Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Econometrics

A study of structural break diagnostic tests and their robust counterparts

Kyle Muller (622311)

Supervisor:	Mikhail Zhelonkin
Second assessor:	-
Date final version:	30th April 2024

Abstract

Structural breaks in regressions break one of the key assumptions of the Ordinary Least Squares estimator, and lead to incorrect assumptions about data if they are not taken into account. For known or suspected structural breaks, the Wald and Chow Break test are commonly used to test their significance. These diagnostic test however fail if the data is contaminated by outliers. To this end, this paper analyses the robustness of the Wald and Chow Break test and suggests robust versions of the Wald test based on a paper by Heritier and Ronchetti (1994). A simulation study and real data example are done to verify the robustness properties, in terms of power and level, of the robust versions of the Wald test. Here it is found that the level of the tests remain consistent and the robust tests remain power for varying structural breaks and sample sizes.

Keywords: Structural Break, Chow Break, Wald, Outlier, Robust Test

Contents

1 Introduction **3**

2 Structural Break Tests **4**

2.1 Chow Break Test 5

2.2 Wald Test 7

2.3 Robustness 8

3 Robust Structural Break Tests **9**

3.1 M-estimator 10

4 Simulation **12**

4.1 Outlier effect 13

4.2 Stability Analysis 13

4.3 Power Analysis 14

4.4 Level Analysis 14

5 Simulation Results **14**

5.1 Outlier effect 14

5.2 Results for Wald and Chow Break test 15

5.2.1 Evaluation of the sensitivity 15

5.2.2 Evaluation of the power 19

5.2.3 Evaluation of the level 20

5.3 Results for Robust Structural Break test 22

5.3.1 Evaluation of the power 22

5.3.2 Evaluation of the level 26

6 Real Data Example **28**

7 Conclusion **33**

References **35**

A Appendix **38**

A.1 Outlier effect simulation results 38

A.2 Power Analysis simulation results 39

A.3 Diagnostic plots real data example 45

1 Introduction

Modern day data analysis relies heavily on the aggregation of data from a variety of sources in order to obtain significant sample sizes, or in order to make generalizing statements about sectors of industry. An example of this would be in a paper by Storfjell, Omoike and Ohlson (2008). In this paper the authors analyse the patient care time versus the cost associated with the care. In order to do this, they consider 14 medical-surgical nursing units, and combine their data. In the act of combining data from various sources, an assumption is made, namely that all parameters are consistent amongst the different medical-surgical units. This assumption is known as the assumption of no structural breaks, or the assumption of consistent regression parameters or variance. An example of a paper that studies structural breaks is the paper by Tabot (2023), which uses the Chow (1960) Break test along with the Wald (1943) test in order to verify whether there are significant variance breaks in stock market returns of varying markets, where the break would occur at a certain time period. In this paper the data from varying markets is not aggregated, but these sets of data are studied separately to identify whether they have a structural break in the dimension of time. The Chow Break and Wald test rely on a predefined structural breakpoint, and in literature there are two ways by which these breakpoints are commonly determined. First, a structural breakpoint can be correlated to large events in world, for example the Great Depression as analysed by Kirkwood (1972), or the more recent COVID-19 global pandemic. The second and more common method by which structural breakpoints are determined is through analysis of the data using specific methods and tests. A method to detect structural breaks is described by Bai and Perron (1998), namely minimizing the sum of squared residuals for a set amount of structural breaks and another method is the CUSUM method explained by Koshti (2011). There are three structural break finding methods related to the Wald test, namely the Mean Wald test and the Exponential Wald test described by Andrews and Ploberger (1994) and the Supremum Wald test created by Quandt (1960). These test to find breaks can be used separately or can be combined as done by Bai, Lumsdaine and Stock (1998) in order to get suspected structural breaks with confidence bounds.

In most papers the presence of outliers is considered to be insignificant, or worse, not considered at all. The use of the Wald and Chow break test is predicated on a set of rather strict assumptions, commonly referred to as the classical assumptions of ordinary least squares which are described by Heij, Dijk, Kloek and Franses (2004). The presence of outliers in diagnostic tests results in the assumptions not being met, which effects the accuracy of the diagnostic tests. For example, the Chow Break test can have a large error in rejection probability when heteroskedasticity is not accounted for (Giles & Scott, 1992). Robust analysis for structural breaks has

been done before, Gagliardini, Trojani and Urga (2005) suggest a robust GMM test for structural breaks. As a different approach to outliers in a structural break setting, Giordani, Kohn and van Dijk (2007) suggest that in the perceived presence of a structural break and outliers, non linearity could be present, and construct a Bayesian framework to estimate models. Chen and Huang (2018) even created a non-parametric test for smooth structural changes in panel data, with limited robustness properties. This shows that the research into robust structural break tests has both practical and scientific relevance.

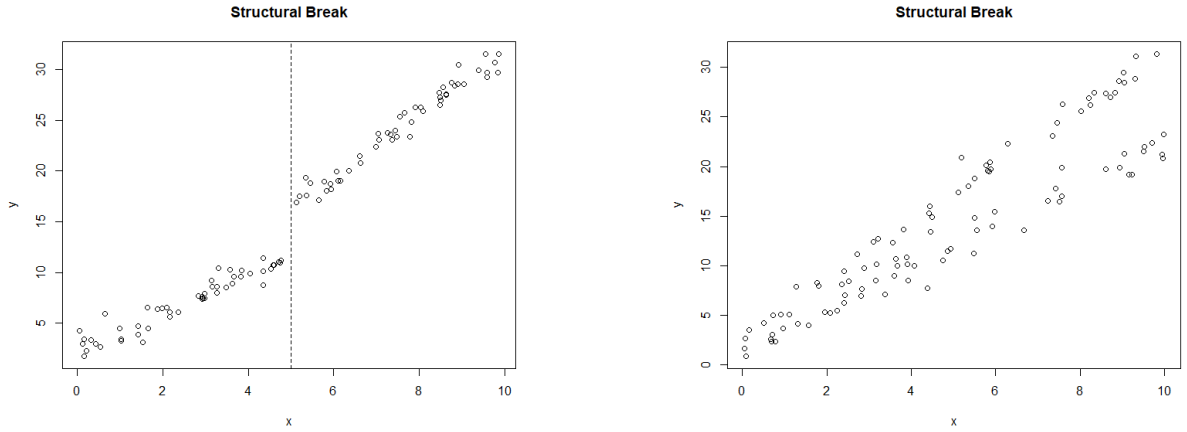
The core focus of this paper will be to investigate the behavior of the Chow (1960) Break test and the Wald (1943) test in the presence of outliers, to describe robust alternatives to these tests, and establish the accuracy, disadvantages and advantages of the robust structural break diagnostic tests.

The paper will be structured in the following way. First in Section 2 a general introduction to structural breaks and Chow Break and Wald tests will be provided. Additionally a measure of robustness, the influence function suggested by Hampel, Stahel, Ronchetti and Rousseeuw (1986), will be discussed, as this is key to defining what a robust structural break diagnostic test looks like. After this, in Section 3, a robust structural break test will be constructed using a framework suggested by Heritier and Ronchetti (1994). With all tests properly defined, in Section 4 a simulation study will be described which will be used to analyse the robustness of the Wald, Chow Break, and robust Wald test and compare their best-use scenarios. The study will analyse the sensitivity, level and power of all the diagnostic tests for a variety of sample sizes, structural break scenarios, structural breakpoints and outlier types, which will be presented in Section 5. This will be followed in Section 6 by a real data example, where the diagnostic tests are used on data from the US labor productivity in the manufacturing/durables sector from February 1947 to April 2001. Finally, Section 7 will summarize the main findings of the paper and provide judgement on the strengths and weaknesses of the paper, along with suggested follow-up research based on the findings or shortcomings of the paper.

2 Structural Break Tests

A structural break occurs when the relation of a dependent variable changes in relation to its regressors. The relation between dependent and independent variable changes across one of the dimensions of the data. Two of those dimensions would be time and space. A space-wise structural break occurs when data passes a threshold in size, for example $x > 5$, as can be seen in Figure 2.1a. For all data generated with independent variables beyond this threshold, the relation with the dependent variable is different. A time-wise structural break occurs when the

relation between the dependent and independent variable changes after a certain point in time, an example of which can be seen in Figure 2.1b.



(a) Space-wise structural break

(b) Time-wise structural break

Figure 2.1: Image showing 2 types of structural break, the vertical line in 2.1a shows where the structural break occurs.

It should be immediately apparent from Figure 2.1b, that visually identifying a structural break is not always possible. Methods to identify the moment of a structural breaks exist, however, for the scope of this paper, we only focus on data where the suspected structural break occurs at a known moment in space or time.

In the case that the space or time of a suspected structural break is known, a variety of diagnostic tests exist to test whether a structural break actually occurs and is significant. The two diagnostic tests that this paper will focus on are the Chow Break and Wald test.

2.1 Chow Break Test

The Chow Break Test introduced by Chow (1960) tests the consistency of coefficients of a linear regression of a dataset. A dataset, D_T consisting of N points is split into two separate datasets of sizes n_1 and n_2, D_1 and D_2 respectively, where $n_1 + n_2 = N$. The type of structural break is defined by how D_T is split. Linear regression of D_T results in Sum of Squared Errors (**SSE**) S_T, linear regression of D_1 yields **SSE** S_1 and the linear regression of D_2 has a **SSE** of S_2.

The following linear regression model is considered

$$y_i = \bar{x}_i^T \bar{\beta} + e_i, \text{ for } i = 1, ..., N, \tag{1}$$

where $\bar{\beta} = \begin{bmatrix}\beta_0 & \beta_1 & \cdots & \beta_k\end{bmatrix}^T$, $\bar{x}_i = \begin{bmatrix}1 & x_{i,1} & \cdots & x_{i,k}\end{bmatrix}^T$ and $e_i \sim \mathcal{N}(0, \sigma^2)$. This model can be written into matrix form as follows

$$Y = X\bar{\beta} + \bar{e}, \tag{2}$$

with $Y = \begin{bmatrix}y_1 & y_2 & \cdots & y_N\end{bmatrix}^T$, $X = \begin{bmatrix}\bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_N\end{bmatrix}^T$ and $\bar{e} = \begin{bmatrix}e_1 & e_2 & \cdots & e_N\end{bmatrix}^T$.
The linear regression model of D_T takes the exact form of (2), which results in the sum of squared errors S_T. Linear regression of D_1 and D_2 uses models $Y_1 = X_1\bar{\beta}_1 + \bar{e}_1$ and $Y_2 = X_2\bar{\beta}_2 + \bar{e}_2$, resulting in the sum of squared errors S_1 and S_2 respectively. For completeness it is important to note that $Y_1 = \begin{bmatrix}y_1 & y_2 & \cdots & y_{n_1}\end{bmatrix}^T$, $X_1 = \begin{bmatrix}\bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_{n_1}\end{bmatrix}^T$ and $\bar{e}_1 = \begin{bmatrix}e_1 & e_2 & \cdots & e_{n_1}\end{bmatrix}^T$, and Y_2, X_2 and $\bar{e}_2$ are the remaining data in D_T. The combination of the two linear regression can be written as

$$\begin{pmatrix}Y_1\\Y_2\end{pmatrix} = \begin{pmatrix}X_1 & \bar{0}\\\bar{0} & X_2\end{pmatrix}\begin{pmatrix}\bar{\beta}_1\\\bar{\beta}_2\end{pmatrix} + \begin{pmatrix}\bar{e}_1\\\bar{e}_2\end{pmatrix}. \tag{3}$$

The null hypothesis H_0 of the Chow Break test is that of no structural break, meaning that from (3), $\bar{\beta}_1 = \bar{\beta}_2 = \bar{\beta}$, which is equivalent to stating that $S_T = S_1 + S_2$. This means that under the null (3) takes the form of

$$\begin{pmatrix}Y_1\\Y_2\end{pmatrix} = \begin{pmatrix}X_1\\X_2\end{pmatrix}\beta + \begin{pmatrix}\bar{e}_1\\\bar{e}_2\end{pmatrix}. \tag{4}$$

The Chow Break test statistic is defined as

$$\frac{S_T - (S_1 + S_2)/k}{(S_1 + S_2)/(n_1 + n_2 - 2k)} \tag{5}$$

This test statistic follows a F distribution with k and $n_1 + n_2 - 2k$ degrees of freedom under the null. The full derivation of the test statistic can be found in the paper by Chow (1960).
The Chow Break test was shown to perform poorly in the presence of heteroskedasticity and autocorrelation as shown by Giles and Scott (1992). A solution to these particular issues was suggested by Sun and Wang (2022), but will not be considered in the paper, as the focus is on the effects of outliers on the diagnostic test and heteroskedasticity and autocorrelation should not be an issue.

2.2 Wald Test

The Wald test is used in general to test a hypothesis on a parameter. For a single hypothesis on a single parameter, the Wald test takes the form of:

$$W_{uni} = \frac{(\hat{\theta} - \theta_0)^2}{var(\hat{\theta})} \tag{6}$$

Here $\hat{\theta}$ is the maximum likelihood estimator of the parameter θ and θ_0 is the hypothesis that is tested on the parameter. When expanding the Wald test to multiple parameters, and possibly multiple hypotheses, it is possible to write the hypotheses as $R\hat{\theta}_n - r$. In the single parameter case from (6), $R = 1$ and $r = \theta_0$.

In order to use the Wald test to test for structural breaks, we have to define matrix R and vector r. Using the Wald test in a structural break scenario means that we equivocate $\bar{\beta}_1$ and $\bar{\beta}_2$ from (3). This means that $\hat{\theta}_n$ takes the form of $\begin{bmatrix}\hat{\bar{\beta}}_1 & \hat{\bar{\beta}}_2\end{bmatrix}'$, a vector of $2k$ parameters, resulting in the following definitions for R and r;

$$R = \begin{bmatrix} I_k & -I_k \end{bmatrix} \tag{7}$$

$$r = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}' \tag{8}$$

where I_k is a k by k identity matrix. This means that R is a k by $2k$ matrix, and r is a 1 by k vector of zeros, where k is the amount of coefficients contained in $\bar{\beta}$. The structural break Wald test takes the form of (Martin, 2013a, p.141):

$$W_{mult} = (R\hat{\theta}_n - r)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\hat{\theta}_n - r) \tag{9}$$

For $n \to \infty$ we can replace $\sigma^2 R(X'X)^{-1}$ by any consistent estimator for the covariance matrix V, namely $\hat{V}_n$ and then the Wald test takes the form of

$$W_{mult} = (R\hat{\theta}_n - r)'[R(\hat{V}_n/n)R']^{-1}(R\hat{\theta}_n - r) \tag{10}$$

The Wald test only uses the unrestricted model, unlike the Chow Break test, where the restricted model is also determined. A strict downside to the Wald test is that it is not invariant under reformulation, as changing the scale of one parameter in $\hat{\theta}_n$ has no effect on R or r, but could effect the outcome of the test statistic (Parker, 2015).

A key assumption of the noted Wald test in (10) is that $VAR(R\hat{\theta}_n) = R[VAR(\hat{\theta}_n)]$. This holds because $R\hat{\theta}_n$ results in M equations of the form $\hat{\bar{\beta}}_{1,i} - \hat{\bar{\beta}}_{2,i}$ for $i = 1, \cdots, M$, where $\hat{\bar{\beta}}_{1,i}$ and $\hat{\bar{\beta}}_{2,i}$ are completely independent. Under the null hypothesis the Wald test has a χ^2_M distribution,

with M the number of restrictions being tested. This distribution is based on the relation that

$$Z_1 = \frac{\hat{\theta}_1 - \theta_0}{se(\hat{\theta})} \sim \mathcal{N}(0,1), \tag{11}$$

which requires that $\hat{\theta}$ is an unbiased estimator for θ_0. If Z is indeed normally distributed, then (6) follows a χ_1^2 distribution. In order to show that (10) follows a χ_M^2 distribution, an adapted deravation of Martin (2013b) is used. The equation can be rewritten to

$$W_{mult} = (R\hat{\theta}_n - r)'(S^{-1})'S^{-1}(R\hat{\theta}_n - r). \tag{12}$$

In (12), the following Cholesky decomposition is used:

$$SS' = R(\hat{V}_n/n)R'. \tag{13}$$

Using (13), (12) can be rewritten to

$$W_{mult} = (R\hat{\theta}_n - r)'(S^{-1})'S^{-1}(R\hat{\theta}_n - r) = Z'Z = \sum_{i=1}^{M} Z_i^2, \tag{14}$$

which proves that under the null hypothesis W_{mult} is the sum of M squared normal distributions, which is equivalent to a χ_M^2 distribution.

