Malicious Intent and Multiple Testing in Regression Discontinuity Estimation

Bachelorarbeit zur Erlangung des Grades

BACHELOR OF SCIENCE (B.Sc.)

im Studiengang Volkswirtschaftslehre

an der Rheinischen Friedrich-Wilhelms-Universität Bonn

Themensteller:

Prof. Dr. Joachim FREYBERGER

vorgelegt im Juni 2020 von:

Jakob Ralph JÜRGENS

Matrikelnummer:

Contents

List of Figures in the Main Part

List of Tables in the Main Part

For a full set of Figures and Tables for each data generating process or access to the code used in this thesis, send an email to **jakob.r.juergens@gmail.com**.

1 Introduction

Since the late 1990s Regression Discontinuity (RD) designs have become an important method in economics and related fields to estimate causal effects of an intervention, where assignment of treatment is determined by an observed "assignment" variable. Although RD designs have some advantages over other approaches such as matching on observables and IV methods, they are not a panacea. As nearly every method in statistical inference, RD designs are vulnerable to undisclosed multiple testing by dishonest scientists — part of what is called p-hacking in the scientific community.

If a scientist with malicious intent uses a plethora of methods to estimate a discontinuity and only presents the results that fit their intentions without acknowledging the multiple testing, the presented results and the given confidence levels can differ wildly from the truth. In this thesis I am going to analyse the effects of undisclosed multiple testing of hypotheses in regression discontinuity estimation. To do so, I am first going to give an overview on RD design and describe a method to estimate a discontinuity parametrically. I am then going to present a simulation framework in which the underlying data generating process (dgp) does not contain a discontinuity. Using the estimates obtained in these simulations, I construct three "malicious-intent" (MI) estimators that mimic the behaviour of dishonest scientists. These estimators described in more detail in section 3 overestimate significance and/or effect strength and therefore can be described as malicious since they are used to deceive the reader in some way or another.

2 Regression Discontinuity Design

Introduced by Thistlethwaite and Campbell (1960), Regression Discontinuity designs became an important tool in analysing treatment effects in the 1990s. Applicable in settings where the allocation of treatment is determined by an observed assignment variable according to a threshold rule, they are useful as threshold rules are common in the real world. To give some examples: passing a test at school, winning an election ($(50-\varepsilon)\%$ vs. $(50+\varepsilon)\%$ of voter share), the Maastricht criteria or a minimum height to become an astronaut are thresholds, where it is important on which side you find yourself.
To model these dependencies in a simple linear i.i.d. case one introduces a dummy variable $D \in \{0,1\}$ indicating on which side of the threshold a particular observation is situated into the standard regression model. This results in Equation (1) as a dgp.

$$Y_i = \mu + D_i\tau + X_i\beta + \varepsilon_i \quad i = 1,2,\ldots,n \tag{1}$$

Elements (Y_j, D_j, X_j) of the observed sample $\{(Y_i, D_i, X_i), i = 1,\ldots,n\}$ are i.i.d., ε_i is an error term, $D_i = 1$ if $X_i \geq c$ and $D_i = 0$ if $X_i < c$. X is the assignment variable and c is the threshold which decides whether an individual receives treatment.

For this to fulfil the assumptions of the classical regression model following chapter one of Hayashi (2000), I additionally specify strict exogeneity,

$$\mathbb{E}[\varepsilon_i|D_1,\ldots,D_n,X_1,\ldots,X_n] = \mathbb{E}[\varepsilon_i|X_i] = 0 \quad i = 1,2,\ldots,n \tag{2}$$

no multicollinearity (simplified for the univariate case shown above),

$$rank(\mathbb{X}) = 3, \quad \text{with} \quad \mathbb{X} = \begin{bmatrix} 1 & D_1 & X_1 \\ \vdots & \vdots & \vdots \\ 1 & D_n & X_n \end{bmatrix} \in \mathbb{R}^{n\times 3} \tag{3}$$

and mean-independence of X_i and ε_i

$$\mathbb{E}[X_i|\varepsilon_i] = \mathbb{E}[X_i] \quad \text{and} \quad \mathbb{E}[\varepsilon_i|X_i] = \mathbb{E}[\varepsilon_i] \quad i = 1,2,\ldots,n \tag{4}$$

to allow for a causal interpretation of $\hat{\tau}$.

A case like this is called a sharp RD design, since receipt of treatment is decided strictly according to the threshold. A visualisation for sharp RD design can be seen in Figure 1. As shown there, the probability of receiving treatment jumps from 0 to 1 at $c = 0$.

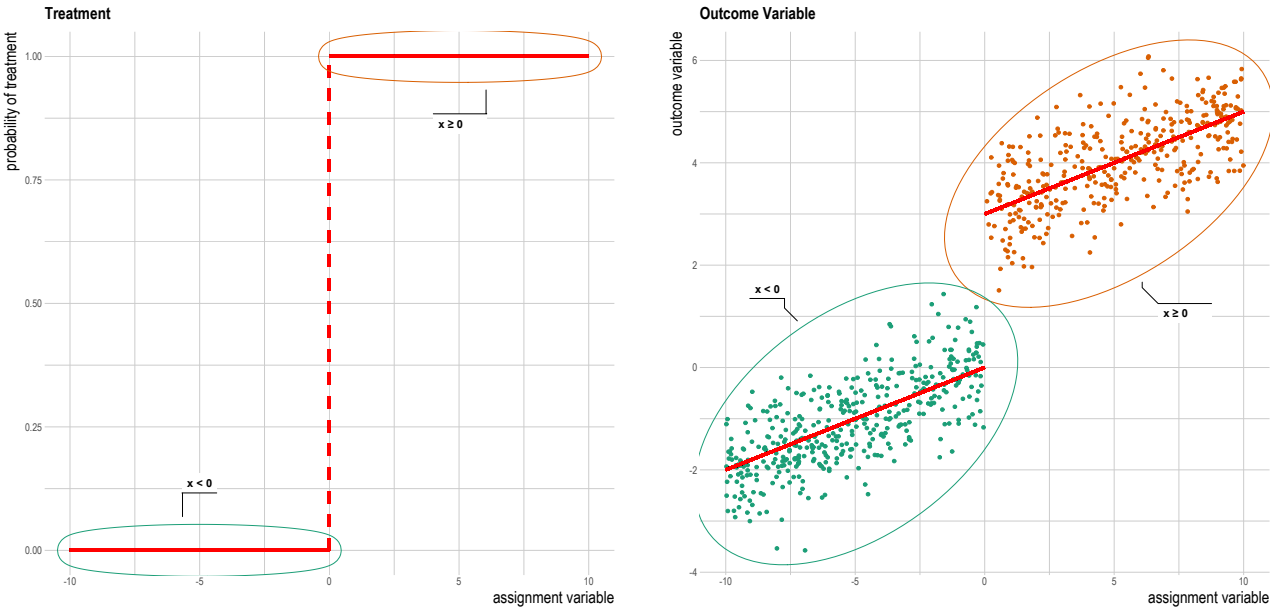

Figure 1: Visualisation for Sharp Regression Discontinuity

In practice, models like this could describe a situation, where all individuals for which X exceeds c are assigned to the treatment group and individuals with a value of X lower than c are assigned to the control group. Similarly to the standard regression framework, this can be extended to allow for heteroskedasticity and multiple covariates.

