwise heteroscedasticity and $\lambda \in \{0.01, 0.02, \ldots, 0.09, 0.1\} \cup \{0.2, 0.3, \ldots, 1.9, 2\}$ for the variance increasing with a regressor. In the first case, we consider index ordering, while in the second case, ordering according to the regressor x_1.

In the level analysis, we construct the box plots with 100 simulation runs and 1000 replications in each run. In the power analysis, we construct the power curves with 1000 replications for each value of λ. Three contamination scenarios are considered: a sample without any contamination and two datasets including outliers, either vertical outliers (placed at $y^* = -100$) or bad leverage points (placed at $y^* = x_1^* = x_2^* = -50$). The point mass contamination is added according to Equation (12), in level evaluation, and Equation (13), in power evaluation, with a degree of contamination $\epsilon = 0.01$. We consider the nominal level of the test $\alpha = 0.05$.

Figure A.1 shows the boxplots under three contamination scenarios. Irrespective of the scenario considered, the modified Goldfeld-Quandt test is oversized, with a median level of approximately 0.16 considerably above the nominal level of $\alpha = 0.05$. The proposed modification to the classical test does not result in a size-correct robust test.

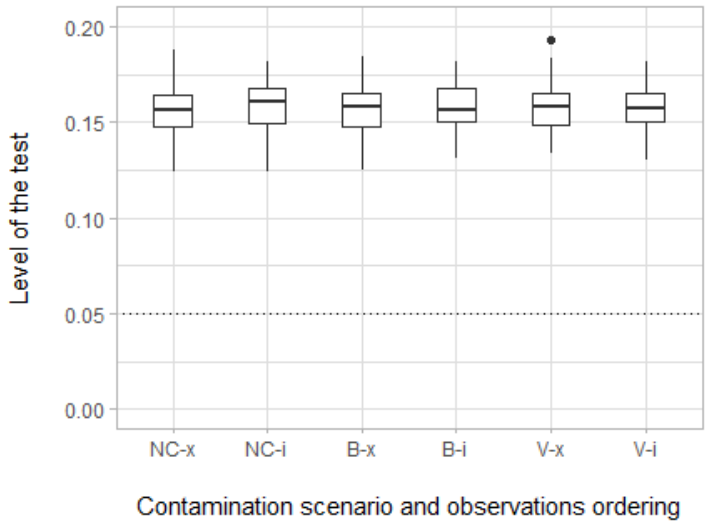
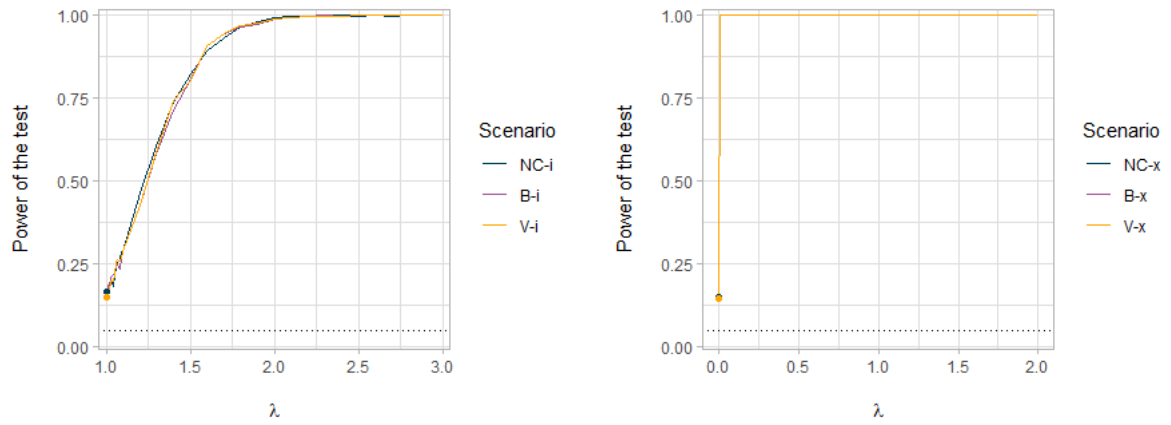


Figure A.1. The level of the modified Goldfeld-Quandt test for sample size $N = 500$ under three contamination scenarios. NC denotes the scenario without contamination, B denotes the model contaminated with bad leverage points, and V with vertical outliers. Ordering according to a regressor x_1 is denoted with x and index ordering with i. The level of the test $\alpha = 0.05$ is shown with the black dotted line.