2.3 Robustness

In order to measure the robustness of the Wald and Chow Break test, we make use of the influence function first introduced in Hampel (1974). The influence function is defined as

$$IF(\{\mathbf{x}, \mathbf{y}\}, T, F) = \lim_{\epsilon \to 0^+} \frac{T\{(1-\epsilon)F + \epsilon\Delta_{\mathbf{x},\mathbf{y}}\} - T\{F\}}{\epsilon}. \tag{15}$$

In the influence function, $T\{\}$ is a statistical function, F is the distribution from which a parameter will be estimated, $\mathbf{x}$ is a vector of observations and ϵ is the contamination level, which describes the fraction of observations which are outliers. In (15), the outliers originate from the point mass contamination $\Delta_{\mathbf{x}}$, as this is also used later in the paper, but in more general terms, the outliers are drawn from a distribution G, with the only restriction being that $F \neq G$. The verification of the robustness of the Wald and Chow Break test does not require the construction of a unique influence function for each of the diagnostic tests. Rather, the robustness of the test statistic of the Wald and Chow Break test is inherited from the robustness of the estimator used according to Heritier and Ronchetti (1994). A regression estimator is considered robust to

local outliers if its influence function is bounded (Hampel et al., 1986). So, in order to check the robustness of the two diagnostic test, only the robustness of the Ordinary Least Squares estimator needs to be checked.

The OLS estimator uses the regression model described in (1). The solution to the OLS estimator can be written in matrix notation as

$$\hat{\beta} = arg\min_{\beta} E[(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)], \tag{16}$$

and in the contaminated case

$$\hat{\beta} = arg\min_{\beta}(1-\epsilon)E[(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)] + \epsilon(\mathbf{y}_i - \mathbf{x}_i\beta)'(\mathbf{y}_i - \mathbf{x}_i\beta), \tag{17}$$

where $(\mathbf{y}_i, \mathbf{x}_i)$ is the contaminated point. This leads to the following influence function according to Hampel et al. (1986);

$$IF(\{\mathbf{x}_i, \mathbf{y}_i\}, T, F) = \left(\int \mathbf{X}'\mathbf{X}dF\right)^{-1} \mathbf{x}_i^T(\mathbf{y}_i - \mathbf{x}_i^T\beta). \tag{18}$$

It is clear that the influence function of OLS is not bounded in either the $\mathbf{x}$ or $\mathbf{y}$ dimension. Therefore, OLS is not robust against any outliers. This means that in the presence of outliers the OLS parameter estimation will be biased, which in turn will make the test statistic for both the Wald and Chow Break test biased. The effect of outliers will be expanded on in the results of the simulation study in Section 5.

3 Robust Structural Break Tests

For the creation of a robust structural break test, a robust Wald test is constructed using the framework from Heritier and Ronchetti (1994). In accordance with Heritier and Ronchetti (1994) the Wald test will take the form of

$$W_n^2 = (\mathbf{T}_n)_{(2)}^t V(\Psi, F_\theta)_{(22)}^{-1}(\mathbf{T}_n)_{(2)}. \tag{19}$$

For the sake of consistent notation, it is important to note that W_n^2 is equivalent to W_{mult} from (10), that is to say, they both refer to the Wald test.

In Heritier and Ronchetti (1994) the M-estimators are defined by

$$\sum_{i=1}^{n} \Psi(\mathbf{z}_i, \mathbf{T}_n) = 0, \tag{20}$$

where $\mathbf{z}_i$ are a sample of iid random vectors, analogously the independent variable x from the data generating processes described in Section 4. $\mathbf{T}_n$ is the M-estimator for β, and $\Psi()$ is the score function. In Heritier and Ronchetti (1994) the complete parameter of interest β is defined by the null hypothesis H_0: $\beta = \beta_0$, with $(\beta_0)_{(2)} = 0$ and $(\beta_0)_{(1)}$ unspecified. Linking this to the Wald test described in Section 2.2, $(\mathbf{T}_n)^t_{(2)}$ is the robust estimator for $R\hat{\theta}_n - r$ from (10), and $V(\Psi, F_\theta)^{-1}_{(22)}$ is the variance covariance matrix for $R\hat{\theta}_n - r$ having used $(\mathbf{T}_n)^t_{(2)}$. As was stated in 2.2, using the independence of parameters in $R\hat{\theta}_n - r$, $\hat{\theta}_n$ can be robustly estimated using an M-estimator, and $V(\Psi, F_\theta)^{-1}_{(22)}$ can be written as $R(\hat{V}_n/n)R'$ with $(\hat{V}_n/n)$ the variance covariance matrix estimated using the robust $\hat{\theta}_n$. The fact that this can be done greatly simplifies the calculations needed for the M-estimators, which will be described below.

3.1 M-estimator

A regular M-estimator $(\mathbf{T}_n)^t_{(2)}$ takes the form of

$$\sum_{i=1}^{n} \psi\left(\frac{r(\hat{\theta}_n)}{\hat{\sigma}}\right)\mathbf{x}_i = 0, \tag{21}$$

where $r(\hat{\theta}_n)$ is shorthand for the residuals of (3), $\hat{\sigma}$ is the estimated variance of the residuals, and $\psi(\cdots)$ is the chosen down-weighting function. The most common down-weighting functions are the Huber and Tukey bisquare function. The Huber down-weighting function has the form of

$$\psi(x;c) = \begin{cases} x, & \text{if } |x| \leq c, \\ c\ sign(x), & \text{if } |x| > c. \end{cases} \tag{22}$$

The Tukey bisquare down-weighting function has the form of

$$\psi(x;c) = \begin{cases} x\left(\left(\frac{x}{c}\right)^2 - 1\right)^2, & \text{if } |x| \leq c, \\ 0, & \text{if } |x| > c. \end{cases} \tag{23}$$

The parameter c in the down-weighting functions is used to determine the asymptotic efficiency and hence is known as the tuning constant. In order to achieve a 95% efficiency values of $c = 1.345$ and $c = 4.685$ are chosen for the Huber and Tukey bisquare function respectively. A downside of a pure M-estimator is that the influence function is only bounded for y, not for x, meaning that when bad leverage outliers are considered, the estimator is not robust (Khan, Ali, Ahmad, Manzoor & Hussain, 2021).

Since bad leverage outliers are considered, a M-estimator that is robust to these kinds of outliers

should be considered. A Mallows type M-estimator is robust against bad leverage outliers (Carroll & Pederson, 1993). The Mallows type M-estimator takes the form of

$$\sum_{i=1}^{n} \psi \left(\frac{r(\hat{\theta}_n)}{\hat{\sigma}} \right) w(\mathbf{x}_i)\mathbf{x}_i = 0, \tag{24}$$

with $w(\mathbf{x}_i)$ a chosen weight function, and all other parameters equivalent to those in the M-estimator described in (21). The weight function is used to reduce the effect of outliers in the variance covariance matrix of the estimator, creating a robustness against bad leverage outliers. Though there exist many weight functions, for the purpose of this paper two particular weight functions are considered, one based on the Hat matrix $\mathbf{H}$ and one based on the Mahalonobis distance $d(\mathbf{x}_i)$. For the weights based on the Hat matrix, we define $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and the weight function is defined as

$$w(\mathbf{x}_i) = \sqrt{1 - H_{ii}}, \tag{25}$$

where H_{ii} is the i'th diagonal element of $\mathbf{H}$. For the weights based on the Mahalonobis distances, the weight function takes the form of

$$w(\mathbf{x}_i; \tilde{c}) = \begin{cases} \mathbf{x}_i, & \text{if } |x| \leq \tilde{c}, \\ \mathbf{x}_i \frac{\tilde{c}}{d(\mathbf{x}_i)}, & \text{if } |x| > \tilde{c}. \end{cases} \tag{26}$$

In (26), $d(\mathbf{x}_i)$ is the robustly estimated Mahalonobis distance, and $\tilde{c}$ is a tuning constant based on a chosen tolerance level, for example, if a tolerance level $\delta = 0.05$ is chosen, $\tilde{c}$ is equal to the 0.95 quantile of a χ^2 distribution with p degrees of freedom, with p the dimension of $\mathbf{x}_i$.
The Mahalonobis distance has the form of $d(\mathbf{x}, \mu, \Sigma) = \sqrt{(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)}$, where μ and Σ are the robustly estimated mean and covariance matrix of $\mathbf{x}$. For use in this paper, the mean and covariance matrix are robustly estimated using the Minimum Covariance Determinant (MCD) from Rousseeuw (1985).
Similarly to the Wald and Chow Break test, the robustness of the M-estimator follows from its influence function

$$IF(\mathbf{z}, \Psi, F_\theta) = M(\Psi, F_\theta)^{-1}\Psi(\mathbf{z}, \theta), \tag{27}$$

where it is clear that so long as $\Psi(\mathbf{z}, \theta)$ is bounded, the influence function is bounded, and therefore robust. In case an M-estimator is used, $\Psi(\mathbf{z}, \theta)$ is only bounded in the dependent variable dimension, so it is not robust to bad leverage outliers, as previously stated. For the Mallows type M-estimator, the additional weight function bounds $\Psi(\mathbf{z}, \theta)$ in the independent variable dimension, making the estimator robust to both vertical and bad leverage outliers. For

this reason the robust structural break test is constructed using a Mallows type M-estimator. In the simulation study four different robust structural break tests will be considered;

- Mallows type M-estimator of (3) with a Huber down-weighing function and a weight function based on the Hat matrix $\mathbf{H}$,
- Mallows type M-estimator of (3) with a Huber down-weighing function and a weight function based on the Mahalonobis distance,
- Mallows type M-estimator of (3) with a Tukey bisquare down-weighing function and a weight function based on the Hat matrix $\mathbf{H}$,
- Mallows type M-estimator of (3) with a Tukey bisquare down-weighing function and a weight function based on the Mahalonobis distance.

4 Simulation

In order to check stability of the structural break tests, a sensitivity analysis, power analysis and level analysis will be performed. Data will be generated according to the following three data generating processes;

$$y_i = \beta_0 + \beta_1 x_i + e_i \text{ for } i = 1, ..., N, \tag{28}$$

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + e_i \text{ for } i = 1, ..., n_1, \\ \beta_2 + \beta_3 x_i + e_i \text{ for } i = n_1 + 1, ..., N, \end{cases} \tag{29}$$

$$y_i = \begin{cases} \beta_0 + \beta_1 x_{i,1} + e_i \text{ for } i = 1, ..., n_1, \\ \beta_2 + \beta_3 x_{i,2} + e_i \text{ for } i = n_1 + 1, ..., N, \end{cases} \tag{30}$$

where $n_1 + n_2 = N$, $\beta_0 = 2$, $\beta_1 = 2$, $\beta_2 = 2$, $\beta_3 = 2.1$, $e_i \sim \mathcal{N}(0, 1)$, $x_i \sim \mathcal{U}_{[-10,10]}$, $x_{i,1} \sim \mathcal{U}_{[-10,0]}$ and $x_{i,2} \sim \mathcal{U}_{[0,10]}$. Data generated with (28) provides a dataset with no structural break and will be called Scenario 1. Data generated by (29) and (30) create datasets with a structural break. Namely, data generated by (29) has a structural break in time and will be called Scenario 2. Data generated by (30) has a structural break in space, as all points where $x \geq 0$ will be subject to different coefficients than points where $x < 0$, and will be called Scenario 3.

These scenarios are used as they encompass all possible structural breaks. In order to analyse the effect of outliers on the structural break diagnostic test in the given structural break scenarios,

we consider the Tukey-Huber contamination model

$$F_\epsilon = (1 - \frac{\epsilon}{\sqrt{N}})F + \frac{\epsilon}{\sqrt{N}}G \tag{31}$$

with G a point mass contamination $\Delta_{\hat{y},\hat{x}}$ at $\hat{y}$ and $\hat{x}$. Three different outliers will be considered.

- Positive vertical outlier: $\hat{y} = 100$.
- Negative vertical outlier: $\hat{y} = -100$.
- Bad leverage outlier: $\hat{x}_1 = -10$, $\hat{y} = 100$.

The bad leverage outlier is changed for Scenario 3, where the outlier has 2 values; $\hat{x}_1 = -10$, $\hat{y} = 100$ for a point where $x < 0$ and $\hat{x}_1 = 0$, $\hat{y} = 100$ for a point where $x \geq 0$.
All forms of analysis are performed at varying sample sizes N, and at varying break point n_1. In order to ensure that the breakpoint of the Scenario 3 is consistantly at a certain value, n_1 points will be generated with $x_i < 0$ and n_2 points with $x_2 \geq 0$. This means that for all scenarios the breakpoint will be considered at n_1.

4.1 Outlier effect

Before analysis on the diagnostic tests is performed, first the effect of the three different outliers on a simple OLS regression is determined. This is done so the effect that the outliers ultimately have on the diagnostic tests can be more easily explained. In order to check the effect of the outliers, data will be generated using (28), where, similar to Scenario 1, we define $\beta_0 = 2$, $\beta_1 = 2$ and consider three variations of x_i, namely $x_i \sim \mathcal{U}_{[-10,0]}$ and $x_i \sim \mathcal{U}_{[0,10]}$. In these regression one point will be replaced with an outlier, and the value of $\hat{\beta}_0$ and $\hat{\beta}_1$ will be estimated using OLS. These values will than be compared to the original β_0 and β_1 in order to analyse the effect the outliers have.

4.2 Stability Analysis

For the sensitivity analysis we consider the three data scenarios with sample sizes $N = 1000$, $N = 200$ and structural breakpoints $n_1 = \frac{1}{2}N$, $n_1 = \frac{3}{4}N$. In order to test the stability of the diagnostic test one of the generated points is replaced with a single outlier, and the effect these outliers have on the p-value of the diagnostic tests is then analyzed.
The stability analysis is only performed on the Wald and Chow Break test, as the effect on the robust test should be negligible.

4.3 Power Analysis

The power of a diagnostic test is an indication of how correctly the test rejects the null hypothesis of no structural breaks. In order to check the power, Scenario 2 and 3 have their β_3 redefined as $\beta_3 = \lambda\beta_1$. For all scenarios the null hypothesis of the structural break test only holds for $\lambda = 1$, so in order to test the power λ is varied between 0.75 and 1.25. The power is expected to be symmetrical around $\lambda = 1$, and the approach 1 the further away from that value λ gets. For every value of λ the scenario is repeated 1000 times in order to get an accurate value for the power at that value of λ.

4.4 Level Analysis

The level of a diagnostic test indicates how often the test incorrectly rejects a correct null hypothesis, in other words, how often a statistical type II error occurs. In order to analyse the level, Scenario 1 is repeated 1500 times, after which the percentage of type II errors is stored. This process is repeated 500 times in order to create a boxplot of the level of the test.

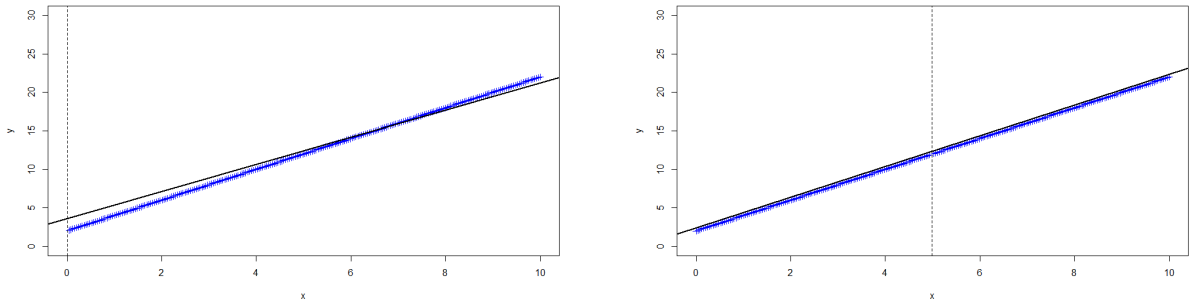
5 Simulation Results

The results of the simulation study will be analysed starting in the optimal case for each scenario, with $N = 1000$ and the structural breakpoint $n_1 = \frac{1}{2}N$. In order to keep the section clear, all results for other values of N and n_1 will be compared to the ideal case. This means that not all results will be presented in this section, and only notable differences shall be discussed. All results for the simulation tests can be found in Appendix A.1 and A.2.

5.1 Outlier effect

The analysis of the outlier effect on OLS will be done for $N = 500$ points, in order to create an idea of the effect of outliers. The effect of outliers for different N will be extrapolated from these results. Figure A.1 shows the effect the outliers have on both β_0 and β_1 for three distributions of x_i. Since these figures are difficult to interpret, a more easily interpretable analysis will be provided. It is important to note that all effects described below rely on the distribution of x_i being symmetrical, i.e Uniform or Normally distributed. Should x_i not be symmetrical, the conclusions drawn will not hold in most cases.