For the mathematical description of sharp RD designs and Potential Outcomes given below I follow Lee and Lemieux (2010) and Guido W. Imbens and Thomas Lemieux (2008). There are requirements for the usage of an RD framework: the most important being smoothness around the threshold and unconfoundedness. First, all factors other than the treatment dummy must influence the outcome variable smoothly around the threshold. This is motivated by the idea, that individuals in the proximity of the threshold should be comparable even if they are on opposite sides of it to allow for reasonable interpretation of the estimated discontinuity. Since this thesis will only cover the case, where only the assignment variable X systematically influences Y, I am going to limit my description to the univariate case in the following. Mathematically this can be visualized using the so-called Potential Outcomes framework.
Imagine that for each value of X there are two expected values of Y:

- One if the individual receives treatment $\mathbb{E}[Y|X=x,D=1] =: \mathbb{E}[Y(1)|X=x]$
- One if it does not $\mathbb{E}[Y|X=x,D=0] =: \mathbb{E}[Y(0)|X=x]$

Then the expected value of Y given X can be written as shown in Equation (5).

$$\mathbb{E}[Y|X=x] = \text{Prob}(D=1|X=x)\mathbb{E}[Y(1)|X=x] + \text{Prob}(D=0|X=x)\mathbb{E}[Y(0)|X=x] \quad (5)$$

But for each value of X we only ever observe data generated by one of those functions.

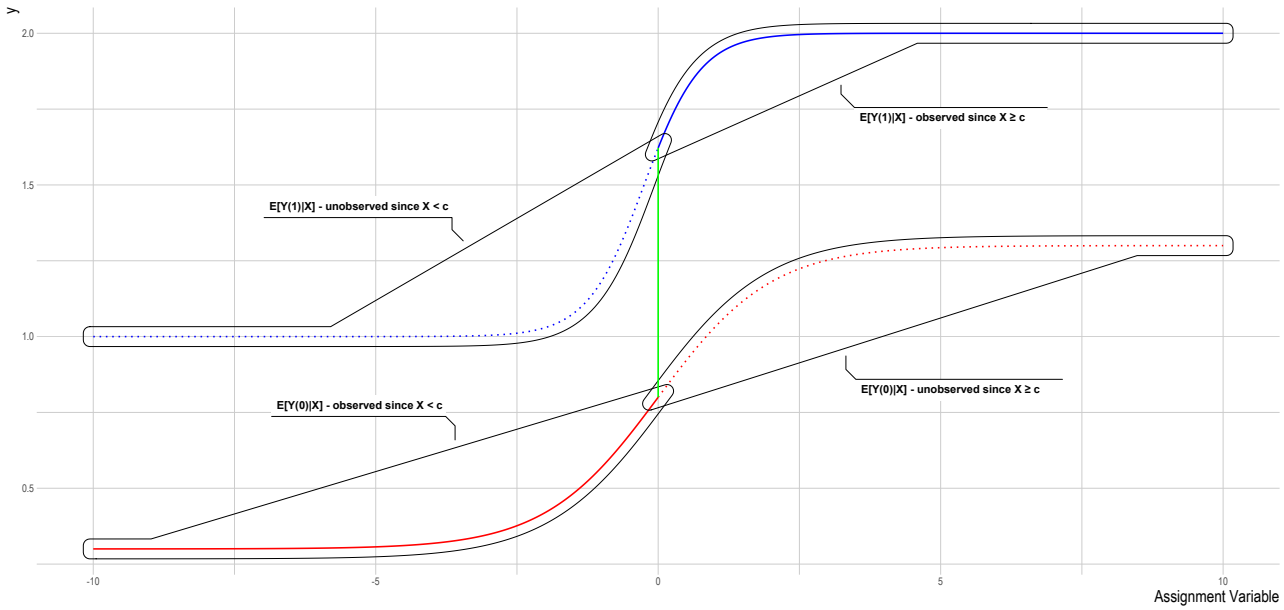

Figure 2: Visualisation for Potential Outcomes Framework

Second, we need unconfoundedness, described in mathematical terms in Equation (6).

$$Y(0), Y(1) \perp D|X \quad (6)$$

This is trivially fulfilled since conditional on X the variance of D is zero. In the case of sharp RD design the average causal effect can be estimated because of the relationship shown in Equation (7), if one assumes unconfoundedness and smoothness of the underlying functions $\mathbb{E}[Y(0)|X=x]$ and $\mathbb{E}[Y(1)|X=x]$ as described above. Estimation of the average causal effect τ at c using the procedures described later is then motivated as shown below.

$$\begin{aligned}\tau = \mathbb{E}[Y(1)|X=c] - \mathbb{E}[Y(0)|X=c] &= \lim_{x\downarrow c} \mathbb{E}[Y(1)|X=x] - \lim_{x\uparrow c} \mathbb{E}[Y(0)|X=x] \\ &= \lim_{x\downarrow c} \mathbb{E}[Y|X=x] - \lim_{x\uparrow c} \mathbb{E}[Y|X=x]\end{aligned} \tag{7}$$

If the receipt of treatment is not decided strictly according to a threshold in an observed variable, but there is still a jump in the probability of receipt of treatment at c, this is called fuzzy RD design. An example can be seen in Figure 7 in the appendix. As shown there, the probability of receiving treatment still jumps at $c = 0$ but not by 1 and therefore treated individuals can be found on the left as well as untreated individuals on the right of the threshold. There are some differences compared to the estimation procedures addressed later, since not only the discontinuity in Y has to be estimated but also the discontinuity in the probability of treatment. In this thesis I am not going to address fuzzy RD designs.

3 Malicious Intent and Multiple Testing

To motivate the idea of multiple testing, think of a simple Gauss test at a significance level of 5% for the null hypothesis $H_0 : \theta_1 = 0$ for some parameter θ_1 estimated by $\hat{\theta}_1$. Under the null, one would expect that $\hat{\theta}_1$ deviates from 0 in a statistically significant way in 5% of cases. Imagine now, that there are multiple parameters $\theta_1, \ldots, \theta_K$ for which similar tests are conducted. Each test creates false positives in about 5% of cases[1] and there is often no reason to assume perfect correlation of those false positives between tests. Therefore, the probability of obtaining at least one false positive, called the family-wise error rate (FWER), is likely larger than 5%. In many contexts it would be reasonable to adjust α to ensure that the FWER is smaller than or equal to 5%. A similar problem arises testing the same hypothesis multiple times for estimates created by different specifications of the functional form. Assuming that different specifications each generate false positives in about 5% of cases and assuming that these are not perfectly correlated, the fraction of false positives is inflated if one only reports the estimates fitting a narrative best.
Suppose one wants to test n hypotheses by n independent tests at a significance level α. Let $\text{fp}_{n,\alpha}$ describe the expected number of false positives in this scenario.

Then it is easy to see that Equation (8) holds.

$$\text{fp}_{n,\alpha} = 1 - (1-\alpha)^n \tag{8}$$

[1] Meaning that the null is falsely rejected.