Figure A.2 demonstrates the power curves for the case of groupwise heteroscedasticity, see Figure A.2a, and the case of heteroscedasticity increasing with one of the regressors, see Figure A.2b. We observe that in the second case, the test achieves a power of 1, once error terms are heteroscedastic and $\lambda > 0$, irrespective of the contamination scenario. However, in the case of groupwise heteroscedasticity, the power curves achieve a value of 1 for $\lambda > 2$. It points out that the modified Goldfeld-Quandt test can detect heteroscedastic

error terms only when the variance in one half is twice as big as in the other half of the sample.

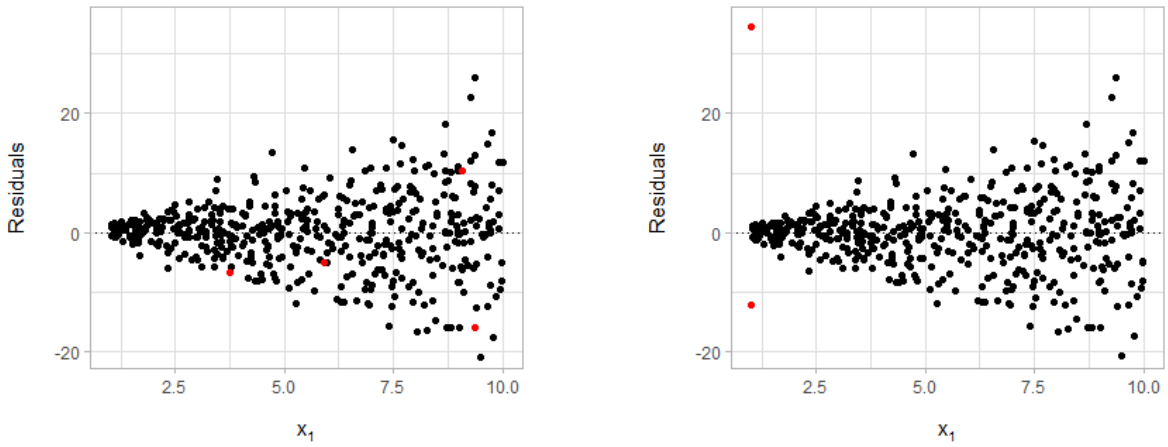


(a) $\sigma_i^2 = \sigma^2$ for $i \leq N/2$ and $\sigma_i^2 = \lambda\sigma^2$ for $i > N/2$ *(b)* $\sigma_i^2 = \lambda\sigma^2 x_{i1}^2$

Figure A.2. Power curves for the modified Goldfeld-Quandt test performed on the residuals from linear regression (see Equation (9)) with (*a*) groupwise heteroscedastic error terms, and (*b*) error terms characterised with heteroscedasticity increasing with one of the regressors, for the sample size $N = 500$ under three contamination scenarios ($\epsilon = 0.01$). NC denotes the scenario without contamination, B denotes the model contaminated with bad leverage points, and V with vertical outliers. Ordering according to a regressor x_1 is denoted with x and index ordering with i. The level of the test $\alpha = 0.05$ is shown with the black dotted line.

The evaluation of the level of the modified Goldfeld-Quandt test indicates that the modification proposed by Rana et al. (2008) to robustify a non-robust component of the test does not result in a size-correct test. Even though the test achieves high power for heteroscedastic error terms when contamination is present, the lack of robustness in terms of the test level precludes us from acknowledging that the test is a robust alternative to the classical tests.

A.2 Auxiliary Figures Section 4.4



(a) Without contamination

(b) Vertical contamination

Figure A.3. Illustrative example of how the regression residuals change when the quasi-symmetrical vertical contamination is present. Residuals from the linear model, $y_i = x_{i1} + e_i$ $(i = 1, \ldots, 500)$, with heteroscedastic error terms, where $e_i \sim \mathcal{N}(0, \sigma_i^2)$ and $\sigma_i^2 = x_{i1}^2$. Red points denote the points changed to vertical outliers in the contaminated sample.

A.3 Additional Results Section 4.5.1

Table A.1. *The level of the Breusch-Pagan test and eight specifications of the robust test for sample sizes $N \in \{500, 1000\}$. In the robust test, the explanatory variables in the variance function include either x_1, x_2, or x_1, x_2, x_1^2, x_2^2 as defined in Section 4.4. Weight function ω - **H** stands for the hat matrix, and MCD stands for the robust Mahalanobis distance with the MCD. The nominal level of the test $\alpha = 0.01$. 10000 replications. In every row, the level values of a robust test with a level closest to the nominal level of 0.01 and in the range of 0.008-0.012 are underlined.*