The effect of outlier can be described as a weight that pulls the regression line towards it. In the case of our simple regression line $y_i = \beta_0 + \beta_1 x_i$, the effect of the outlier will be described in terms of changes on β_0 and β_1, but it is worth to note that if the regression line contained



(a) (b)

Figure 5.1: The figure shows a regression line for data generated using the data generating process $y = 2 + 2x$ with an outlier placed at $y = 100$ at, (a) $x = 0$, (b) $x = 5$. The location of the outlier is marked by the vertical dashed line.

k independent variables, the effect on $\beta_2, \cdots, \beta_k$ is similar to the effect on β_1. The effect of a positive vertical and negative vertical outlier on a regression line is always opposite. A positive vertical outlier always move the regression line up, which means that in most cases β_0^*, where the $*$ superscript indicates contamination, will be larger than the original β_0. On the other hand, the effect of a positive vertical outlier on β_1^* depends on the position of the outlier with respect to the range of the dependent variable. If the outlier is positioned in the center of the range, the effect on β_1^* will be negligible. If the outlier is located on the left hand side of the center, a positive vertical outlier will reduce β_1^*, and a negative vertical outlier will increase β_1. Figure 5.1a aims to show this effect, whereas figure 5.1b shows the effect of a single outlier in the center of the independent variable range, which simply increases β_0.

5.2 Results for Wald and Chow Break test

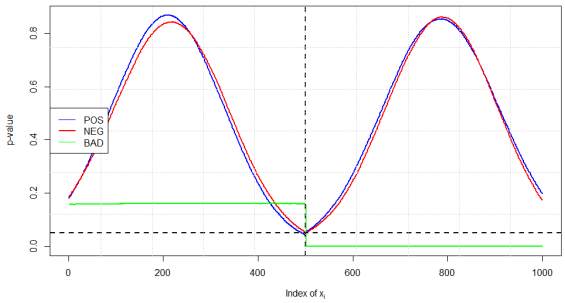
5.2.1 Evaluation of the sensitivity

First the sensitivity of the Chow Break test is analysed. Figure 5.2a shows that for neither the defined positive or negative outliers the p-value of the test falls below the level $\alpha = 0.05$. It can be seen that at n_1 the effect of the outlier is greatest in case of the vertical outliers, and had the outlier been larger, the p-value would have fallen below 0.05. The effect of the bad leverage outlier is always consistent within a dataset, and in this case, if the bad leverage outlier is part of second dataset, $x_{n_1+1}, \cdots, x_N$, the Chow Break test will always reject the null hypothesis. Figure 5.2b shows that the Wald test is slightly more robust to singular vertical outliers as the p-value does not even come close to 0.05. Similar to the Chow test, the Wald test does fail when the bad leverage outlier is in $x_{n_1+1}, \cdots, x_N$. It should be noted that for Scenario 1, the ordering of the data creates a very strong effect of the bad leverage outlier. Data from $x_{n_1+1}, \cdots, x_N$

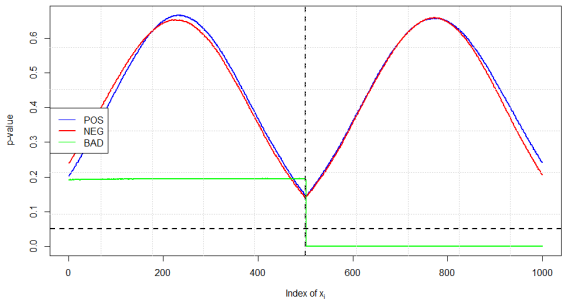
ranges from approximately $[0, 10]$, the bad leverage outlier located at $(\hat{x}_1, \hat{y}) = (-10, 100)$ will have an exacerbated effect on the regression of that data. In other words, the ordering of the data creates the worst possible results both the Wald and Chow Break test could have in Scenario 1.

For Scenario 2 the sensitivity of the Chow Break and Wald test, in Figures 5.2c and 5.2d respectively, is nearly identical, so only Figure 5.2c will be discussed. Before and after the structural break at n_1 the effect of the positive and negative vertical outlier is mirrored. For the structural break that occurs, the intercept is left the same, $\beta_0 = \beta_2$, but the slope of the regression increases, $\beta_1 < \beta_3$, which means that if β_1 is increased, the p-value increases, and if β_3 is decreased, the p-value increases, and visa versa. From Figure 5.2c it can be seen that for $i < n_1/2$, the negative outlier causes a p-value above 0.05, incorrectly accepting the null of no structural break. This is caused by the negative outlier increasing the value of β_1 to be closer to that of β_3. The positive outlier has the same effect for $n_1/2 < i < n_1$. This effect is in line with the outlier effect discussed in 5.1. After n_1, the effect the outlier has to have to increase the p-value above 0.05 is opposite that of before. This is why for $n_1 < i < n_1 + n_2/2$ the positive outlier increases the p-value of the test above the 0.05 level and for $i > n_1 + n_2/2$ the negative outlier has this effect. For values close to $i = n_1/2$ and $i = n_1 + n_2/2$, the p-value is situated below 0.05, since the values of β_1 and β_3 are not significantly affected. The bad leverage outlier decreases the value of β_1 for $i < n_1$, and therefore decreases the p-value below the level. For $i > n_1$ the bad leverage outlier similar decreases the value of β_3 to be closer to β_1, increasing the p-value above the level. These effects are in line with what was discussed in 5.1. What is of note is that the effect the outliers have on β_0 and β_2 is seemingly insignificant.

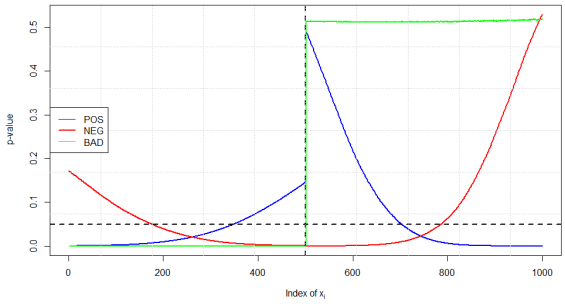
In Scenario 3 the effect of outliers is greater, due to the limited range of x_i before and after $i = n_1$. Figures 5.2e and 5.2f show that the outliers more significantly affect the Chow Break test than Wald test, especially when considering positive vertical outliers. The peaks in the Figures indicate that a complex interaction between the outlier and its effect on both the intercept and slope of the regressions. The peaks of the positive vertical outlier occur when they have no effect on β_0 and β_2. The peaks of the negative vertical outlier occur at a less clearly defined moment. Without diving into the specific values of the parameters, the reason positive outliers have a smaller effect of the p-value is because both β_1 and β_3 are positive values. Reasoning for why the bad leverage outliers have the effect that they do on the p-value is equivalent to the reasoning for their effect in Scenario 2. The main difference between the Chow Break and Wald



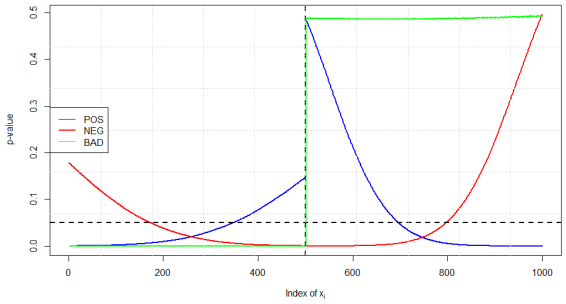
(a) Chow Break Scenario 1



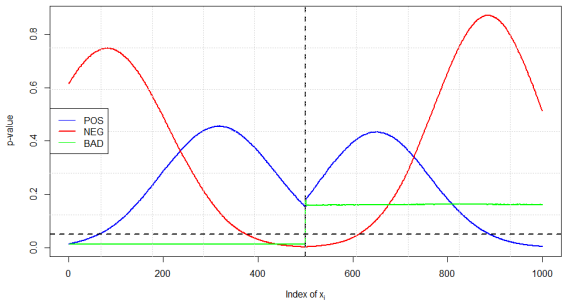
(b) Wald Scenario 1



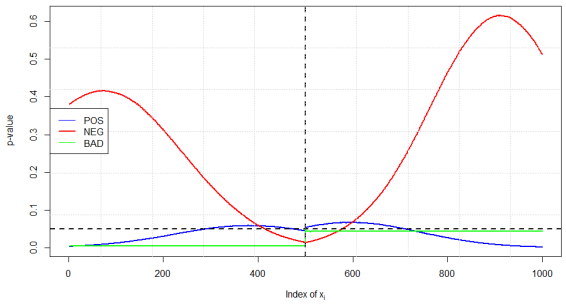
(c) Chow Break Scenario 2



(d) Wald Scenario 2



(e) Chow Break Scenario 3

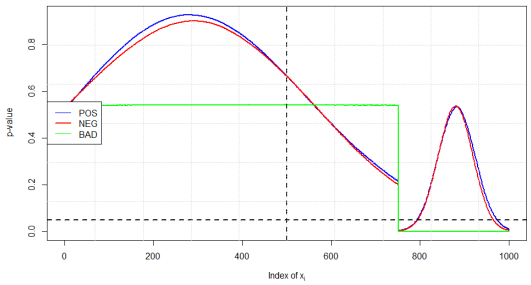


(f) Wald Scenario 3

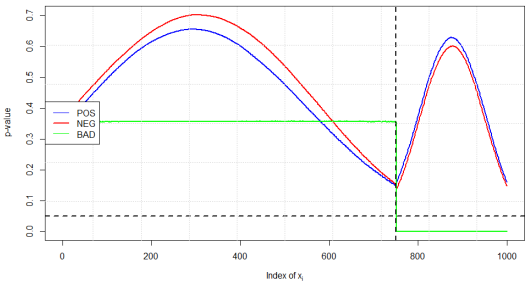
Figure 5.2: Sensitivity analysis of the p-value of the Chow Break test, (a),(c),(e), and the Wald test, (b),(d),(f), for the three data generating scenarios with $N = 1000$. The x-axis indicates the index of the point that is made an outlier. The y-axis gives the average p-value of 100 test statistics. For Scenario 1, x_i is ordered in increasing order so $x_1 < \cdots < x_{n_1} < x_{n_1+1} < \cdots < x_N$. For Scenario 2, x_i is ordered in increasing order so $x_1 < x_2 < \cdots < x_{n_1}$ and $x_{n_1+1} < \cdots < x_N$, in both cases x_i ranges from $[-10, 10]$. For Scenario 3, x_i is ordered in increasing order so $x_1 < x_2 < \cdots < x_{n_1}$ ranging from $[-10, 0]$ and $x_{n_1+1} < \cdots < x_N$ ranging from $[0, 10]$. The horizontal dotted line indicates the level of the test $\alpha = 0.05$ and the vertical dotted line indicates the structural break point $n_1 = 500$. BAD indicates that the outlier used was a bad leverage outlier. POS indicates that the outlier used was a positive vertical outlier. NEG indicates that the outlier used was a negative vertical outlier.

test from Figures 5.2e and 5.2f is that the Wald test is less effected by the outliers, and in the case of positive outliers, is less likely to provide a false positive result.

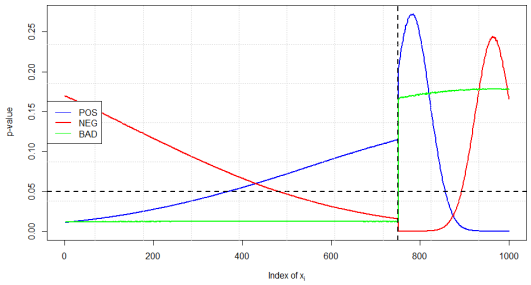
This difference in effect is caused by the main difference between the Wald and Chow Break



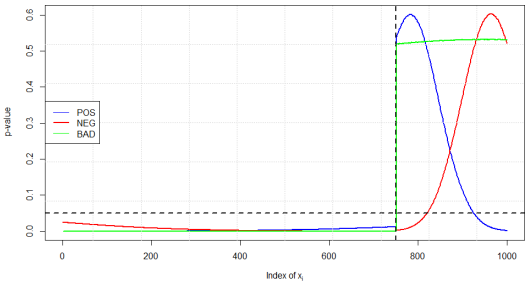
(a) Chow Break Scenario 1



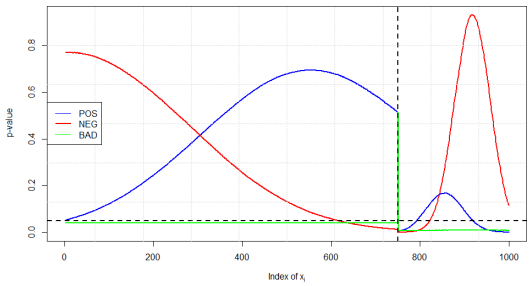
(b) Wald Scenario 1



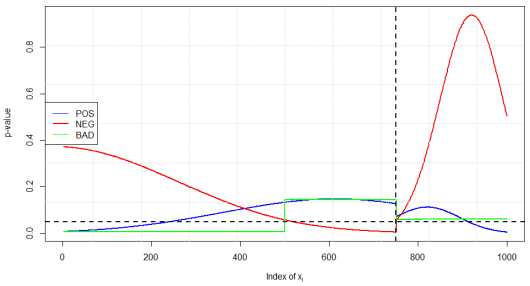
(c) Chow Break Scenario 2



(d) Wald Scenario 2



(e) Chow Break Scenario 3



(f) Wald Scenario 3

Figure 5.3: Sensitivity analysis of the p-value of the Chow Break test, (a),(c),(e), and the Wald test, (b),(d),(f), for the three data generating scenarios with $N = 1000$. The x-axis indicates the index of the point that is made an outlier. The y-axis gives the average p-value of 100 test statistics. For Scenario 1, x_i is ordered in increasing order so $x_1 < \cdots < x_{n_1} < x_{n_1+1} < \cdots < x_N$. For Scenario 2, x_i is ordered in increasing order so $x_1 < x_2 < \cdots < x_{n_1}$ and $x_{n_1+1} < \cdots < x_N$, in both cases x_i ranges from $[-10, 10]$. For Scenario 3, x_i is ordered in increasing order so $x_1 < x_2 < \cdots < x_{n_1}$ ranging from $[-10, 0]$ and $x_{n_1+1} < \cdots < x_N$ ranging from $[0, 10]$. The horizontal dotted line indicates the level of the test $\alpha = 0.05$ and the vertical dotted line indicates the structural break point $n_1 = 750$. BAD indicates that the outlier used was a bad leverage outlier. POS indicates that the outlier used was a positive vertical outlier. NEG indicates that the outlier used was a negative vertical outlier.

test. In case of the Chow Break test, a singular outlier at $i < n_1$ effects two regressions, the regression over the segment of data, i.e $x_1, \cdots, x_{n_1}$, and the regression over the entire dataset $x_1, \cdots, x_N$, the regression on $x_{n_1+1}, \cdots, x_N$ is unaffected. The Wald test simply looks at the difference between β_0 and β_2, and β_1 and β_3, and based on these differences assesses the presence of a structural break. The Chow Break test on the hand compares the sum of squared

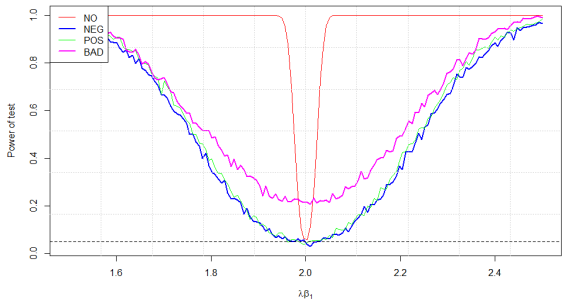
residuals of the regressions. In situations where the Wald test has a p-value below 0.05, and the Chow Break test has a value above 0.05, it is found that the **SSR** is also effected significantly in regression on the whole dataset. In Figure 5.3 the outlier effect is shown for $N = 1000$ and $n_1 = \frac{3N}{4} = 750$. When comparing the results from Figure 5.2 with those where the structural breakpoint is changed to $n_1 = 3 * N/4 = 750$, it is most notable that the effect of the outliers decrease before the structural breakpoint and increase after, when considering the Chow Break test. Most conclusions drawn for $n_1 = N/2$ can be drawn for $n_1 = \frac{3N}{4}$, but in case of Scenario 3, the Chow Break test incorrectly accepts the null hypothesis more often if the outlier lies before the structural breakpoint, as can clearly be seen in Figure 5.3e. For the Wald test the effect of the outlier is drastically reduced if found before the structural breakpoint, most notably so in Scenario 2, where the Wald test is seemingly robust against the singular outlier, as can be seen in Figure 5.3d. This effect is caused by the increase in sample size before the structural break. As the sample size increases, the effect of the outlier on the data decreases, which is why the effect after n_1 is larger for the Wald test. For the Chow Break test, this effect is not as large, as the effect of the outlier on the parameters β_0 and β_1, might decrease, the effect of the outlier on the **SSR** does not.

The effect of outliers when decreasing the sample size is simply increased when compared to the effect visible in Figures 5.2 and 5.3.