For $\alpha = 0.05$ this leads to the following values: (rounded for $n \geq 2$)

n	1	2	3	4	5	10	25	50	75	100
$\text{fp}_{n,0.05}$	0.05	0.0975	0.143	0.185	0.226	0.401	0.723	0.923	0.979	0.994

Table 1: Expected fraction of false positives for independent tests at $\alpha = 0.05$

There are multiple ways to correct for these problems, the easiest and most well known being the Bonferroni correction introduced by Dunn (1961). The basic idea is to divide the chosen level of significance α by the number of tests. This more restrictive level of significance is used to test each null hypothesis to ensure that the FWER is smaller or equal to α.
There are different methods to estimate a discontinuity in an RD setting and sometimes many functional specifications for each of those estimators. This indicates the possibility of problems like multiple testing or p-hacking in this context. Imagine a ruthless scientist trying to find a discontinuity where in reality there is none. They could use a multitude of estimators and tests, then only report the results fitting their narrative. Given the calculations above, the more specifications they use, the more likely it might be that they might find a significant result.

3.1 Simulated Data and Monte Carlo Method

To test the effects of multiple testing in a sharp RD setting, I deliberately chose a setting without a discontinuity, namely data as described below.

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \text{ i.i.d. with } X_i \sim unif(-1,1),\ Y_i \sim \mathcal{N}(0,1) \text{ and } X_i \perp Y_i \quad i = 1,2,\ldots,n \tag{9}$$

I then searched for a discontinuity at $c = 0$. Therefore, the typical null hypothesis in these scenarios $\tau = 0$ is true. I did this to conduct a Monte-Carlo-Simulation on the properties of three "malicious-intent" (MI) estimators under the null. These estimators are constructed to mimic the behaviour of dishonest scientists using covert multiple testing to their personal advantage. Developed in the 1930s and 1940s by Enrico Fermi, Stanislaw Ulam, and John von Neumann during the Manhattan project, the Monte Carlo method uses randomness to its advantage if the derivation of exact results is unfeasible.

As Shonkwiler and Mendivil (2009) put it:

> *"The Monte Carlo method is a technique for analyzing phenomena by means of computer algorithms that employ, in an essential way, the generation of random numbers."*

To give an idea of the procedure, imagine an estimator with unknown properties for a given sample size n. To obtain information on the distribution of its properties, one could calculate the estimator for m random samples. For $m \to \infty$ the empirical distribution will converge to the theoretical distribution, therefore making arbitrarily precise estimates of its properties possible. When talking of p-values in the following I refer to the p-value of the null hypothesis that $\tau = 0$ for a specific estimate. My simulation to study the MI estimators consisted of $m = 10000$ data sets with $n = 1000$ observations each.

3.2 "Wanna be certain" - MI1 Estimator

The first MI estimator is constructed to only report the estimate corresponding to the lowest p-value out of a set of estimates. Mathematically, for a set of K estimates $\hat{\tau}_k$ and K corresponding p-values p_k, MI1 is calculated as shown below.

$$\mathbf{MI1}\left(\{(\hat{\tau}_k, p_k) : k = 1, \ldots, K\}\right) = (\hat{\tau}_j, p_j) \quad \text{where} \quad j = \underset{k=1,\ldots,K}{\arg\min}\, p_k \tag{10}$$

I call this estimator the "Wanna be certain" estimator with a bit of irony since people often confuse the p-value with the probability of H_0 being true. Therefore, people might use these estimates to argue that it is very unlikely that there is no discontinuity, even though this is not a rational claim. In the following this estimator will be abbreviated as MI1.

3.3 "Talk Show" - MI2 Estimator

The second MI estimator is constructed, in a way that it always reports the largest estimate in absolute terms for the discontinuity. For a set of K estimates with K corresponding p-values, the MI2 Estimator as defined as described below.

$$\mathbf{MI2}\left(\{(\hat{\tau}_k, p_k) : k = 1, \ldots, K\}\right) = (\hat{\tau}_j, p_j) \quad \text{where} \quad j = \underset{k=1,\ldots,K}{\arg\max}\, |\hat{\tau}_k| \tag{11}$$

I gave MI2 the name "Talk Show" estimator as pure effect size is more relevant in some contexts. One example could be a talk show, where a big number might be more impressive to spectators than a more reasonable, statistically significant estimate. But in some scientific contexts creating an impressive narrative might be more important than statistical rigour too. An example could be policy advice, where stories of huge benefit or crippling damage might be more influential than precise mathematical argumentation. In the following this estimator will be abbreviated as MI2.

3.4 "p-Hacking" - MI3 Estimator

The third MI estimator is inspired by p-hacking and reports the estimate with the highest effect size out of all significant estimates.

So given a set of K estimates with K corresponding p-values, MI3 is defined as follows:

$$\textbf{MI3}\left(\{(\hat{\tau}_k, p_k) : k = 1, \ldots, K\}\right) = \begin{cases} (\hat{\tau}_l, p_l), & \text{if } \{p_k : p_k \leq 0.05 \text{ for } k = 1, \ldots, K\} = \emptyset \\ (\hat{\tau}_j, p_j), & \text{otherwise} \end{cases} \tag{12}$$

$$\text{where} \quad j = \underset{k=1,\ldots,K \text{ with } p_k \leq 0.05}{\arg\max} |\hat{\tau}_k| \quad \text{and} \quad l = \underset{k=1,\ldots,K}{\arg\max} |\hat{\tau}_k|$$

I call this the "p-Hacking" estimator, since its construction is motivated by the same incentives that lead to p-hacking in parts of modern science. In publishing the p-value is often only relevant up to a point — a study might only be published if the reported p-value is smaller than 0.05[2]. Its exact value might not be relevant after that. Therefore, a ruthless scientist might want to report the maximum effect strength out of all significant estimates to maximize the perceived importance of their findings. In the following this estimator will be abbreviated as MI3.

3.5 One-sided Modification of the Malicious Intent Estimators

In some settings dishonest scientists might only be interested in reporting findings with a specific sign or direction. This could be motivated by political intentions or other personal motives. To reflect this, I define a set of modified MI Estimators only looking for positive discontinuities $\tau \geq 0$ without loss of generality due to the underlying symmetry in the dgp. In mathematical terms these MI_+ Estimators can be written as shown below.