Explanatory variables in the variance function x_{i1}, x_{i2}						
N	Contamination	Classical BP	Huber, **H**	Tukey, **H**	Huber, MCD	Tukey, MCD
500	None	0.0104	0.0084	<u>0.0094</u>	0.0087	0.0092
	Vertical, 1%	0.0072	0.0079	<u>0.0094</u>	0.0077	0.009
	Bad leverage, 1%	0.2066	<u>0.0112</u>	0.0071	0.0073	0.0078
1000	None	0.0108	<u>0.0101</u>	0.0109	0.0102	0.011
	Vertical, 1%	0.0081	0.0095	0.0118	<u>0.0097</u>	0.0115
	Bad leverage, 1%	0.283	0.0137	0.0085	<u>0.0098</u>	0.009
Explanatory variables in the variance function $x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2$						
N	Contamination	Classical BP	Huber, **H**	Tukey, **H**	Huber, MCD	Tukey, MCD
500	None	0.0104	0.008	<u>0.0096</u>	0.0078	<u>0.0096</u>
	Vertical, 1%	0.0072	<u>0.0087</u>	0.0137	0.0085	0.0122
	Bad leverage, 1%	0.2066	<u>0.0098</u>	0.0076	0.0079	0.0074
1000	None	0.0108	<u>0.0099</u>	0.0092	<u>0.0101</u>	0.0094
	Vertical, 1%	0.0081	<u>0.011</u>	0.0205	0.0112	0.0189
	Bad leverage, 1%	0.283	0.013	0.0084	<u>0.0108</u>	0.0089

A.4 Additional Results Section 5

Table A.2. *Descriptive statistics of the continuous variables from the credit card dataset (Greene, 1992) used in the regression in Section 5.1. All variables refer to values observed for a single applicant. Average monthly credit card expenditure in USD, age in years plus twelfths of a year, yearly income in USD 10,000.*

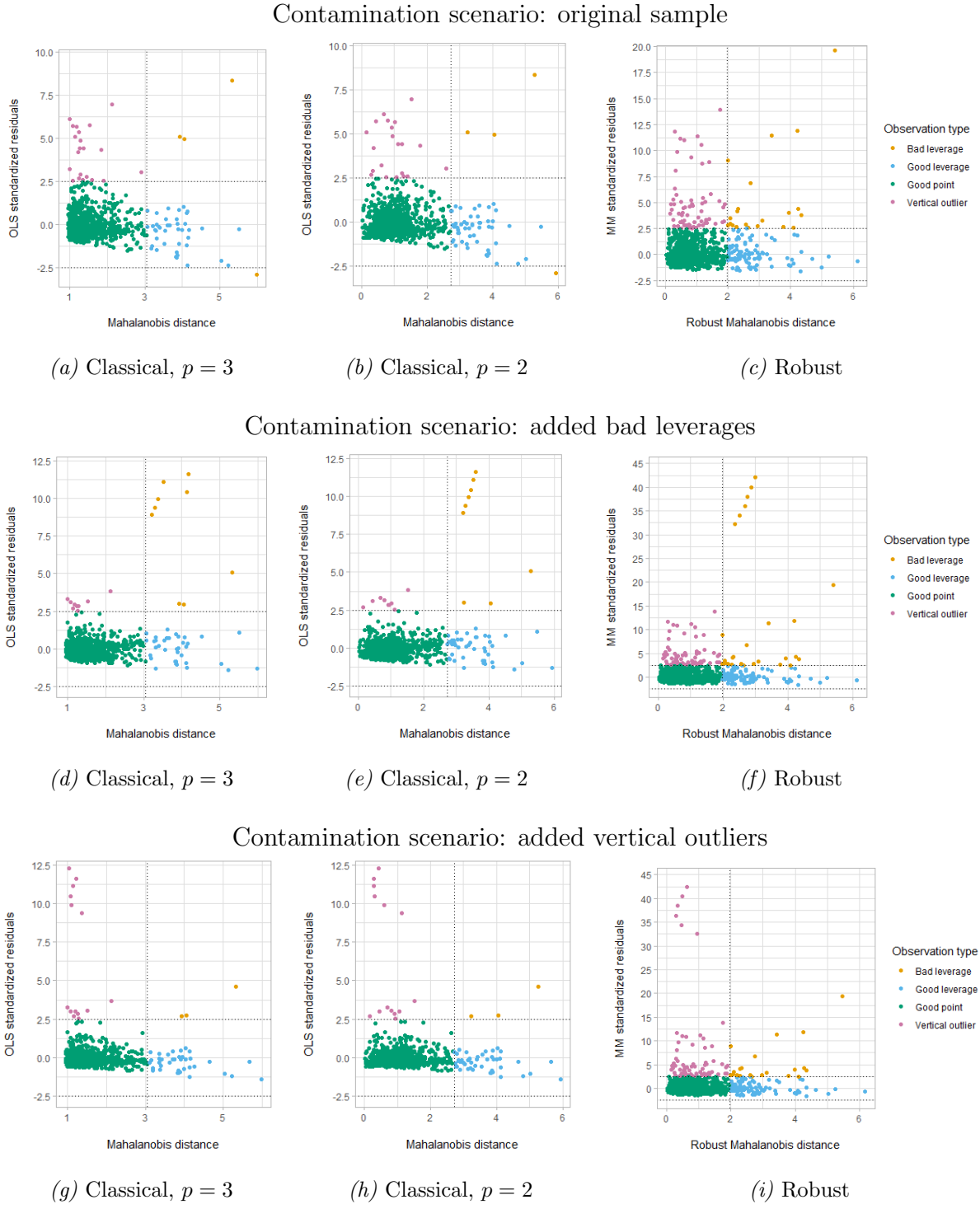
Variable	Count	Min	0.25 Quantile	Median	0.75 Quantile	Max
credit card expenditure	1002	0.312	70.596	156.578	319.200	3099.505
age	1002	0.167	25.354	31.167	39.646	83.500
yearly income	1002	0.210	2.350	3.000	4.000	13.500