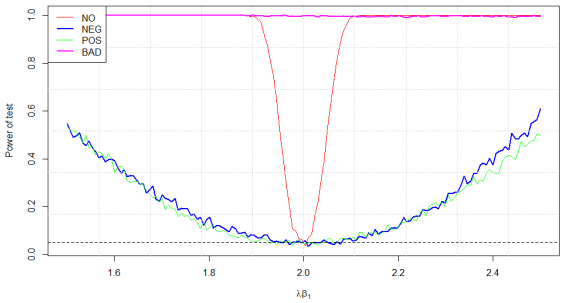
5.2.2 Evaluation of the power

Analysis of the power curves of the Chow Break and Wald test is done for $N = 1000$ in figure 5.4. The power of the tests indicates how often the test rejects the null hypothesis of no structural breaks. This means that if the power is 1, the diagnostic test always rejects the null hypothesis and finds that a structural break is present.

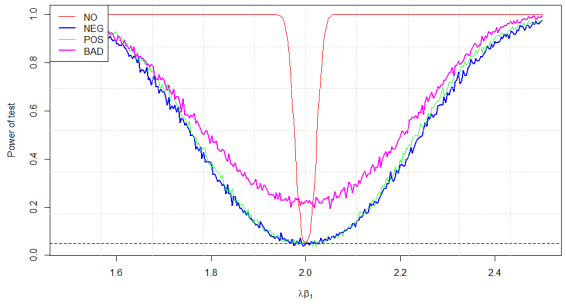
Figures 5.4a and 5.4b show that the power of the Chow Break test is greater when a time-wise structural break is present, as the power reaches 1 around $\lambda\beta_1 = 1.95 \vee 2.05$ in figure 5.4a and only at $\lambda\beta_1 = 1.9 \vee 2.1$ in figure 5.4b. For Scenario 2 the power of the bad leverage contamination fails around $\lambda\beta_1 = 2$, as it does not approach the level of the test $\alpha = 0.05$. The scenario with positive and negative outliers do approach the level at $\lambda\beta_1 = 2$, but the power takes a long time to approach 1, not even reaching it as $\lambda\beta_1 = 1.5 \vee 2.5$. This holds for both the Wald test and the Chow Break test. In Scenario 3, for both the Wald and Chow Break test, the bad leverage contamination scenario causes a complete failure of the test statistic, as the null hypothesis of no structural breaks is always rejected, as can be seen in figures 5.4b and 5.4d. Additionally, the



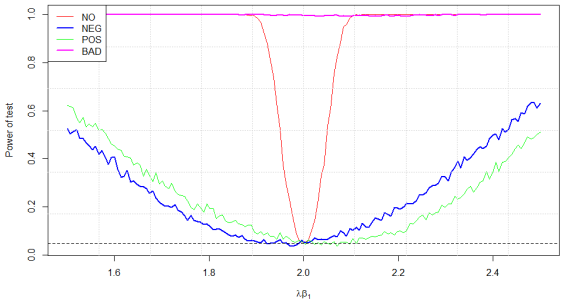
(a) Chow Break Scenario 2



(b) Chow Break Scenario 3



(c) Wald Scenario 2



(d) Wald Scenario 3

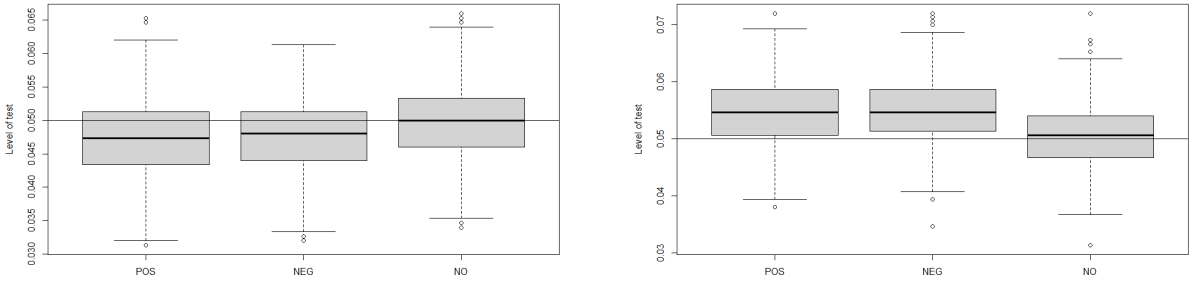
Figure 5.4: Power analysis of the Chow Break test, (a),(b), and the Wald test, (c),(d), for the two structural break data generating scenarios with $N = 1000$ and structural break point $n_1 = 500$. The x-axis indicates the value of β_3 after the break, $\lambda\beta_1$. The y-axis gives the average power of 1000 replications. The horizontal dotted line indicates the level of the test $\alpha = 0.05$. The contamination level was set to $\epsilon = 0.01$. BAD indicates that the outliers used were bad leverage outliers. POS indicates that the outliers used were positive vertical outliers. NEG indicates that the outliers used were negative vertical outliers. NO indicates the data was uncontaminated.

power of both tests is weaker than it is in Scenario 2. In case of vertical contamination, either positive or negative, the power barely approaches 0.6 as $\lambda\beta_1$ gets to $1.5 \vee 2.5$. In Scenario 3, the Wald test is slightly less powerful than the Chow Break test in the case of no contamination. In either Scenario 2 or 3, the Chow Break and Wald test have similar weaknesses to outlier contamination. Both completely fail in the presence of bad leverage outliers and have strongly reduced power in the presence of vertical outliers.

5.2.3 Evaluation of the level

The level of the diagnostic tests analyses the presence of false negative results, meaning that for a level of 0.05, a diagnostic tests rejects the null hypothesis 5% of the time, even though it should not. The level of the bad leverage contamination scenario is not shown in Figure 5.5, as this is always 1 with how the data is ordered. This can be verified by looking at the power of in Figures 5.4b and5.4d. The level of the test falls below 0.05 for the Chow Break test. This is

caused by the outliers increasing the SSR of all 3 regressions performed to such an extend that the consistency of the outliers results in the Chow Break test over accepting the null hypothesis. The Wald test on the other hand over rejects the null for positive and negative outliers, as the average level is above 0.05.



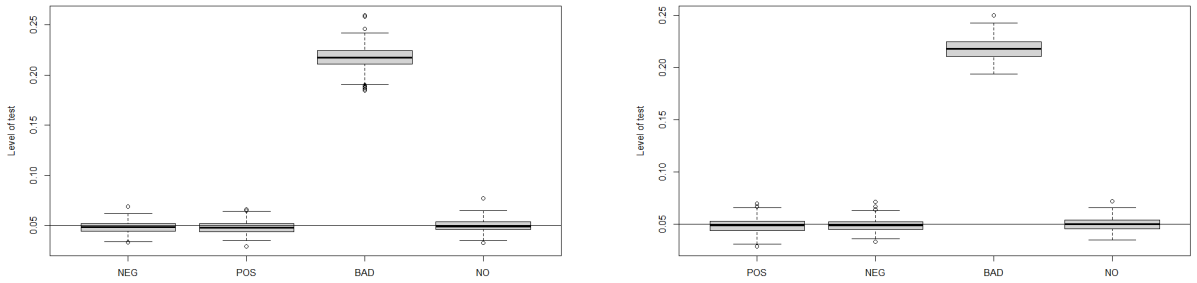
(a) Chow Break

(b) Wald

Figure 5.5: Level analysis of the Chow Break test, (a), and the Wald test, (b), for the two structural break data generating scenarios with $N = 1000$ and structural break point $n_1 = 500$. The x-axis indicates the type of outlier contamination for which the level is being considered. The contamination level was set to $\epsilon = 0.01$. For these levels data was ordered along the x-axis, namely $x_1 < \cdots < x_N$

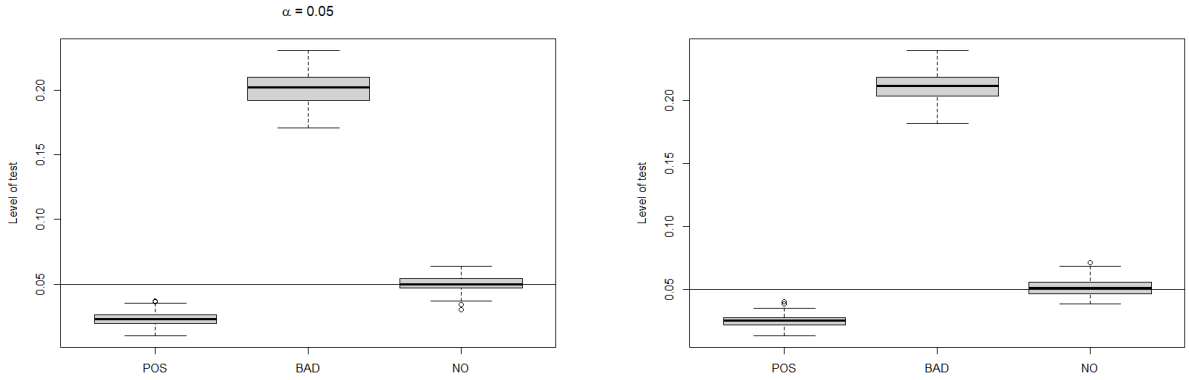
Neither test is robust to bad leverage outliers, but they both are relatively robust to vertical outliers, as their median level sits at approximately 0.047 and 0.055 for the Chow Break and Wald test respectively. The robustness of the level to vertical outliers however only applies when every point has an equal chance of being an outlier, and can be expected that if the outliers are grouped in some manner, the level of the Chow Break and Wald test will fail to be around 0.05. If the data is ordered randomly when the level is tested, once again we see that both the Chow Break and Wald test are robust to vertical outliers with their median levels around 0.05. The level of both tests is still not robust against bad leverage outliers, with the median level close to 0.23 as can be seen in Figure 5.6.

If the suspected breakpoint is changed, the level of the Chow Break and Wald test does not change significantly. When the sample size is lowered, both the Chow Break and Wald test lose their apparent robust to vertical outliers. This can be seen in Figure 5.7. The results for the structural breakpoint $n_1 = \frac{3N}{4} = 150$ produces results similar to those seen in Figure 5.7.



(a) Chow Break (b) Wald

Figure 5.6: Level analysis of the Chow Break test, (a), and the Wald test, (b), for the two structural break data generating scenarios with $N = 1000$ and structural break point $n_1 = 500$. The x-axis indicates the type of outlier contamination for which the level is being considered. The contamination level was set to $\epsilon = 0.01$. For these levels data was ordered randomly



(a) Chow Break (b) Wald

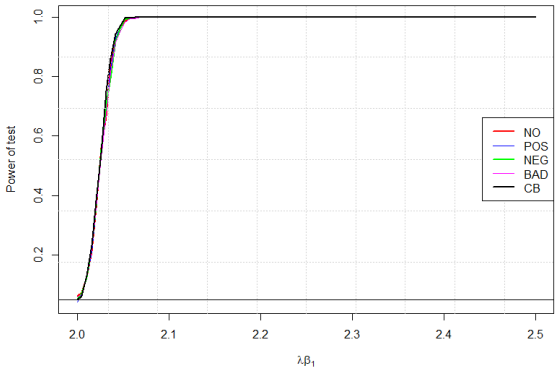
Figure 5.7: Level analysis of the Chow Break test, (a), and the Wald test, (b), for the two structural break data generating scenarios with $N = 200$ and structural break point $n_1 = 100$. The x-axis indicates the type of outlier contamination for which the level is being considered. The contamination level was set to $\epsilon = 0.01$. For these levels data was ordered randomly

5.3 Results for Robust Structural Break test

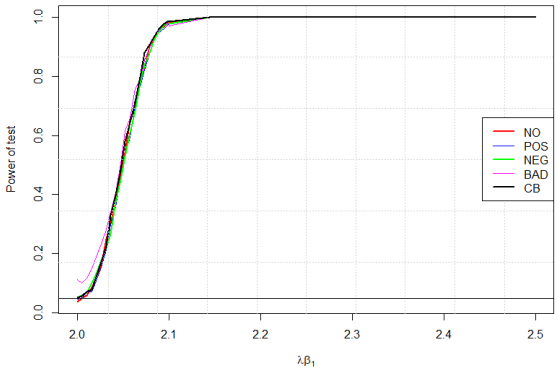
5.3.1 Evaluation of the power

The power of the Robust Structural Break test is analysed using data from Figures 5.8 and 5.9. Since the power of diagnostic tests was previously symmetrical around $\lambda\beta_1 = 2$, only $2 \leq \lambda\beta_1 \leq 2.5$ is considered.

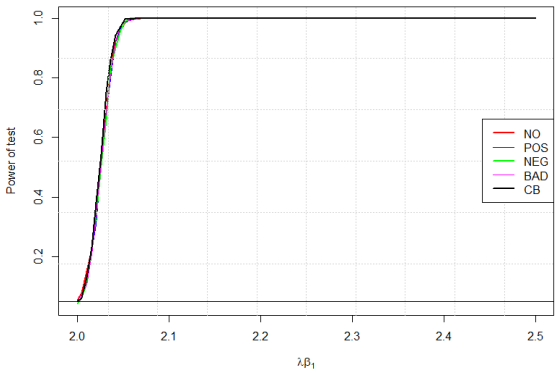
The power of all four specified robust structural break tests is equivalent in both Scenario 2 and Scenario 3. The power of the robust tests is equivalent to that of the Wald and Chow Break in case no data contamination is present. The power of the robust tests is significantly better in cases of any contamination. Most notably, bad leverage outliers no longer break the diagnostic tests. The data from Figures 5.8 and 5.9 was generated using a specified range of $\lambda\beta_1$, namely



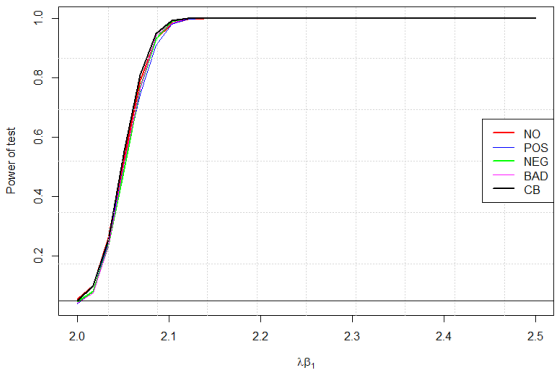
(a) Huber Hat Scenario 2



(b) Huber Hat Scenario 3



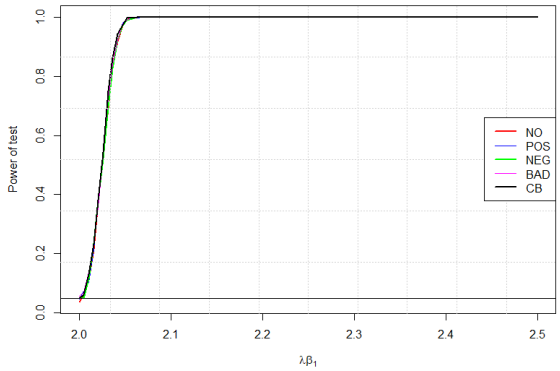
(c) Huber Mah Scenario 2



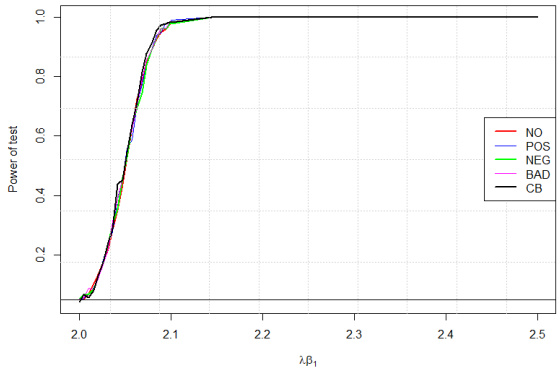
(d) Huber Mah Scenario 3

Figure 5.8: Power analysis of the robust structural break test using the Huber down-weighing function for the two structural break scenarios. The sample size of the model used was $N = 1000$, and the breakpoint was at $n_1 = 500$. Hat indicates the Hat matrix based weight function was used. Mah indicates the Mahalonobis distance based weight function was used. On the x-axis the points of data were $\lambda\beta_1 \in \{2, 2.005, \cdots, 2.1, 2.15, \cdots, 2.5\}$. In the legend, NO indicates the no contamination scenario, POS the positive vertical outlier contamination, NEG the negative vertical outlier contamination, BAD the bad leverage contamination and CB indicates the performance of the Chow Break test in the no contamination scenario. The contamination level was set to $\epsilon = 0.01$.