$$\textbf{MI1}_+\left(\{(\hat{\tau}_k, p_k) : k = 1, \ldots, K\}\right) = \begin{cases} (\hat{\tau}_l, p_l), & \text{if } \{\hat{\tau}_k : \forall k \in \{1, \ldots, K\} \text{ with } \hat{\tau}_k \geq 0\} = \emptyset \\ (\hat{\tau}_j, p_j), & \text{otherwise} \end{cases} \tag{13}$$

$$\text{where} \quad p_j = \min\{p_k : \forall k \in \{1, \ldots, K\} \text{ with } \hat{\tau}_k \geq 0\} \quad \text{and} \quad l = \underset{k=1,\ldots,K}{\arg\max}\, \hat{\tau}_k$$

$$\textbf{MI2}_+\left(\{(\hat{\tau}_k, p_k) : k = 1, \ldots, K\}\right) = (\hat{\tau}_j, p_j) \quad \text{where} \quad j = \underset{k=1,\ldots,K}{\arg\max}\, \hat{\tau}_k \tag{14}$$

$$\textbf{MI3}_+\left(\{(\hat{\tau}_k, p_k) : k = 1, \ldots, K\}\right) = \begin{cases} (\hat{\tau}_l, p_l), & \text{if } \{p_k : p_k \leq 0.05 \text{ for } k = 1, \ldots, K\} = \emptyset \\ (\hat{\tau}_j, p_j), & \text{otherwise} \end{cases} \tag{15}$$

$$\text{where} \quad j = \underset{k=1,\ldots,K \text{ with } p_k \leq 0.05}{\arg\max}\, \hat{\tau}_k \quad \text{and} \quad l = \underset{k=1,\ldots,K}{\arg\max}\, \hat{\tau}_k$$

For MI3_+ one could argue about the relative importance of positivity and significance. As can be seen in the construction of MI3_+ I chose to prioritize significance over positivity to reflect the importance of p-values in publishing. Different choices could be reasonable.

[2] One could argue that this in itself is a discontinuity that could be studied using RD designs.

4 Estimation of Sharp Discontinuities

Since these MI estimators need a set of estimates to choose from, I used a number of Ordinary Least Squares (OLS) estimators with different specifications of the functional form to estimate the discontinuity. I did so, since OLS is the best linear unbiased estimator for data generated by a dgp like the one used in this thesis. In other settings it might be more appropriate to choose different estimators, for example non-parametric approaches, which I will shortly address in the outlook. To increase the number of specifications of the functional form, I started with a constant specification and added powers of x to the regression. Choosing powers of x instead of other terms such as logarithms or trigonometric functions is reasonable, since sufficiently smooth functions can be expressed as a power series, called the Taylor series of that function. Including powers up to K gives us an approximation of the function called its Taylor polynomial of degree K. I stopped adding higher powers at x^6 since R reported a violation of the rank condition when adding x^7 to the regression model. This is a well known property[3] which is studied in numerical mathematics and will not be addressed further in this thesis.
Nevertheless, it is possible to add higher powers of x to the regression. For example, using Legendre polynomials instead of powers of x makes it possible to include powers up to $n-2$ to the regression as these are orthogonal by definition. With enough computational resources this could be a way to study the properties of the MI estimators when using equivalents of higher K in the regression model. Other approaches such as using a Fourier series as theoretical motivation might also be reasonable if the context is appropriate[4] and could be studied in another simulation.

One important property of these estimators is that they are unbiased. Therefore, I am going to ignore errors because of an incorrect specification of the functional form in the following analysis. To estimate the discontinuity I used a pooled regression (in the sense, that I estimated the regression on both sides of the threshold in one step) for each specification of the functional form. The advantage of this approach compared to estimating both sides independently is that it delivers standard errors using standard OLS theory. I allowed for differing constants and slope parameters on different sides of the threshold since $\mathbb{E}[Y(1)|X=x]$ and $\mathbb{E}[Y(0)|X=x]$ do not necessarily have to share these characteristics. There are some points that can be made for restricting these parameters to match across treatment and control group, but aside from special cases (where it is theoretically justified) the standard approach is not to restrict them. This is done to ensure that estimates of Y on the left of c only rest on observations to the left of c and vice versa as implied by Equation (7).

From now on K signifies the order of the polynomial specification of one specific estimator while J refers to the highest order of polynomial estimator included in the calculation of the MI estimators. As a special case $K=0$ refers to the estimator with constant specification and $J=0$ refers to the case where only the constant estimator is included.

[3] Not specifically of x^7 but the general approach of adding powers of x to a regression.
[4] For example in a time series context where cyclicity might reasonably be assumed.

4.1 OLS Estimator with Constant Specification

Using the OLS estimator described above while including only constant terms is equivalent to calculating the arithmetic mean on the left and right of the suspected discontinuity. The estimate for the discontinuity is the difference of those means. Equation (16) describes the estimated relation.

$$Y = \hat{\alpha} + D\hat{\tau} + \hat{\varepsilon} \tag{16}$$

If one would stop testing after this step or limit the analysis to any other value of K, there would be no problem of multiple testing, but it would also be unlikely (5%) that one would find a discontinuity that significantly differs from zero.

4.2 OLS Estimator with General Polynomial Specification

Similar to the constant approach this is equivalent to estimating an OLS regression including powers of x up to a limit K on either side of $c = 0$ and then to calculate the difference of those regressions at zero.

$$Y = \hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_{l,i} X^i + D\hat{\tau} + D\sum_{i=1}^{k} (\hat{\beta}_{r,i} - \hat{\beta}_{l,i}) X^i + \hat{\varepsilon} \tag{17}$$

Looking at the underlying dgp each of these functional forms is correctly specified, which results in a Bias of zero. Nevertheless, due to randomness, the estimators find large discontinuities in some data sets, especially for higher order polynomials due to the increased variance. The maximum discontinuity for each specification including constant is shown in Figure 6 in the appendix. Comparing the estimates created by estimators for different choices of K gives an overview of what the MI estimators get to work with in this thesis.

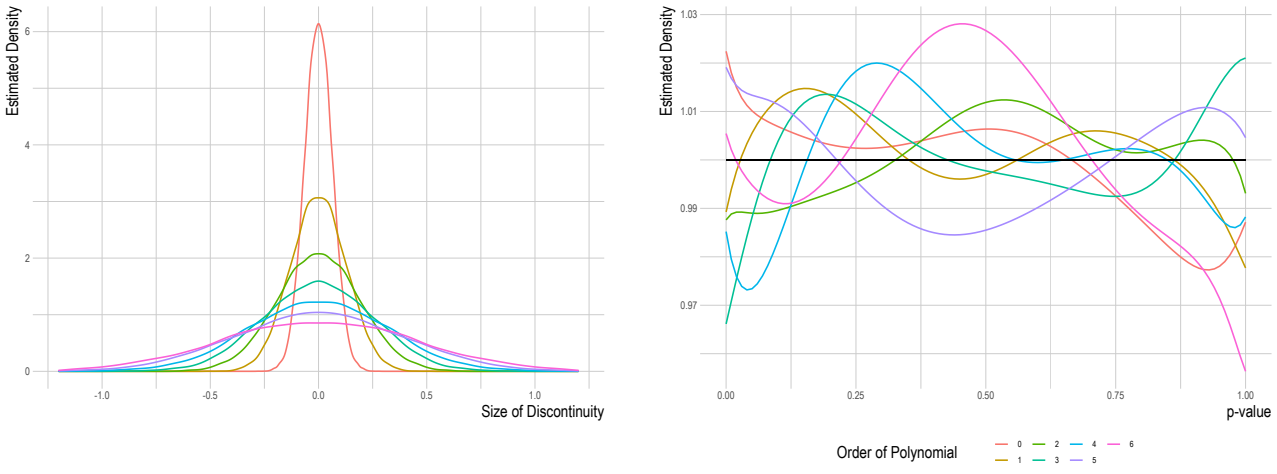

Figure 3: Estimated Densities for Discontinuities and p-values by order of Polynomial

Figure 3 shows the estimated densities of the discontinuity estimates for different choices of K in the left panel[5] and the estimated densities for the p-values in the right panel. Since the latter are bounded to $[0,1]$ I used a density estimator for bounded data provided by the R package *bde*[6]. As expected the estimated discontinuities show a bigger variance for higher orders of polynomial specification and the p-values approximately follow a uniform distribution on $[0,1]$ under the null. As a comparison, Figure 8 shows the histograms for the p-values of each specification which also point to a uniform distribution.