Table A.3. *Count and types of points in the regression for three contamination scenarios of the credit card dataset (Greene, 1992). Points identified with three different regression diagnostic plots: the robust regression diagnostic plot with standardised residuals of the MM regression estimator vs the robust Mahalanobis distance with the cutoff values from $\mathcal{F}$ distribution (Robust), the classical regression diagnostic plot with standardised residuals of the OLS regression estimator vs the Mahalanobis distance with the cutoff values $\sqrt{\chi^2_{3,0.975}}$ (Classical, $p = 3$), and the classical regression diagnostic plot with standardised residuals of the OLS regression estimator vs the Mahalanobis distance with the cutoff values $\sqrt{\chi^2_{2,0.975}}$ (Classical, $p = 2$). Sample size $N = 1002$.*

Contamination scenario: original sample				
Diagnostic plot	Good point	Good leverage	Bad leverage	Vertical outlier
Robust	590	321	45	46
Classical, $p = 3$	936	41	4	21
Classical, $p = 2$	925	52	4	21
Contamination scenario: added bad leverages				
Diagnostic	Good point	Good leverage	Bad leverage	Vertical outlier
Robust	592	315	51	44
Classical, $p = 3$	948	36	9	9
Classical, $p = 2$	937	47	9	9
Contamination scenario: added vertical outliers				
Diagnostic	Good point	Good leverage	Bad leverage	Vertical outlier
Robust	590	317	45	50
Classical, $p = 3$	948	37	3	14
Classical, $p = 2$	937	48	3	14

Table A.4. *Descriptive statistics of the continuous variables from the teacher ratings dataset (Hamermesh & Parker, 2005) used in the regression in Section 5.2. All variables refer to values obtained for a single course evaluation. The teaching evaluation score on a scale of 1 (very unsatisfactory) to 5 (excellent), the instructor's physical appearance rating on a scale of 1 (lowest) to 10 (highest), shifted to have a mean of zero.*

Variable	Count	Min	0.25 Quantile	Median	0.75 Quantile	Max
teaching evaluation score	463	2.1	3.6	4.0	4.4	5
physical appearance rating	463	-1.45	-0.66	-0.06	0.55	1.97



(a) Classical, $p = 3$ *(b)* Classical, $p = 2$ *(c)* Robust

(d) Classical, $p = 3$ *(e)* Classical, $p = 2$ *(f)* Robust

(g) Classical, $p = 3$ *(h)* Classical, $p = 2$ *(i)* Robust

Figure A.4. Regression diagnostic plots for three contamination scenarios of the credit card dataset (Greene, 1992). Points identified with three different regression diagnostic plots: the robust regression diagnostic plot with standardised residuals of the MM regression estimator vs the robust Mahalanobis distance with the cutoff values from $\mathcal{F}$ distribution (Robust), the classical regression diagnostic plot with standardised residuals of the OLS regression estimator vs the Mahalanobis distance with the cutoff values $\sqrt{\chi^2_{3,0.975}}$ (Classical, $p = 3$), and the classical regression diagnostic plot with standardised residuals of the OLS regression estimator vs the Mahalanobis distance with the cutoff values $\sqrt{\chi^2_{2,0.975}}$ (Classical, $p = 2$). Sample size $N = 1002$.