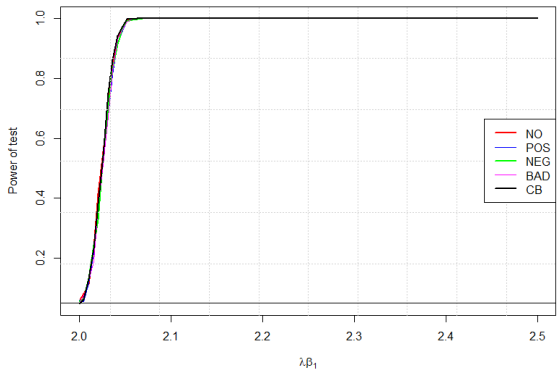
$\lambda\beta_1 \in \{2, 2.005, \cdots, 2.1, 2.15, \cdots, 2.5\}$, as the behavior of the robust test for values of $\lambda\beta_1$ close to 2 is of most interest. It can be seen in Figures 5.8 and 5.9 that even for smaller values of λ, the robust estimators remain consistently robust. From the power analysis, no robust structural break test is clearly preferred for large sample size and centered structural breakpoint, nor is there a significant loss of power when compared to the Chow Break test in a no contamination scenario. In order to analyse whether this is truly the case, the power of the robust Wald test using the Tukey-bisquare down-weighing function and the weights based on the Hat matrix is plotted for $\lambda\beta_1 \in \{2, 2.001, 2.002, \cdots, 2.1\}$ in Figure 5.10.



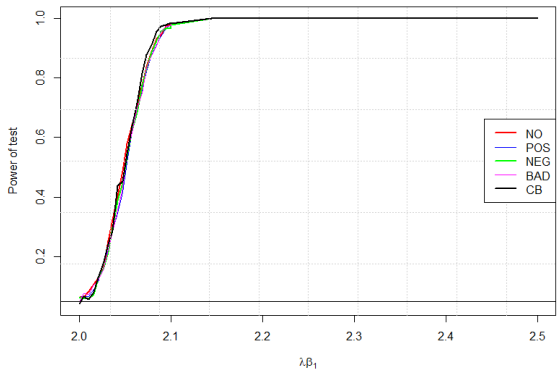
(a) Tukey Hat Scenario 2



(b) Tukey Hat Scenario 3



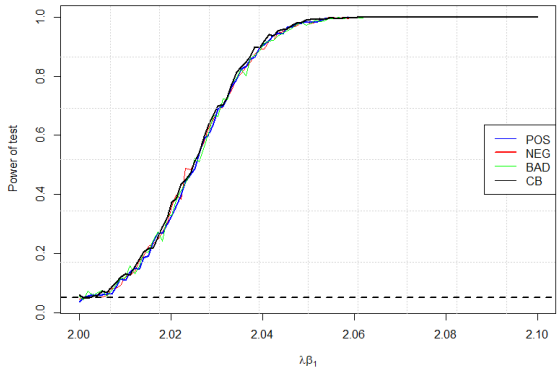
(c) Tukey Mah Scenario 2



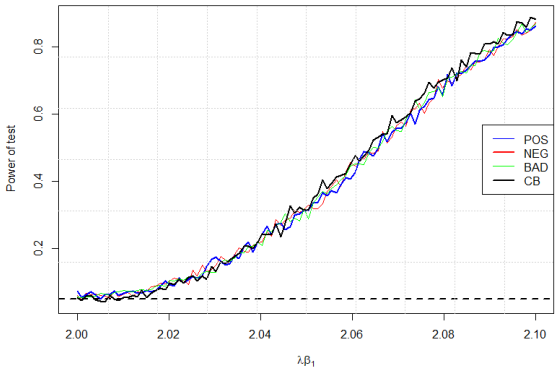
(d) Tukey Mah Scenario 3

Figure 5.9: Power analysis of the robust structural break test using the Tukey bisquare down-weighing function for the two structural break scenarios. The sample size of the model used was $N = 1000$, and the breakpoint was at $n_1 = 500$. Hat indicates the Hat matrix based weight function was used. Mah indicates the Mahalonobis distance based weight function was used. On the x-axis the points of data were $\lambda\beta_1 \in \{2, 2.005, \cdots, 2.1, 2.15, \cdots, 2.5\}$. In the legend, NO indicates the no contamination scenario, POS the positive vertical outlier contamination, NEG the negative vertical outlier contamination, BAD the bad leverage contamination and CB indicates the performance of the Chow Break test in the no contamination scenario. The contamination level was set to $\epsilon = 0.01$.

Figure 5.10 shows that the power of the robust tests is equivalent to that of the standard Chow Break test, but when the structural breakpoint shifts, the power of the robust versions of the Wald test greatly decrease when compared to the Chow Break test. In further analysis, a smaller sample size and different structural breakpoint are considered. Figure 5.11 shows the results of the power analysis of the robust Wald test using the Tukey bisquare down-weighing function and the weights based on the Hat matrix. Conclusions drawn from figure 5.11 can also be drawn for the other robust Wald tests, as their performance is nearly identical, as can be seen in Appendix A.2. It should be noted that for more in-dept analysis, the negative vertical outlier contamination was no longer considered, as its effect on the diagnostic test was found to



(a) Structural Break at $\frac{N}{2}$



(b) Structural Break at $\frac{3N}{4}$

Figure 5.10: Power analysis of the robust structural break test using the Tukey-bisquare down-weighing function and the weights based on the Hat matrix, with the power determined for $\lambda\beta_1 \in \{2, 2.001, 2.002, \cdots, 2.1\}$. Data was generated using Scenario 2. The contamination level was set to $\epsilon = 0.01$.

be equivalent to that of the positive vertical outlier contamination. From figure 5.11 it can be seen that shifting the structural breakpoint decreases the power as the difference in sample size before and after the break increases. The decrease in power does not come from the increased

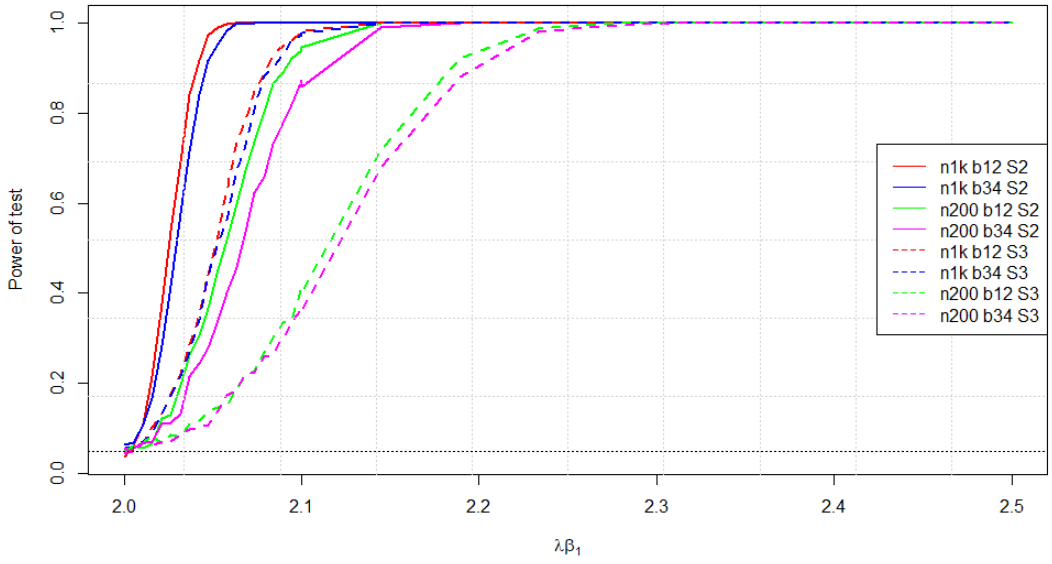
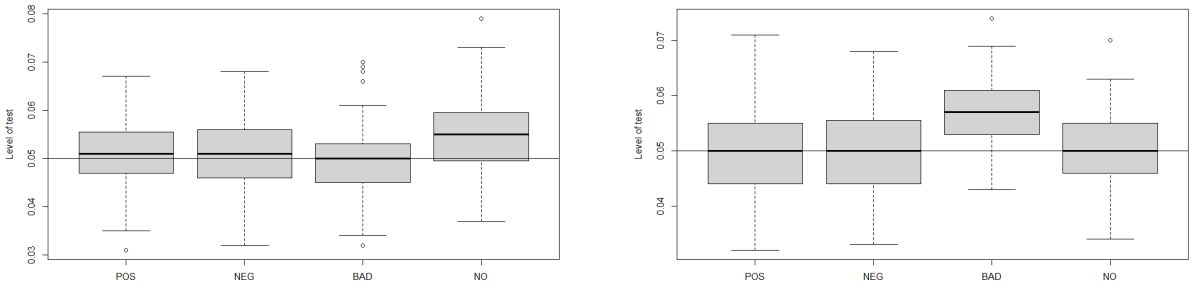


Figure 5.11: Power analysis of the robust structural break test using the Tukey bisquare down-weighing function and the weights base on the Hat matrix. The sample size of the model used is indicated by $n1k \rightarrow N = 1000$ and $n200 \rightarrow N = 200$, and the breakpoint is indicated by $b12 \rightarrow n_1 = N/2$ and $b34 \rightarrow n_1 = 3N/4$. $S2$ indicates data was generated using Scenario 2. $S3$ indicates data was generated using Scenario 3. On the x-axis the points of data were $\lambda\beta_1 \in \{2, 2.005, \cdots, 2.1, 2.15, \cdots, 2.5\}$. For this power analysis no contamination was present. The contamination level was set to $\epsilon = 0.01$.

difference between n_1 and n_2 in the test, but rather the decrease in n_2, as is apparent by the decrease in power between the tests using sample size $N = 1000$ and $N = 200$. The decrease in power is more pronounced for the robust tests performed in the space-wise break scenario, as their power reaches 1 as $\lambda\beta_1 \approx 2.15$ for $N = 1000$ and $\lambda\beta_1 \approx 2.25$ for $N = 200$. For the time-wise break scenario the power reaches 1 as $\lambda\beta_1 \approx 2.075$ and $\lambda\beta_1 \approx 2.15$ for $N = 1000$ and $N = 200$ respectively. The power of the robust tests when compared with the Chow break or Wald test is equivalent in both the case that the structural breakpoint is at $\frac{N}{2}$ or $\frac{3*N}{4}$, similar to behavior seen in Figure 5.10. This means that regardless of sample size or structural breakpoint, the robust versions of the Wald test remain powerful.

5.3.2 Evaluation of the level

The level of the robust structural break tests is anylysed using the boxplots from figures 5.12 and 5.13. These boxplots were constructed of 100 different level values produced by 1000 replications. The contamination level was set to $\epsilon = 0.01$ and the data was randomly ordered. Positive and negative vertical contamination was considered, along with bad leverage contamination and a no contamination case.

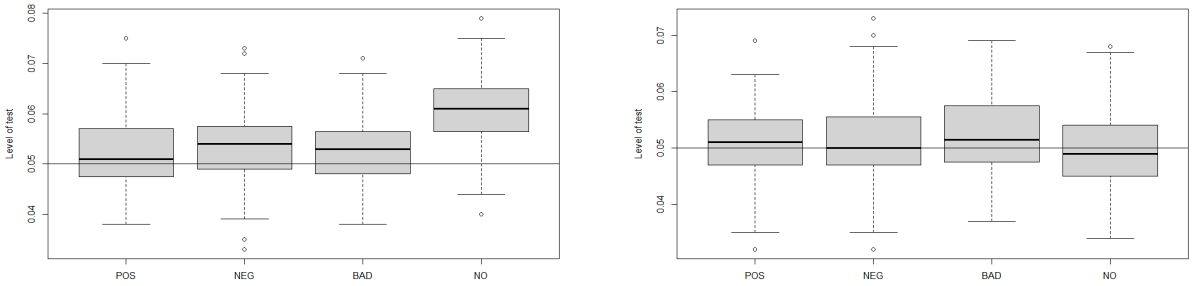


(a) Mahalonobis (b) Hat

Figure 5.12: Level analysis of the Robust Structural Break test using the Huber down-weighing function and the weight function based on the; (a) robust Mahalonobis distance, (b) Hat matrix, for the two structural break data generating scenarios with $N = 1000$ and structural break point $n_1 = 500$. The x-axis indicates the type of outlier contamination for which the level is being considered. The contamination level was set to $\epsilon = 0.01$.

The level of the robust test using the Huber down-weighing function produces levels close to the nominal level of $\alpha = 0.05$. A notable outlier is the bad leverage contamination scenario when using the robust Hat matrix weights, where the level lies close to 0.06. Additionally, in the no contamination case, the robust test using the Mahalonobis weights has a level that is above the nominal level.

For the robust test using the Tukey down-weighing function once again the no contamination



(a) Mahalonobis (b) Hat

Figure 5.13: Level analysis of the Robust Structural Break test using the Tukey bisquare down-weighing function and the weight function based on the; (a) robust Mahalonobis distance, (b) Hat matrix, for the two structural break data generating scenarios with $N = 1000$ and structural break point $n_1 = 500$. The x-axis indicates the type of outlier contamination for which the level is being considered. The contamination level was set to $\epsilon = 0.01$. For these levels data was ordered randomly

scenario with the Mahalonobis weights has a level closer to 0.06 than the nominal level, as can be seen in figure 5.13b. From the levels of the tests in figures 5.12 and 5.13, a preference for the robust test with the Tukey bisquare down-weighing function and Hat matrix weights is found. This particular robust test has a level around the nomimal $\alpha = 0.05$ in all contamination scenarios.

Table 5.1 shows the levels for all considered structural breakpoints and sample sizes, similar to the in-depth power analysis, the effect of negative vertical outliers is not considered as it is similar to that of positive vertical outliers. From Table 5.1 it can be seen that the robust versions of the Wald test maintain good performance at the structural breakpoint $n_1 = \frac{3N}{4}$ and at the small sample size $N = 200$. For the small sample size, the level of the tests is always above the nominal level of $\alpha = 0.05$. The level is not significantly above the nominal level, but this behavior could be an indicator that at smaller sample sizes a small sample correction would be required. From Table 5.1 we once again draw the conclusion that robust Wald test using the Tukey bisquare down-weighing function and the weights based on the Hat matrix performs best when considering the level, but the other robust versions of the Wald test also perform well.

Table 5.1: The level of the Chow Break (CB) and the four robust versions of the Wald test. H indicates that the weights based on the Hat matrix were used in the test, and M indicates that the weights based on the robust Mahalonobis distance were used. The nominal level of the test was set to $\alpha = 0.05$, and the noted level is the result of 100.000 replications. For every scenario of N, n_1 and contamination type, the level closest to the nominal level is emboldened.

N	Breakpoint n_1	Contamination	CB	H Huber	M Huber	H Tukey	M Tukey
1000	$\frac{N}{2} = 500$	None	**0.050**	**0.050**	0.055	0.049	0.06
		Positive vertical	0.048	**0.050**	**0.050**	0.051	0.053
		Bad leverage	0.217	0.057	**0.050**	0.052	0.053
	$\frac{3N}{4} = 750$	None	**0.050**	**0.050**	0.052	0.051	0.052
		Positive vertical	0.047	**0.051**	**0.051**	**0.051**	**0.051**
		Bad leverage	0.217	0.059	0.058	0.053	**0.052**
200	$\frac{N}{2} = 100$	None	**0.049**	0.052	0.053	0.053	0.057
		Positive vertical	0.024	0.053	0.054	**0.051**	0.058
		Bad leverage	0.202	0.060	0.060	**0.053**	0.059
	$\frac{3N}{4} = 150$	None	**0.050**	0.057	0.058	0.057	0.057
		Positive vertical	0.072	0.057	0.057	**0.051**	0.057
		Bad leverage	0.149	0.062	0.064	**0.058**	0.06

6 Real Data Example

In this section we compare the performance of the Chow Break test and the robust versions of the Wald test described in Section 3 in a real data application. The data is available through the *strucchange* package in R (Zeileis, Leisch, Hornik & Kleiber, 2002). The data has previously been analysed by Hansen (2001) and Zeileis, Leisch, Kleiber and Hornik (2005). In previous analyses the presence of outliers was not considered. The presence of outliers, or the lack thereof, will be verified. Once this is done, the data will be contaminated with additional, realistic outliers, and the effect on the analysis will be discussed.

The data that used in the papers by Hansen (2001) and Zeileis et al. (2005) concerns the US labor productivity in the manufacturing/durables sector from February 1947 to April 2001. This is a monthly time series of the average weekly labor hours in that month. The data is fitted to simplest dynamic model, namely a first order autoregression of the form:

$$y_t = \alpha + \beta y_{t-1} + e_t, \text{ for } t = 1, ..., 650, \tag{32}$$

with $e_t \sim \mathcal{N}(0, \sigma^2)$. Since the data contains 651 point, y_t will be a vector of 650 points, with y_0 defined as the average weekly labor hours in February 1947. Hansen (2001) states that there was a suspected break in the year of 1973, but using the Chow Break test found the test statistic below the required critical value, whereas performing the test with a suspected break at 1975 yields a result above the critical value. The paper then goes on to describe several causes for

Table 6.1: Tally of types of points in the regression for five structural break scenarios. The dataset contains 650 points of data and the cut-off points for the standardized errors and robust Mahalonobis distance are $|\frac{r_i}{\hat{\sigma}}| = 2.5$ and $d_i = \sqrt{\chi^2_{2,0.975}}$ respectively.