K / K	0	1	2	3	4	5	6
0	1	0.15	0.0764	0.0342	0.0202	0.00368	0.00927
1	0.501	1	0.316	0.169	0.0982	0.0684	0.0579
2	0.344	0.671	1	0.415	0.248	0.162	0.117
3	0.253	0.504	0.742	1	0.489	0.315	0.231
4	0.201	0.405	0.589	0.789	1	0.562	0.375
5	0.169	0.336	0.492	0.656	0.831	1	0.608
6	0.141	0.291	0.421	0.565	0.708	0.855	1

Table 2: Correlation between $|\hat{\tau}|$ for different K (purple) and p-values of different K (green)

Table 2 shows the correlations between discontinuity estimates created by estimators for different K under the main diagonal and correlations between p-values for estimates created by those estimators above the main diagonal. Overall the correlation seems to be a relevant factor and might lead to a slowdown in the increase of false positives with increasing J, since additional specifications are increasingly correlated with the ones calculated before. As shown there too, correlation between estimates for k' and $k'+1$ seems to be increasing as k' increases. The same can be observed for the corresponding p-values, indicating that the postulated slowdown in the increase of false positives gets stronger as J increases. This could be an interesting subject to study in another simulation for higher values of K. This could be used to analyse the effects on the MI estimators for larger values of J too. It might be possible that for $J = n-2$ (the highest order that can possibly fulfil the no-collinearity condition) the fraction of false positives does not go to one as $n \to \infty$.

K	0	1	2	3	4	5	6
correlation coefficient	-0.946	-0.951	-0.95	-0.947	-0.941	-0.942	-0.938

Table 3: Correlation between $|\hat{\tau}|$ and their corresponding p-values for different K

As Table 3 shows, the correlation for each chosen value of K between $|\hat{\tau}|$ and its corresponding p-value is close to -1. This means that false positives are typically generated by large discontinuity estimates motivating similarities between MI1 and MI3, and MI2.

[5]Due to the nature of these estimators I mirrored the estimates at zero to obtain a more precise estimate.
[6]https://cran.r-project.org/web/packages/bde/index.html

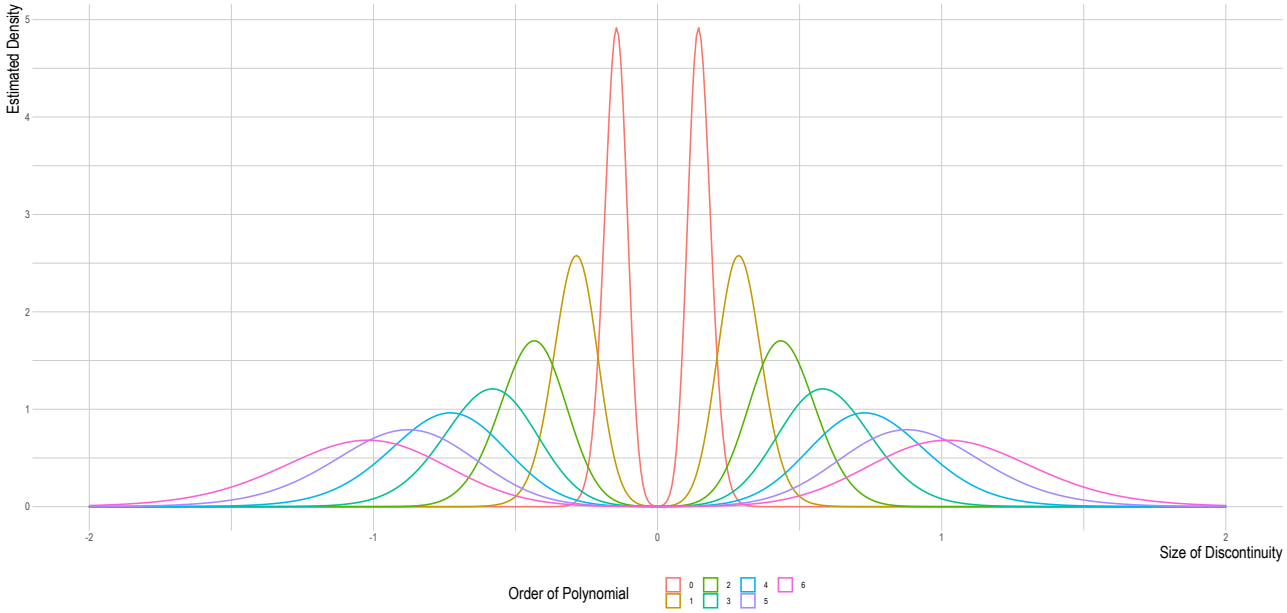

Figure 4: Estimated Densities for significant Discontinuity Estimates

Figure 4 shows density estimates for only the significant realisations of the OLS estimators for $K = 0, \ldots, 6$. These are interesting from a practical standpoint since a scientist with malicious intent might only publish significant results, as explained before. Therefore, these distributions and those realised by the significant estimates chosen by the MI estimators might be more relevant for the effects on published science in the real world.

As can be seen there, each density is bimodal, which makes sense, since only the absolute value of the estimated discontinuity is relevant for its p-value but not its sign. Since the peaks of the distributions wander outwards as K increases, there might be a possibility to identify peaks caused by certain underlying estimators in the significant estimates chosen by the MI estimators. Additionally, this identification idea could be supported using the decomposition by chosen polynomial order K for each MI estimator. A theoretical analysis of these distributions seems feasible and could be an approach to derive theoretical distributions of the MI estimators constructed in this thesis.

4.3 Findings for MI Estimators

Calculating the MI estimators described in section 3 for the estimates created by these polynomial OLS estimators ($K = 0, \ldots, 6$) gives an idea of what effects undisclosed multiple testing might have on published science. In the following I am going to describe the observed properties of the MI estimators. First I will look at the observed fraction of false positives for sets of included estimators with polynomial specification up to order J, shown in Figure 5 and Table 4. If these were to differ systematically from 5%, this would be a first indicator of problematic behaviour connected to multiple testing. Then I am going to address the MI estimators individually.