Structural break scenario	Good point	Good leverage	Bad leverage	Vertical outlier
One structural break	505	102	16	27
Two structural breaks	501	101	14	34
Three structural breaks	504	98	15	33
Four structural breaks	501	94	20	35
Five structural breaks	499	96	20	35

this perceived inaccuracy, and describes a method by Bai (1997) to find multiple breaks in a singular dataset. Using this method Zeileis et al. (2002) describe 5 breakpoint scenarios, namely scenarios with 1 to 5 structural breakpoints. The breakpoints for each of these scenarios are as follows:

- December 1981 ($t = 418$),
- July 1965 and April 1991 ($t = 221$ and $t = 530$),
- August 1956, November 1965 and April 1991 ($t = 114$, $t = 225$ and $t = 530$),
- August 1956, July 1965, December 1981 and May 1991 ($t = 114$, $t = 221$, $t = 418$ and $t = 531$),
- August 1956, July 1965, September 1973, December 1981 and May 1991 ($t = 114$, $t = 221$, $t = 319$, $t = 418$ and $t = 531$).

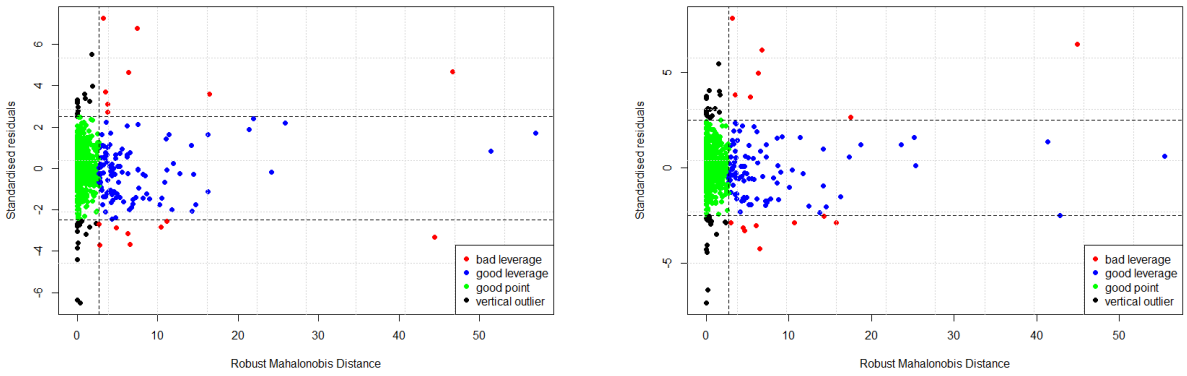
For each of these structural breakpoint scenarios, we will analyse the test values for the Chow (1960) Break test, Wald (1943) test and the four robust versions of the Wald test. Additionally, for each scenario a robust regression diagnostic plot will be made to determine the presence of outliers within the original dataset. A robust regression diagnostic plot has to be made for every scenario as the context of the data changes as the amount of structural breakpoints changes. In order to create robust regression diagnostic plots we consider the robust Mahalonobis distance and the standardized residuals from a high-breakdown regression estimator as per Rousseeuw and van Zomeren (1990). Rousseeuw and van Zomeren (1990) use the least median of squares estimator as their high-breakdown regression estimator to determine the residuals r_i, the Median Absolute Deviation to determine the robust scale estimate $\hat{\sigma}$ and the Minimum Volume Ellipsoid estimator as their method to robustly estimate the Mahalonobis distance d_i. In the robust regression diagnostic plots the procedure of Rousseeuw and van Zomeren (1990) will be followed, except in order to robustly estimate the Mahalonobis distance, the covariance will be

Table 6.2: Test statistic for every break of every break scenario for the Chow Break test, Wald test and the four robust structural break tests. M denotes that the robust Mahanolobis distance was used to determine the weights, and H denotes that the Hat matrix was used to determine the weights. Values that are emboldened fall below the critical value, and hence indicate no structural break. The critical value for the Wald tests is $\chi^2_{2,0.95}$, and for the Chow Break test it is $F(0.95, 2, N-4)$, where N differs per break scenario and break considered.

Break	Chow Break	Wald	M Huber	M Tukey	H Huber	H Tukey
First Break	9.186	22.312	24.682	23.826	26.040	27.690
First Break	7.356	14.441	18.008	17.940	28.009	32.106
Second Break	12.178	25.258	33.157	28.138	31.864	27.410
First Break	4.491	9.014	**0.445**	**0.148**	**1.505**	**1.795**
Second Break	8.627	14.314	15.434	11.894	27.723	32.479
Third Break	12.047	25.037	32.910	27.800	31.636	27.071
First Break	4.487	9.023	**0.476**	**0.098**	**1.535**	**1.842**
Second Break	5.066	9.394	10.447	7.685	17.194	18.975
Third Break	3.834	8.554	**5.307**	**4.686**	6.957	6.327
Fourth Break	7.286	14.772	15.488	13.662	17.824	15.743
First Break	4.487	9.023	**0.476**	**0.098**	**1.535**	**1.841**
Second Break	4.076	9.163	10.900	8.937	16.146	19.176
Third Break	**0.918**	**1.842**	**1.920**	**1.675**	**2.386**	**2.345**
Fourth Break	5.012	10.104	6.824	**5.674**	9.249	8.243
Fifth Break	7.285	14.771	15.488	13.662	17.824	15.743

estimated using the Minimum Covariance Determinant.

The robust regression diagnostic plots require cut-off values for the robust Mahalonobis distance and the standardized residuals in order to identify outliers. The cut-off value for the robust Mahalonobis distance d_i is $\sqrt{\chi^2_{2,0.975}}$, and the cut-off value for the standardized residuals is $|\frac{r_i}{\hat{\sigma}}| < 2.5$, similarly to Rousseeuw and van Zomeren (1990).



(a) One Break Scenario

(b) Three Break Scenario

Figure 6.1: Robust diagnostic plots for the US labor productivity in the manufacturing/durables sector from February 1947 to April 2001, the cut-off points for the standardized errors and robust Mahalonobis distance are $|\frac{r_i}{\hat{\sigma}}| = 2.5$ and $d_i = \sqrt{\chi^2_{2,0.975}}$ respectively.

Figure 6.1 shows the results for the robust diagnostic plot for the scenario with one structural

Table 6.3: Tally of types of points in the regression for five structural break scenarios. The dataset contains 650 points of data of which four are artificially contaminated and the cut-off points for the standardized errors and robust Mahalonobis distance are $|\frac{r_i}{\hat{\sigma}}| = 2.5$ and $d_i = \sqrt{\chi^2_{2,0.975}}$ respectively.

Structural break scenario	Good point	Good leverage	Bad leverage	Vertical outlier
One structural break	497	104	17	32
Two structural breaks	499	100	17	34
Three structural breaks	498	97	18	37
Four structural breaks	494	96	22	38
Five structural breaks	493	98	22	37

break, and the diagnostic plots for the other scenarios can be found in Appendix A.3. The results for these robust diagnostic plots are summarized in Table 6.1. In Table 6.1 it can be seen that for the different breakpoint scenarios the types of points change. Analysis from Zeileis et al. (2002) indicates that the three structural break scenario is the best fit for the data. From Table 6.1 a similar conclusion could be drawn, however, the single break scenario would also be a contender for best scenario as it contains the least outliers. In order to check the validity of the breaks suggested by Zeileis et al. (2002), we check the test statistics of the Chow (1960) Break test, Wald (1943) test, and the four robust versions of the Wald test. Table 6.2 shows that the regular structural break tests find that all breaks all significant except for the break at September 1973 ($t = 319$) in the five structural break scenario. The robust structural break tests however indicate that the suspected break around August 1956 ($t = 114$) is not significant in any case, and the break at December 1981 ($t = 418$) is rejected by some of the robust tests, indicating that there might be a break around this point in time, but not exactly on this point in time. The results from Table 6.2 indicate that, had the breakpoints been estimated robustly, they would have been found at different points in time, as the breakpoint at August 1956 $t = 114$, is never found significant.

Since the considered model in (32) is an AR(1)-model, when artificially contaminating the data, we cannot specifically add a bad leverage outliers to the data. In order to artificially contaminate the data, four points of data are selected, a random point in the ranges $[1, 113]$, $[114, 224]$, $[225, 529]$ and $[530, 650]$, being $75, 129, 460$ and 622 respectively. We chose these ranges in an attempt to break the 'ideal' three structural break scenario provided by Zeileis et al. (2002). In order to contaminate the data, points y_{75}, y_{129}, y_{460} and y_{622} are changed to be twice the maximum of $\{y_1, \cdots, y_{650}\}$. The diagnostic plots for the artificially contaminated data can be found in Appendix A.3, their results are summarized in Table 6.3. Table 6.3 shows that there is a decrease in good points in case the data is artificially contaminated for every structural breakpoint scenario. For the effect on the perceived significance of the breaks we look at the values in Table 6.4 we see

Table 6.4: Test statistic for every break of every break scenario for the Chow Break test, Wald test and the four robust structural break tests. M denotes that the robust Mahanolobis distance was used to determine the weights, and H denotes that the Hat matrix was used to determine the weights. The dataset is artifically contaminated at y_{75}, y_{129}, y_{460} and y_{622}. Values that are emboldened fall below the critical value, and hence indicate no structural break. The critical value for the Wald tests is $\chi^2_{2,0.95}$, and for the Chow Break test it is $F(0.95, 2, N-4)$, where N differs per break scenario and break considered.

Break	Chow Break	Wald	M Huber	M Tukey	H Huber	H Tukey
First Break	3.833	7.637	22.141	20.857	12.660	11.309
First Break	**1.490**	**2.816**	13.893	14.698	12.060	18.879
Second Break	7.711	13.330	31.417	26.085	28.769	26.801
First Break	**0.110**	**0.219**	**0.016**	**0.228**	**0.171**	**4.425**
Second Break	**1.029**	**1.549**	8.510	8.145	11.076	33.952
Third Break	7.587	13.184	31.090	25.729	28.590	26.520
First Break	**0.111**	**0.223**	**0.002**	**0.182**	**0.181**	**4.549**
Second Break	**2.616**	**4.122**	6.628	**5.450**	8.220	20.468
Third Break	**2.654**	**4.339**	**5.823**	**4.676**	**4.088**	8.254
Fourth Break	**1.733**	**3.458**	13.693	11.901	15.254	27.272
First Break	**0.111**	**0.222**	**0.002**	**0.182**	**0.181**	**4.549**
Second Break	**1.450**	**3.514**	6.868	6.410	7.904	19.445
Third Break	**0.918**	**1.842**	**1.920**	**1.675**	**2.386**	**2.345**
Fourth Break	**2.498**	**5.129**	7.194	**5.604**	**5.976**	9.938
Fifth Break	**1.733**	**3.458**	13.693	11.901	15.254	27.272

that the standard tests fail for almost every break in every break scenario. The robust structural break tests show similar results to the uncontaminated case from Table 6.2. In the three break scenario, the significance of the test values from the robust test remain consistent in the presence of the artificial contamination. In the scenarios with four or five breaks, certain test values drop below their critical values, however these values were already close to the critical value in the uncontaminated case. It should be noted that with the addition of the artificial contamination, the suggested breakpoints for each of the scenarios is shifted: (Zeileis et al., 2002)

- December 1991 ($t = 538$),
- June 1959 and December 1981 ($t = 148$ and $t = 418$),
- December 1957, January 1966 and December 1982 ($t = 130$, $t = 227$ and $t = 430$),
- December 1957, January 1966, December 1982 and February 1993 ($t = 130$, $t = 221$, $t = 430$ and $t = 552$),
- December 1957, January 1966, April 1974, December 1982 and February 1993 ($t = 130$, $t = 227$, $t = 326$, $t = 430$ and $t = 552$).

From this analysis we can conclude that the method from Bai (1997) that is adapted by Zeileis et al. (2002) is not robust to outliers, and that a slight change in data can completely invalidate

the conclusions drawn by the non-robust diagnostic tests. The robust structural break tests draw consistent conclusions in the presence of contamination, with only slight derivations. This study of a real data example does show the greatest weakness of this paper, namely that the robust Wald tests are not expanded to a robust supremum Wald test, as this would have proven to be interesting additional analysis for this dataset.

7 Conclusion

In this paper the robustness properties of structural break tests were investigated. Two standard structural break tests were analysed; The Wald (1943) test and the Chow (1960) Break test, and several robust structural break tests were made using a framework provided by Heritier and Ronchetti (1994) for a robust Wald test. The Wald and Chow Break test were shown to not be robust to any of the considered outliers in the simulation study, which was theoretically supported by the OLS estimator having an unbounded influence function. For this reason the robust structural break tests were considered. The robust tests were constructed with Mallows type M-estimators in order to ensure that the influence function of the estimator was properly bounded. The simulation results show that this was indeed the case, as the robust tests were robust against all forms of contamination. These conclusions held for large sample sizes and a variety of structural breakpoints. The level of the robust tests is always close to the nominal level for a considered level of $\alpha = 0.05$. In terms of level there is a slight preference for the robust Wald test using the Tukey-bisquare down-weighing function and the weights based on the Hat matrix, however the other versions of the robust Wald test perform good as well. There is no notable drop in the power of the robust diagnostic tests for any of the considered sample sizes or structural breakpoints. The real data example shows that the suggested robust versions of the Wald test are indeed more robust to outliers than the Chow (1960) Break test and Wald (1943) test, yet also highlights one of the key shortcomings of this paper, namely the suggested tests being unable to identify the location of structural breaks.

The scope of this paper is limited mostly by its simulation study. The study focuses on a simple data generating model with a singular regressor. Though adding more regressors in this simple model should not alter the found level and powers of the structural break tests, it would allow more complex structural breaks to be analysed. The paper only covers the nominal level of $\alpha = 0.05$ and does not check whether the structural break tests remain size correct for different values of α, such as $\alpha = 0.01$ or even $\alpha = 0.001$. Additional research could be done into the breakdown point of the robust versions of the Wald test, in order to accurately determine

the best version of this test. Furthermore, a MM-estimator could be considered in place of the Mallows type M-estimator in the robust Wald test. Research into the use of the robust Wald test in the Supremum Wald test might also be interesting, as this would create a robust method to estimate unknown breakpoints.

References

Andrews, D. W. K. & Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, *62*(6), 1383–1414. Retrieved 2024-04-29, from http://www.jstor.org/stable/2951753

Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric Theory*, *13*, 315–352.

Bai, J., Lumsdaine, R. L. & Stock, J. H. (1998). Testing for and dating breaks in integrated and cointegrated time series. *Review of Economic Studies*, *65*.

Bai, J. & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, *66*(1), 47–78. Retrieved 2024-04-29, from http://www.jstor.org/stable/2998540

Carroll, R. J. & Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society. Series B (Methodological)*, *55*(3), 693–706. Retrieved 2024-03-24, from http://www.jstor.org/stable/2345881

Chen, B. & Huang, L. (2018). Nonparametric testing for smooth structural changes in panel data models. *Journal of Econometrics*, *202*(2), 245-267. Retrieved from https://www.sciencedirect.com/science/article/pii/S030440761730221X doi: https://doi.org/10.1016/j.jeconom.2017.10.004

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, *28*(3), 591–605. Retrieved 2023-11-24, from http://www.jstor.org/stable/1910133

Gagliardini, P., Trojani, F. & Urga, G. (2005, Nov). Robust gmm tests for structural breaks. *Journal of Econometrics*, *129*(1–2), 139–182. doi: 10.1016/j.jeconom.2004.09.006

Giles, D. & Scott, M. (1992, February). Some consequences of using the Chow test in the context of autocorrelated disturbances. *Economics Letters*, *38*(2), 145-150. Retrieved from https://ideas.repec.org/a/eee/ecolet/v38y1992i2p145-150.html

Giordani, P., Kohn, R. & van Dijk, D. (2007). A unified approach to nonlinearity, structural change, and outliers. *Journal of Econometrics*, *137*(1), 112-133. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304407606000406 doi: https://doi.org/10.1016/j.jeconom.2006.03.013

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*(346), 383–393. Retrieved 2024-03-22, from http://www.jstor.org/stable/2285666

Hampel, F. R., Stahel, W. A., Ronchetti, E. M. & Rousseeuw, P. (1986). *Robust statistics the approach based on influence functions*. Wiley.

Hansen, B. (2001). The new econometrics of structural change: Dating breaks in u.s. labor productivity. *Journal of Economic Perspectives*, *15*, 117–128.

Heij, C., Dijk, H. K. V., Kloek, T. & Franses, P. H. (2004). *Econometric methods with applications in business and economics*. Oxford University Press.