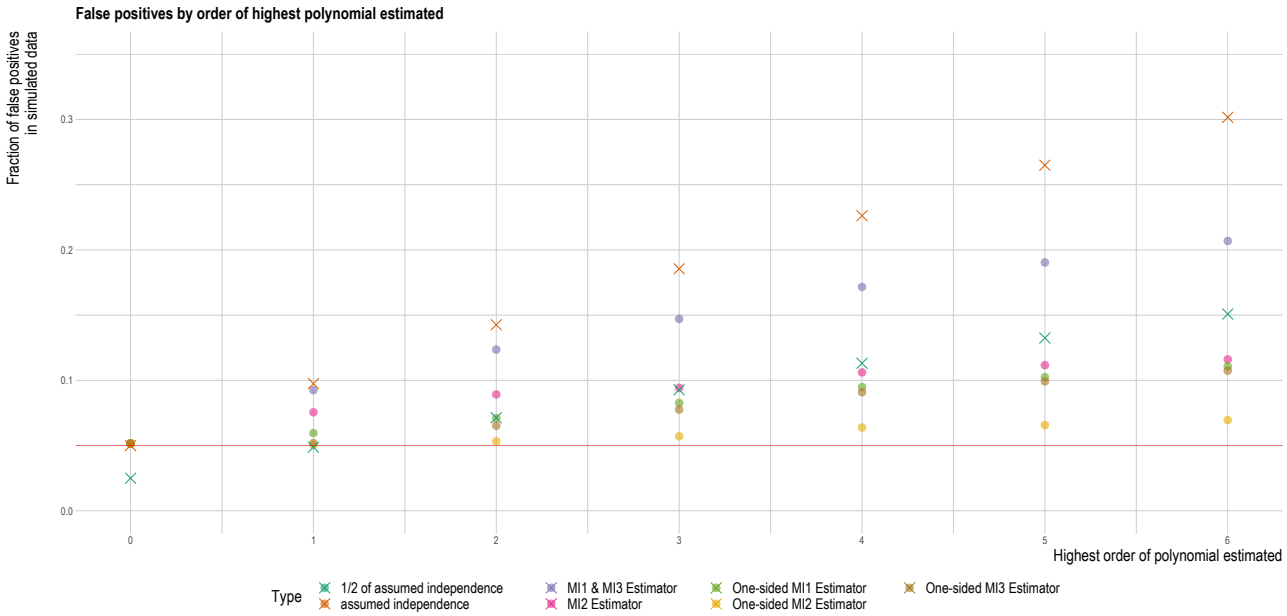

Figure 5: Fraction of False Positives by Maximum Order of Polynomial estimated

n (J)	1 (0)	2 (1)	3 (2)	4 (3)	5 (4)	6 (5)	7 (6)
$\text{fp}_{n,0.05}$	0.05	0.0975	0.143	0.185	0.226	0.265	0.302
MI1 and MI3	0.0516	0.0925	0.1236	0.1471	0.1716	0.1904	0.2068
MI2	0.0516	0.0755	0.0892	0.0943	0.1061	0.1116	0.1161

Table 4: False Positives for the MI Estimators at $\alpha = 0.05$

The term assumed independence in Figure 5 relates to Equation (8) for $\alpha = 0.05$. From a theoretical standpoint, it is clear, that the fraction of false positives of MI1 and MI3 have to be identical, since each would choose an estimate with a p-value smaller than 0.05 if one exists.
As can be seen in Figure 5 and Table 4 the quota of false positives is increasing with J for all MI and MI_+ estimators. Due to correlation between the estimates created by the underlying OLS estimators, the increase is slower than the assumption of complete independence predicts. Compared to MI1 and MI3, MI2 creates fewer false positives, which is expected since large estimates as chosen by MI2 do not necessarily have to be significant. As Table 3 only shows the in-estimator correlations of estimates and p-values, these do not contradict this observation. MI1 and MI3 will choose a significant estimate whenever there is one, a feature that MI2 does not share by construction. MI2 showing a smaller fraction of false positives could indicate that unusually large estimates are rarely responsible for false-positives.
These observations motivate the idea that a correction of α derived from the assumption of complete independence might be conservative especially for large values of J. Studying the joint distribution of these discontinuity estimates from a theoretical standpoint and creating an appropriate correction for the choice of α could be an interesting next step. An outline is given in section 9.5.

MI1 Estimator As shown in Figure 12, MI1 starts to show a bimodal nature for $J \geq 1$ due to its symmetric construction. The estimated density gets flatter and wider as J increases, as well as the modes getting farther away from zero. This makes sense since MI1 includes estimators for higher values of K which show a larger variance. Restricting the analysis to the false positives as shown in Figure 13, MI1 gains another interesting feature as for $J = 1$ the density clearly shows two modes on either side of zero. These could correspond to the two underlying estimators, a hypothesis that is supported by Figure 4. This clarity is lost for higher values of J but the overall structure seems to support the idea. As J increases the estimated density of the false positives gets wider and flatter. Around zero the behaviour is contrary to this observation. Here the density increases with J.
As can be seen in Figure 10 the estimates created for $K = 0$ continuously account for the largest share of estimates chosen by MI1, which could be explained by the limited correlation of p-values for estimates created for small values of K with other p-values. Therefore, the relative importance of $K = 0$ in MI1 could be expected to be relatively stable. Overall the decomposition by order of the chosen estimator is rather flat compared to the ones for MI2 and MI3 and seems to approach a uniform distribution as J increases. This is reasonable since under the null, the p-values of each underlying estimator should approximately follow a uniform distribution on $[0,1]$ and therefore have the same probability of delivering the smallest p-value. The difference to MI2 and MI3 can also be linked to the limited correlation of p-values for different estimators compared to the larger correlation between the corresponding estimates. In the case of MI1 restricting the analysis to the significant observations does not change the decomposition by chosen polynomial estimator meaningfully, indicating that false positives are generated proportionally by the chosen estimators.
Looking at the histograms for the p-values created by different J in Figures 14 and 15 I observe that the histograms for MI1 become increasingly right skewed as J increases. A feature that distinguishes MI1 from the other MI estimators is, that this histograms indicate a convex density for the p-values reported by MI1. Overall the histograms hint to severe problems of multiple testing as these do not approximate a uniform distribution on $[0,1]$ as would be expected under the null.

MI2 Estimator As expected since the correlations in Table 3 are close to -1, the overall shape and development with increasing J of MI2 is similar to the one of MI1. The strong correlations mean, that out of the set of estimates created for a value of K, similar estimates are chosen by both MI1 and MI2. In the following I am therefore going to address the differences in Figures 10 to 15 between MI1 and MI2.
Compared to MI1, the density of MI2 is wider and its modes are farther out, which is explained by its construction. Where MI1 chooses the smallest p-value, MI2 chooses the largest effect size. Even though those are correlated, differences such as those shown by Figure 12 are expected. Looking at the distribution of false positives, MI2 shows a wider density which is reasonable due to the same explanation. For $J = 1$ the distribution again seems to show a differentiation into two modes on either side of zero. In this case the inner bump is smaller

— exactly opposite as in the case of MI1. This is consistent with Figure 4 and the findings of Figure 11. In case there are multiple significant estimates to choose from, it is likely that the largest was created for a high value of K. Comparing the decompositions by chosen estimator for the significant estimates (Figure 11) further supports this idea, since out of the false positives reported by MI2, large portions are chosen from estimators corresponding to higher K. A similar pattern emerges for the complete set of estimates by MI2 as Figure 10 shows. $K = J$ is always the estimator chosen most often. This can be explained by the increasing variance of the underlying estimators as K increases.
Due to these properties it seems reasonable to assume that the modes closer to zero correspond to $K = 0$. For higher values of J this pattern is not obvious, but further theoretical analyses could be made to describe the resulting distribution of false positives. One hypothesis could be that for each increase in J the distribution for $J - 1$ is overlaid by a component for $K = J$ and rescaled with a large weight on the latter.