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

Khan, D. M., Ali, M., Ahmad, Z., Manzoor, S. & Hussain, S. (2021, 11). A new efficient redescending m-estimator for robust fitting of linear regression models in the presence of outliers. *Mathematical Problems in Engineering*, *2021*, 1-11. doi: 10.1155/2021/3090537

Kirkwood, J. B. (1972). The great depression: A structural analysis. *Journal of Money, Credit and Banking*, *4*(4), 811–837. Retrieved 2024-04-29, from `http://www.jstor.org/stable/1991229`

Koshti, V. V. (2011). Cumulative sum control chart. *International Journal of Physics and Mathematical Sciences*, *1*.

Martin, S. H. D., Vance; Hurn. (2013a). *Econometric modelling with time series*. Cambridge University Press.

Martin, S. H. D., Vance; Hurn. (2013b). *Econometric modelling with time series*. Cambridge University Press.

Parker, T. (2015). Finite sample distributions of the wald, likelihood ratio and lagrange multiplier test statistics in the classical linear model. *University of Waterloo*.

Quandt, R. (1960). Test of the hypothesis that a linear regression obeys two separate regimes. *Journal of the American Statistical Association*, *55*, 324–330.

Rousseeuw, P. (1985, 01). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications Vol. B*, 283-297. doi: 10.1007/978-94-009-5438-0_20

Rousseeuw, P. & van Zomeren, B. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, *85*, 633–639.

Storfjell, J. L., Omoike, O. & Ohlson, S. (2008, May). The balancing act. *JONA: The Journal of Nursing Administration*, *38*(5), 244–249. doi: 10.1097/01.nna.0000312771.96610.df

Sun, Y. & Wang, X. (2022). An asymptotically f-distributed chow test in the presence of heteroscedasticity and autocorrelation. *Econometric Reviews*, *41*(2), 177-206. Retrieved from `https://doi.org/10.1080/07474938.2021.1874703` doi: 10.1080/07474938.2021.1874703

Tabot, S. (2023). Exploring structural breaks in international stock markets and its implications. *Journal of Public Administration, Finance and Law*(27), 109–115. doi: 10.47743/jopafl -2023-27-09

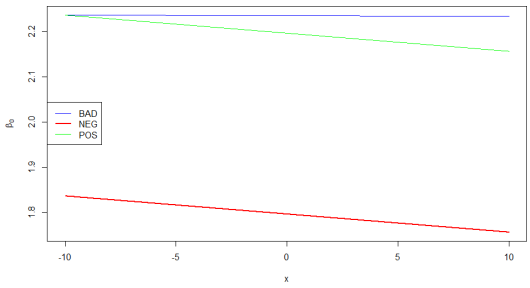
Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(3), 426–482. doi: 10.1090/s0002-9947-1943-0012401-3

Zeileis, A., Leisch, F., Hornik, K. & Kleiber, C. (2002). strucchange: An r package for testing for structural change in linear regression models. *Journal of Statistical Software*, *7*(2), 1–38. doi: 10.18637/jss.v007.i02

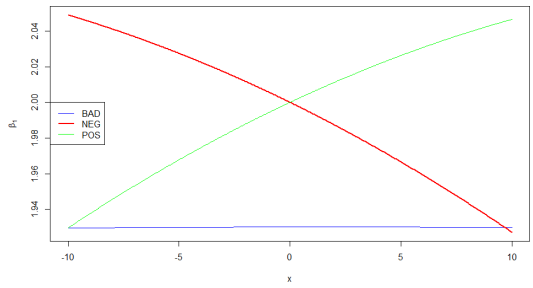
Zeileis, A., Leisch, F., Kleiber, C. & Hornik, K. (2005). Monitoring structural change in dynamic econometric models. *Journal of Applied Econometrics*, *20*, 99–121.

A Appendix

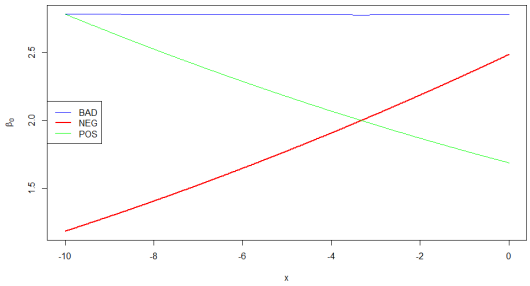
A.1 Outlier effect simulation results



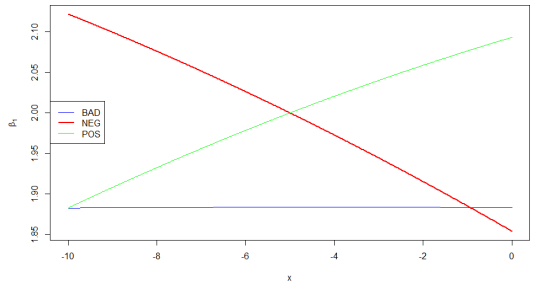
(a) β_0, $x_i \sim \mathcal{U}_{[-10,10]}$



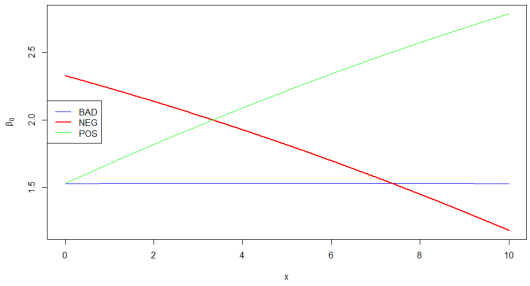
(b) β_1, $x_i \sim \mathcal{U}_{[-10,10]}$



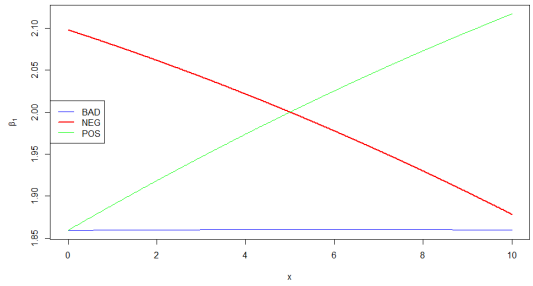
(c) β_0, $x_i \sim \mathcal{U}_{[-10,0]}$



(d) β_1, $x_i \sim \mathcal{U}_{[-10,0]}$



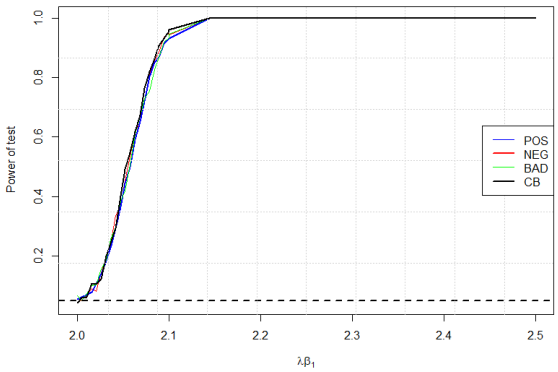
(e) β_0, $x_i \sim \mathcal{U}_{[0,10]}$



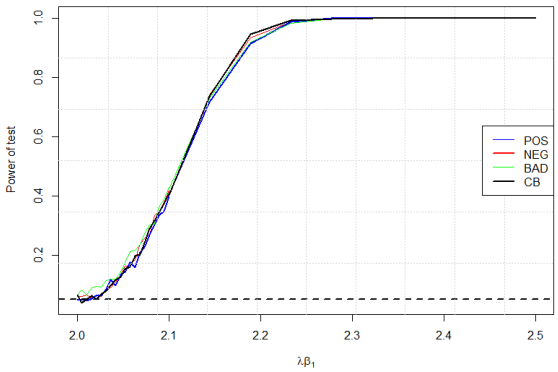
(f) β_1, $x_i \sim \mathcal{U}_{[0,10]}$

Figure A.1

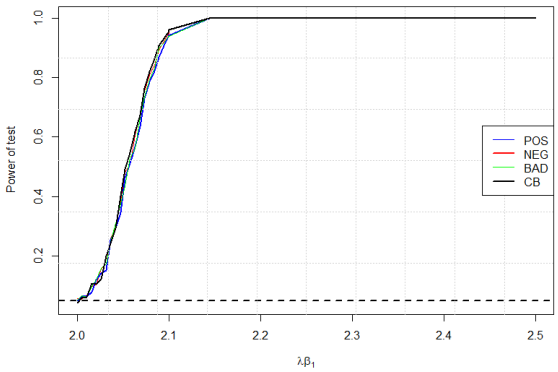
A.2 Power Analysis simulation results



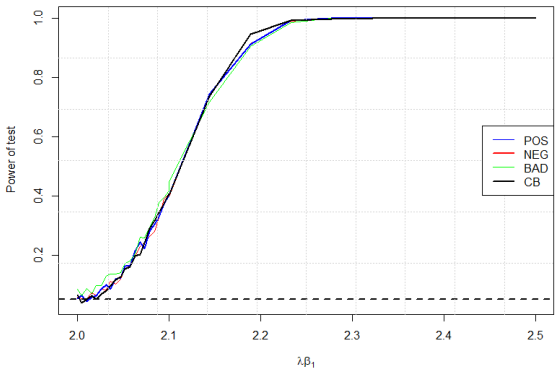
(a) Huber Hat Scenario 2



(b) Huber Hat Scenario 3

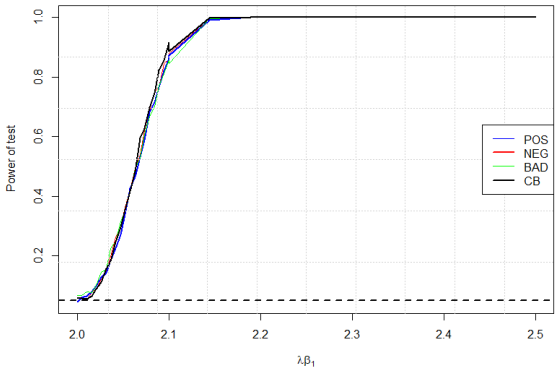


(c) Huber Mah Scenario 2

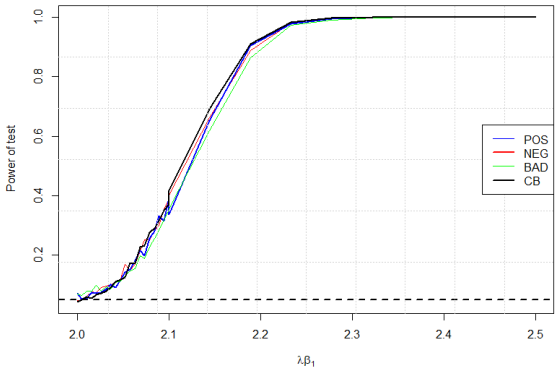


(d) Huber Mah Scenario 3

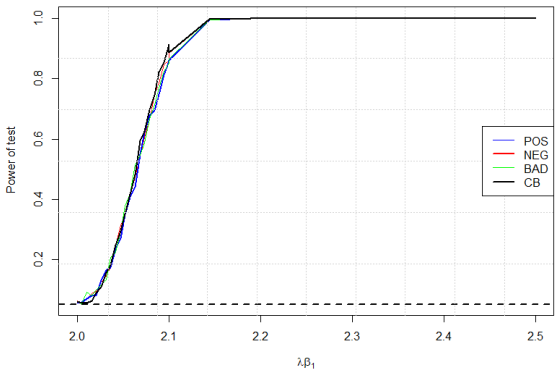
Figure A.2: Power analysis of the robust structural break test using the Huber down-weighing function for the two structural break scenarios. The sample size of the model used was $N = 200$, and the breakpoint was at $n_1 = 100$. Hat indicates the Hat matrix based weight function was used. Mah indicates the Mahalonobis distance based weight function was used. CB indicates the power line for the Chow Break test with no contamination. The contamination level was set to $\epsilon = 0.01$. On the x-axis the points of data were $\lambda\beta_1 \in \{2, 2.005, \cdots, 2.1, 2.15, \cdots, 2.5\}$.



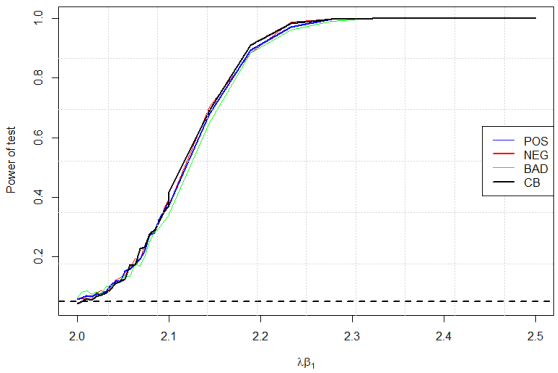
(a) Huber Hat Scenario 2



(b) Huber Hat Scenario 3

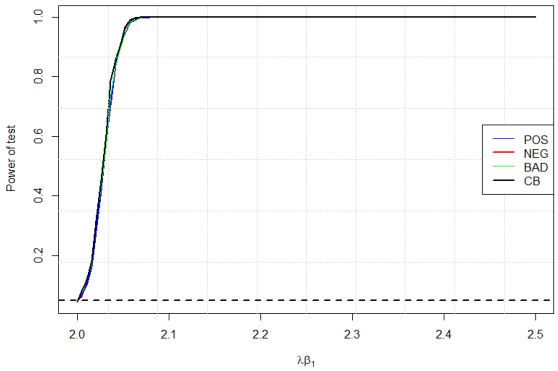


(c) Huber Mah Scenario 2

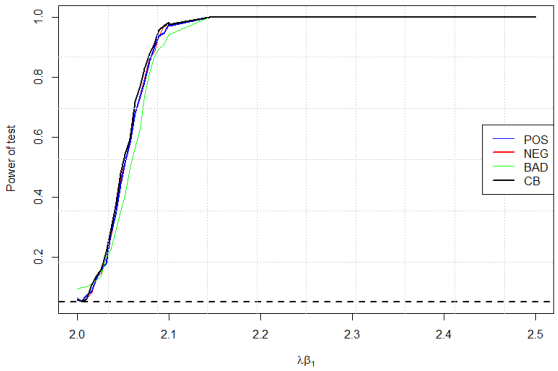


(d) Huber Mah Scenario 3

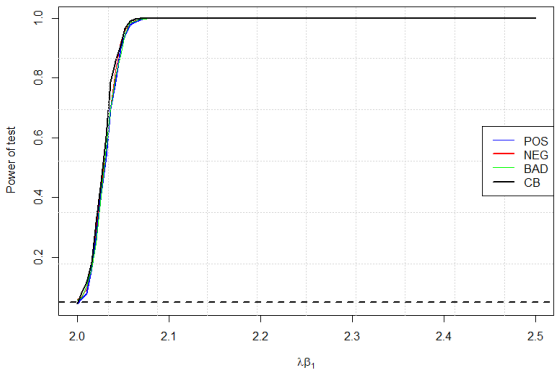
Figure A.3: Power analysis of the robust structural break test using the Huber down-weighing function for the two structural break scenarios. The sample size of the model used was $N = 200$, and the breakpoint was at $n_1 = 150$. Hat indicates the Hat matrix based weight function was used. Mah indicates the Mahalonobis distance based weight function was used. CB indicates the power line for the Chow Break test with no contamination. The contamination level was set to $\epsilon = 0.01$. On the x-axis the points of data were $\lambda\beta_1 \in \{2, 2.005, \cdots, 2.1, 2.15, \cdots, 2.5\}$.