MI3 Estimator Due to its two-stage definition MI3 shares properties of both MI1 and MI2. Looking at Figures 10 and 12 shows that MI3 seems to behave very similar to MI2 both in density and in its choice of underlying estimators. But important differences start to emerge when looking at the false positives — not only does MI3 produce by far more false positives as Table 4 shows, the density of its false-positives and the decomposition by chosen polynomial estimator for significant results show strong differences.
As Figure 13 shows, the estimated density of the reported false positives is quite similar to the one for MI1. But especially for lower values of J such as $J = 1$ to $J = 3$ the densities seem to be more flat. By flat I mean that compared to MI1 and MI2 which had opposite weights on the peaks depicted in Figure 4, these seem to be more evenly weighted for MI3. This is reasonable when contrasted with Figure 11, where the decomposition of MI3 seems to mix characteristics of those for MI1 and MI2.
Looking at the histograms for the p-values of chosen estimates in Figures 14 and 15 shows that as with the other characteristics, MI3 exhibits features of MI1 for significant estimates, creating many false positives — with the exception that MI3 shows more estimates in the bin to the left of 0.05 than in the utmost left bin. This indicates some sort of trade-off: choosing higher p-values to increase reported effect size. This behaviour can be found in the definition of MI3. For the insignificant choices MI3 once again mimics the behaviour of MI2 which is reasonable since their definitions match for the case of no significant estimators to choose from. As these outweigh the significant choices by far, these similarities translate to the overall estimator.

4.4 Findings for One-Sided MI Estimators

The same analysis conducted above for the two-sided MI estimators can be applied to the one-sided MI estimators. In this case due to the nature of the estimators, which essentially discard half of the estimates due to negativity, the expected fraction of false positives is also halved. Since the MI estimators might not be able to choose a positive estimate for small values of J this idea only holds approximately. For large values of J, due to the increasing correlation with increasing K, the same idea applies by a slightly extended argument. Therefore, halving the theoretical value for complete independence is only a guideline and cannot be used as freely as in the two-sided case.

n (J)	1 (0)	2 (1)	3 (2)	4 (3)	5 (4)	6 (5)	7 (6)
$\frac{1}{2}\text{fp}_{n,0.05}$	0.025	0.0486	0.071	0.093	0.113	0.132	0.151
MI1_+	0.0516	0.0596	0.071	0.0827	0.0949	0.1025	0.1108
$\hat{\tau} \geq 0$	0.0266	0.0473	0.0633	0.0766	0.0903	0.0988	0.1072
MI2_+	0.0516	0.0504	0.0533	0.0571	0.0638	0.0657	0.0695
$\hat{\tau} \geq 0$	0.0266	0.0381	0.0456	0.051	0.0592	0.0620	0.0659
MI3_+	0.0516	0.0519	0.0652	0.0776	0.909	0.992	0.1074
$\hat{\tau} \geq 0$	0.0266	0.0473	0.0633	0.0766	0.903	0.0988	0.1072

Table 5: False Positives for the One-Sided MI Estimators at $\alpha = 0.05$

As Figure 5 and Table 5 show, the one-sided variants behave more tame in terms of the fraction of false-positives they report. Due to their nature of essentially dictating a direction of the effect, this is expected and cannot really be taken as a redeeming quality due to its implications for other properties of the estimators for example their non-zero bias.

Nevertheless, as in the two-sided case, the fraction of false positives increases as J increases. Again the rate is significantly slower than for the case of independent tests. This can be explained by the same logic as before by looking at the correlation between estimates and p-values of estimators for different values of K. Restricting the false-positives to be positive, as might reasonably done due to the real life motivation of this modification, has interesting effects on the fraction of false positives. Where the difference is quite large for small values of J, the difference quickly approaches zero as J increases, indicating that although estimates are increasingly (positively) correlated as K increases, the one-sided MI estimators are able to choose positive estimates nearly exclusively for even relatively small values of J. Compared to their two-sided equivalents the increase in false positives as J increases is even slower, motivating the idea that the effectiveness of increasing J to achieve false positives is quite limited for these Estimators. The observation that the correlation between estimators for $K = k'$ and $K = k' + 1$ quickly increases with k' supports this idea, since the additional estimators become increasingly correlated as J increases.

MI1$_+$ Estimator As with the two-sided MI estimators, their one-sided counterparts share their overall shape as Figure 18 shows. The density estimates for MI1 show a strong right skewness which increases with J. Due to their construction, the density is unimodal with the mode being situated slightly to the right of zero and getting farther away from zero as J increases. This can be explained looking at Figure 16, as the decomposition of MI1$_+$ by chosen polynomial estimator shows that $K = J$ continuously accounts for the second largest share of chosen estimates. Therefore, estimators with a larger variance are included as J increases, flattening the density and giving more weight to the tails. As seen for MI1 before, the largest share is continuously chosen from $K = 0$ with an analogous explanation as in the two-sided case.
Looking only at the significant estimates, Figure 19 shows an interesting feature of the one-sided MI estimators. For small values of J there is a mode to the left of zero created by data sets where no positive estimate was generated. This mode quickly shrinks as J increases, once again showing that although the estimates for different values of K are correlated, higher values of J lead to a higher fraction of data sets showing at least one positive estimate. The peak to the left of zero does not seem to change position as J increases motivating the idea that it corresponds to a single estimator. Comparison to Figure 4 could indicate that this peak corresponds to $K = 0$ as their position is similar.
To the right of zero, one finds a similar pattern to the two-sided variant. For small values of J there are $J + 1$ clearly separated modes which probably correspond to the underlying estimators. This feature is not clear for higher J as the distribution becomes wider and flatter. Figure 17 shows strong similarity to Figure 16 for MI1$_+$ indicating a proportionality between the chosen estimators and the observed false positives. Looking at Figures 20 and 21 shows a similar pattern to the two-sided MI1 estimator. Once again the distribution is right skewed and convex. However, the distribution is less skewed than for the two-sided MI1, which can be explained by the points made for the fraction of false positives at the beginning of section 4.4.

MI2$_+$ Estimator Once again, I am going to elaborate on the differences between MI2$_+$ and MI1$_+$ as they share many properties. Compared to MI1$_+$ the density estimates for MI2$_+$ are flatter overall and show a larger share of estimates with a comparatively large effect size as shown in Figure 18. The density estimates for the significant realisations of MI2$_+$ shown in Figure 19 exhibit a similar pattern to MI1$_+$ with an analogous distinction to the one observed for their two-sided equivalents. Where MI1$_+$ has its peak relatively close to zero, MI2$_+$ peaks at larger values.
This can be explained looking at the decomposition by chosen polynomial estimator and Figure 4. MI2$_+$ prefers estimates created by the estimator for the highest K as these often produce the larger estimates due to their increased variance. This pattern is also present for the significant results, explaining the shifted peaks. As expected, the histogram of the p-values is still right skewed as Figures 20 and 21 show, but compared to MI2, the skewness develops far slower with increasing J which follows analogously to MI1$_+$.