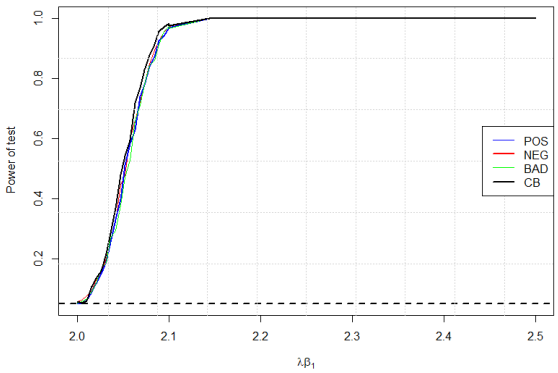
(a) Huber Hat Scenario 2



(b) Huber Hat Scenario 3

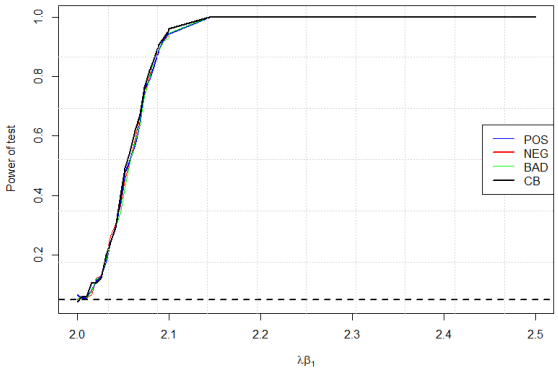


(c) Huber Mah Scenario 2

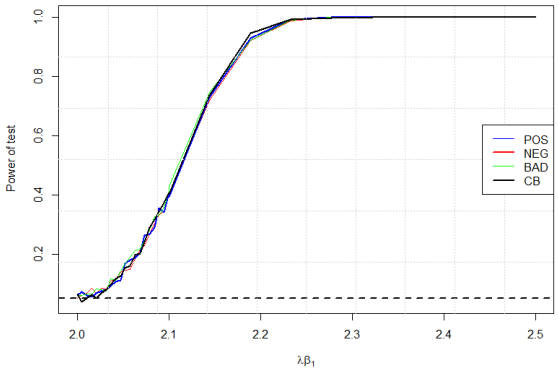


(d) Huber Mah Scenario 3

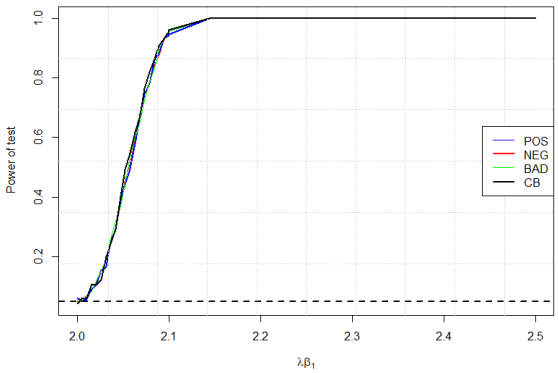
Figure A.4: Power analysis of the robust structural break test using the Huber down-weighing function for the two structural break scenarios. The sample size of the model used was $N = 1000$, and the breakpoint was at $n_1 = 750$. Hat indicates the Hat matrix based weight function was used. Mah indicates the Mahalonobis distance based weight function was used. CB indicates the power line for the Chow Break test with no contamination. The contamination level was set to $\epsilon = 0.01$. On the x-axis the points of data were $\lambda\beta_1 \in \{2, 2.005, \cdots, 2.1, 2.15, \cdots, 2.5\}$.



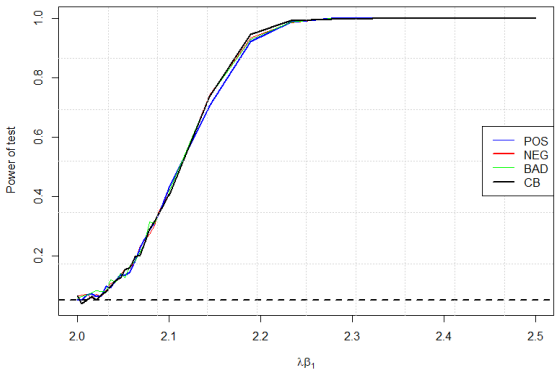
(a) Tukey Hat Scenario 2



(b) Tukey Hat Scenario 3

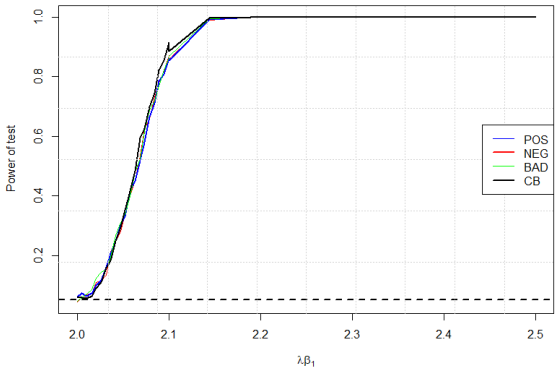


(c) Tukey Mah Scenario 2

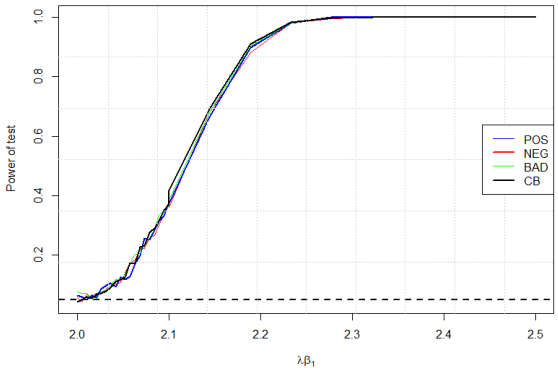


(d) Tukey Mah Scenario 3

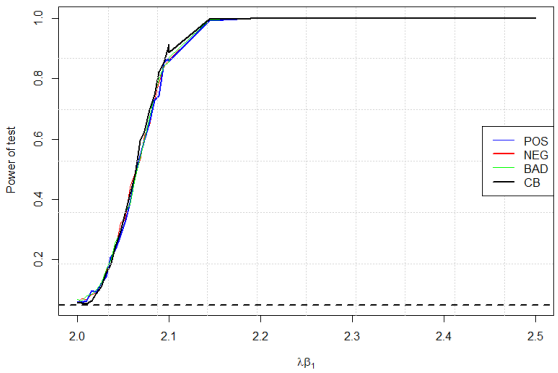
Figure A.5: Power analysis of the robust structural break test using the Tukey bisquare down-weighing function for the two structural break scenarios. The sample size of the model used was $N = 200$, and the breakpoint was at $n_1 = 100$. Hat indicates the Hat matrix based weight function was used. Mah indicates the Mahalonobis distance based weight function was used. CB indicates the power line for the Chow Break test with no contamination. The contamination level was set to $\epsilon = 0.01$. On the x-axis the points of data were $\lambda\beta_1 \in \{2, 2.005, \cdots, 2.1, 2.15, \cdots, 2.5\}$.



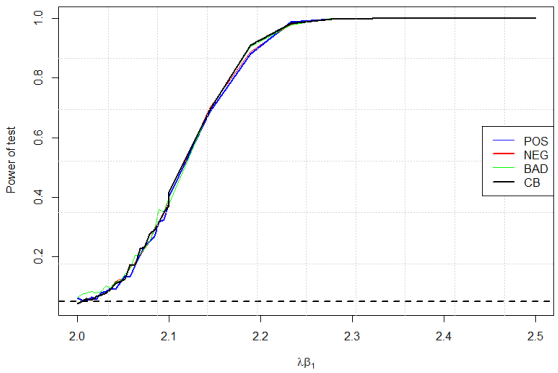
(a) Tukey Hat Scenario 2



(b) Tukey Hat Scenario 3

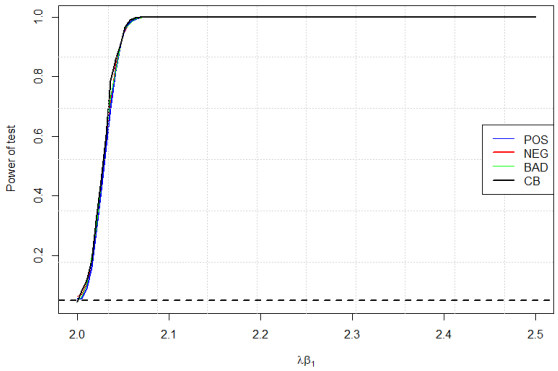


(c) Tukey Mah Scenario 2

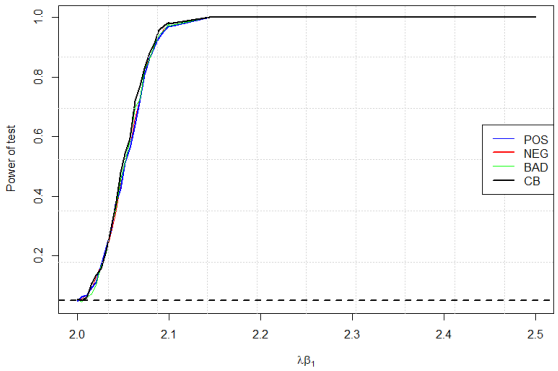


(d) Tukey Mah Scenario 3

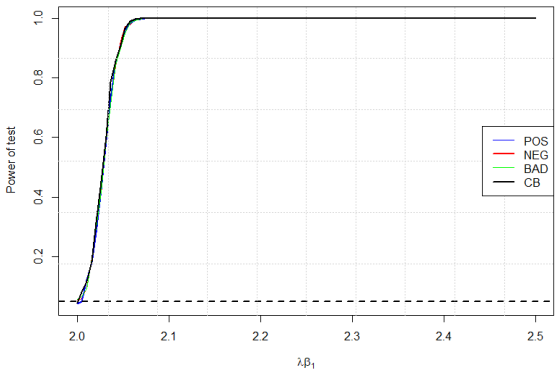
Figure A.6: Power analysis of the robust structural break test using the Tukey bisquare down-weighing function for the two structural break scenarios. The sample size of the model used was $N = 200$, and the breakpoint was at $n_1 = 150$. Hat indicates the Hat matrix based weight function was used. Mah indicates the Mahalonobis distance based weight function was used. CB indicates the power line for the Chow Break test with no contamination. The contamination level was set to $\epsilon = 0.01$. On the x-axis the points of data were $\lambda\beta_1 \in \{2, 2.005, \cdots, 2.1, 2.15, \cdots, 2.5\}$.



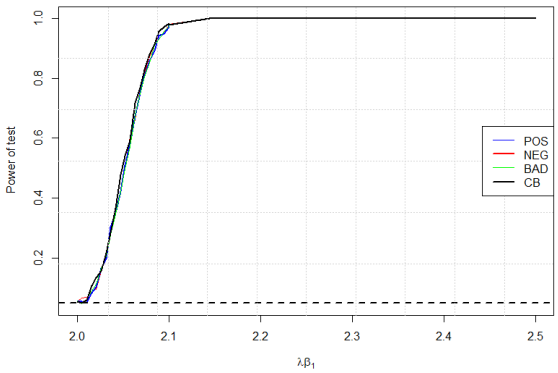
(a) Tukey Hat Scenario 2



(b) Tukey Hat Scenario 3



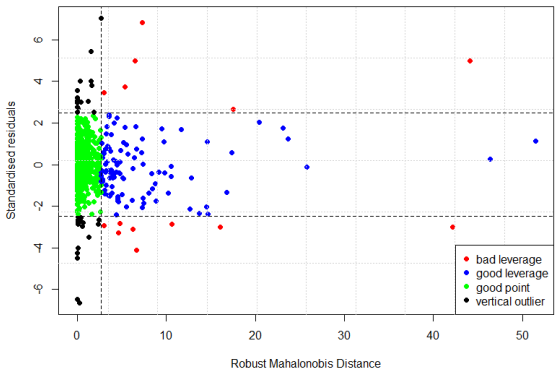
(c) Tukey Mah Scenario 2



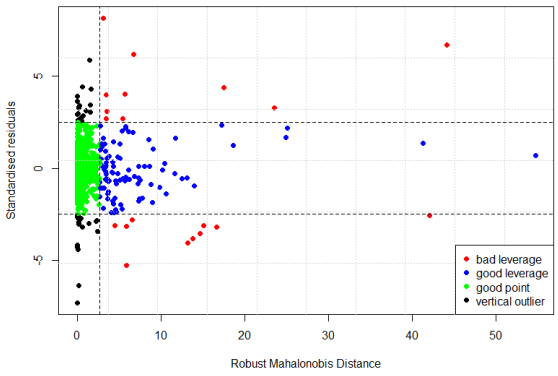
(d) Tukey Mah Scenario 3

Figure A.7: Power analysis of the robust structural break test using the Tukey bisquare down-weighing function for the two structural break scenarios. The sample size of the model used was $N = 1000$, and the breakpoint was at $n_1 = 750$. Hat indicates the Hat matrix based weight function was used. Mah indicates the Mahalonobis distance based weight function was used. CB indicates the power line for the Chow Break test with no contamination. The contamination level was set to $\epsilon = 0.01$. On the x-axis the points of data were $\lambda\beta_1 \in \{2, 2.005, \cdots, 2.1, 2.15, \cdots, 2.5\}$.

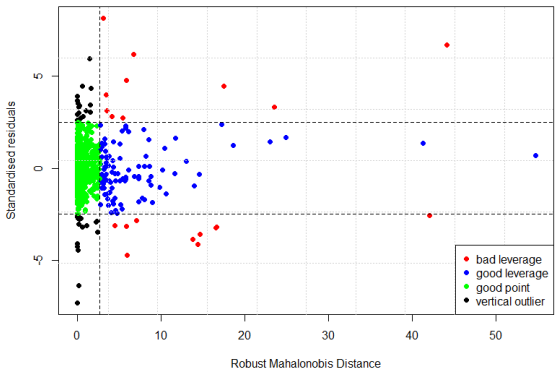
A.3 Diagnostic plots real data example



(a) Two Break Scenario

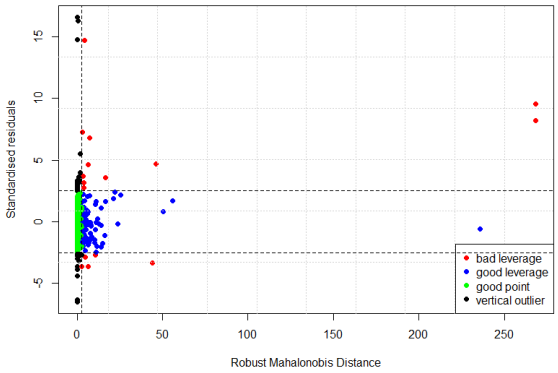


(b) Four Break Scenario

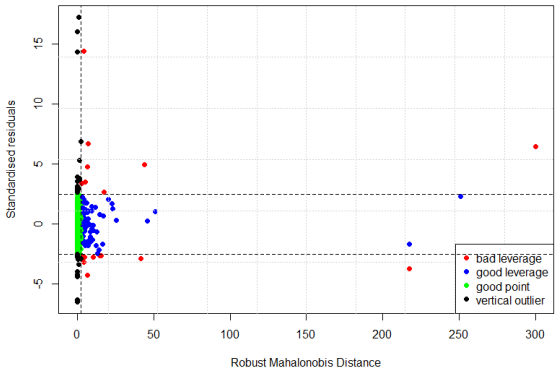


(c) Five Break Scenario

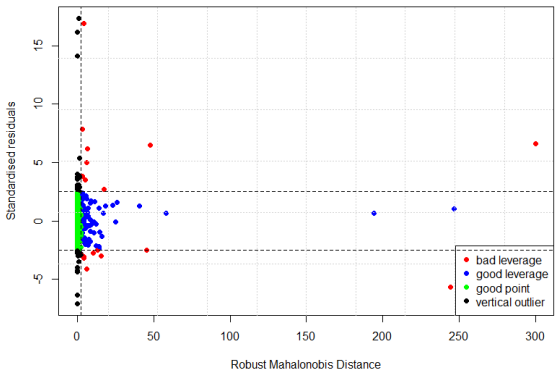
Figure A.8: Robust diagnostic plots for the US labor productivity in the manufacturing/durables sector from February 1947 to April 2001, the cut-off points for the standardized errors and robust Mahalonobis distance are $|\frac{r_i}{\hat{\sigma}}| = 2.5$ and $d_i = \sqrt{\chi^2_{2,0.975}}$ respectively.



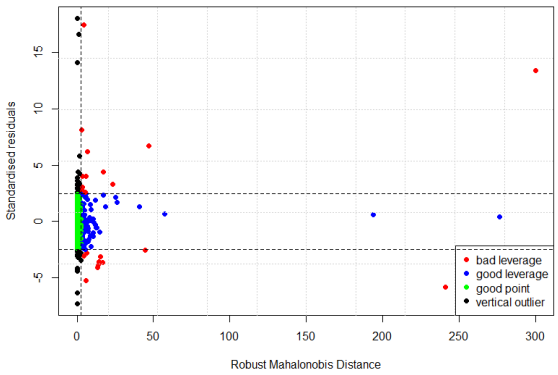
(a) One Break Scenario



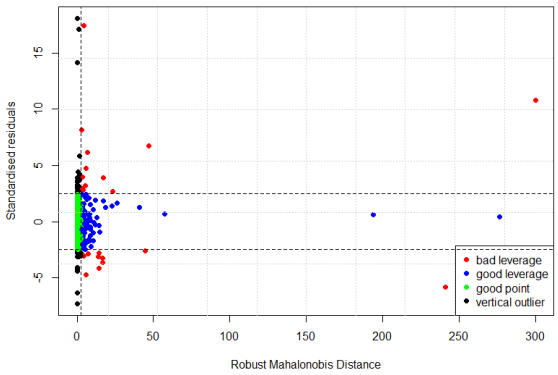
(b) Two Break Scenario



(c) Three Break Scenario



(d) Four Break Scenario



(e) Five Break Scenario

Figure A.9: Robust diagnostic plots for the US labor productivity in the manufacturing/durables sector from February 1947 to April 2001 with artificial contamination at points y_{75}, y_{129}, y_{460} and y_{622}, the cut-off points for the standardized errors and robust Mahalonobis distance are $|\frac{r_i}{\hat{\sigma}}| = 2.5$ and $d_i = \sqrt{\chi^2_{2,0.975}}$ respectively.