MI3$_+$ Estimator As in the two-sided case MI3$_+$ combines properties of both MI1$_+$ and MI2$_+$. Its overall density is very similar to MI1$_+$ and the density of its false positives (especially when focussing on positive values) is more evenly spaced than either MI1$_+$ or MI2$_+$ as shown in Figures 18 and 19. As in the other cases this is reasonable due to the more even distribution of chosen estimators for the false positives shown in Figure 17. The decomposition by chosen K of the complete set of estimates reported by MI3$_+$ again mimics the one of MI2$_+$ as can be seen in Figure 16.
Its p-values unsurprisingly show a strong inflation of false positives as explained under Table 5. Apart from that its behaviour is similar to the one described for MI1$_+$ and MI2$_+$ with the exception that p-values slightly larger than 0.05 seem to be under-represented compared to MI2$_+$ even if one considers rescaling.[7] A cause for this could be that the selection criteria of MI3$_+$ filter out estimates corresponding to different p-values with different intensity. An explanation could be that due to the correlation in effect size and p-value, these estimates are not likely to be chosen if there is no significant estimate as they likely correspond to smaller discontinuity estimates.

5 Modifications

To find out how modifications of the dgp influence the properties of the MI estimators I generalize Definition (9), so that it allows for heteroskedasticity and a better control of outliers. To do so I specify Definition (18).

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \text{ i.i.d. with } X_i \sim unif(-1,1),\ Y_i \sim \begin{cases} \mathscr{F}_1 & \text{with probability } p_i \\ \mathscr{F}_2, & \text{otherwise} \end{cases} \tag{18}$$
$$\text{where } \mathbb{E}[Y_i] = 0,\ p_i = f(x_i) \in [0,1] \text{ and } X_i \perp Y_i \quad i = 1,2,\ldots,n$$

Now Definition (9) is a special case where $\mathscr{F}_1 = \mathscr{N}(0,1)$ and $p_i = 1 \quad \forall i = 1,\ldots,n$. To estimate parameters correctly for these more general processes it is necessary to use robust standard errors. To do so I used the R libraries *sandwich*[8] and *lmtest*[9].

5.1 Heteroskedastic Error Terms

I generated two additional sets of data representing two opposite forms of heteroskedasticity. In both cases I chose $\mathscr{F}_1 = \mathscr{N}(0,1)$ and $\mathscr{F}_2 = \mathscr{N}(0,5)$ but used different functions to determine the probability p_i. The first data set shows a hourglass pattern as $p_i = f(x_i) = |x_i|$, so the variance increases with increasing distance from the suspected discontinuity. The second data set mimics a diamond shape since $p_i = f(x_i) = 1 - |x_i|$, so the variance is larger in proximity of the suspected discontinuity. Figures for both of these dgps can be found in section 9.6.

[7] This effect can be seen even better for heteroskedastic data as shown in Figure 29.
[8] https://cran.r-project.org/web/packages/sandwich/index.html
[9] https://cran.r-project.org/web/packages/lmtest/index.html

5.2 Adding Outliers as a Special Case of Heteroskedasticity

To add outliers the dgp given in Definition (18) can also be used. Again $\mathscr{F}_1 = \mathscr{N}(0,1)$, but $\mathscr{F}_2 = \mathscr{N}(0,10)$. To generate outliers far away from the suspected discontinuity I chose $p_i = f(x_i) = 0.01|x_i|$ and $p_i = f(x_i) = 0.01(1-|x_i|)$ to generate outliers close to the threshold. Looking at data generated by these processes, a practitioner observing these patterns might not choose to use robust standard errors. Instead, they might assume homoskedastic error terms and outliers for unrelated reasons. Nevertheless, to stay true to the dgp I chose to test this scenario using robust standard errors as well. A different approach might reasonably be assumed and could be studied in further simulations. Figures for both dgps can be found in section 9.7.

5.3 Findings for Heteroskedastic Error Terms

For Section 5.1 — Heteroskedasticity Increasing the variance in the proximity of zero (Diamond-Process) unsurprisingly increases the variance of the underlying OLS estimators, whereas an increase in the variance away from zero (Hourglass-Process) does the same but less strong. Because of that, I am first going to focus on the former. The effects are as expected — both MI1 and MI1$_+$ change slightly, but since the distribution of the p-values of the underlying estimators does not seem to be influenced, the effects are limited to a widening in the reported density estimates. MI2 and MI3 change their behaviour significantly as their selection criterion is directly affected. As Figures 26 and 27 show, the choices are more biased towards estimators with a higher value of K. This is reasonable due to the larger increase in variance exhibited by estimators for higher K. Increasing the variance away from zero (Hourglass-Process) strongly decreases the correlation between estimators of different K, increasing the overall number of false positives generated by the MI estimators as shown in Figure 31 and Table 8.

For Section 5.2 — Outliers As Figures 33 and 38 show, adding outliers to the process as described before also widens the distribution of the underlying estimators. Adding them far away from c reduces correlation between estimators as described for the Hourglass-Process in addition to the effects of widening the underlying distributions. Adding estimators close to zero additionally influences them differently. As the right panel of Figure 33 shows, the p-values do not seem to follow a uniform distribution on $[0,1]$ under the null. Therefore, the selection criteria of MI1 and MI1$_+$ might be influenced. As a comparison of Figures 10 and 11 with 36 and 37 shows, this influence seems to be negligible for both MI1 and MI1$_+$. This hypothesis is further supported as the p-values in Figure 33 seem to be influenced similarly for different K and therefore selection criteria based on these might not necessarily be influenced. Influence on the fraction of false positives seems to be limited as a comparison of the observed fraction of false positives shows. This is interesting as the p-values of the underlying estimators are shifted towards smaller values, indicating that this effect might be weak for p-values that are smaller than 0.05.

6 Conclusion, Possible Modifications and Extensions

Overall the findings of this thesis illustrate the problems of multiple testing and some aspects of p-hacking in the context of RD designs. I do not expect that real scientists use the MI estimators I envisioned, but as they mimic patterns of dishonest behaviour incentivised by the same structures found in the real world, they might tell something about possible distortions in scientifically published results. An inflated number of false positives, unreasonably high effect sizes and equally unreasonable p-values as shown in this simulation can be the consequence. In the last years meta analyses and publications that take an even broader look at published science like Alexander A. Aarts et al. (2015) showed that these patterns can also be found in reality.

The underlying idea of this thesis does not only apply to RD designs and there are many more ways to game the system than choosing different functional forms, like post-hoc formulation and testing of multiple hypotheses, or choosing data subsets that fit pre-existing ideas etc. This thesis might motivate critical thought on publication bias and the incentives created by the current state of publishing, as well as the steps scientists might take due to these conditions. A factor contributing to the problem is the p-value which is used extensively in this thesis (as well as nearly every quantitative science). While not problematic in itself, relying on this statistic alone is — even without considering outright incorrect application or false interpretation which is also common. For more information on this Wasserstein and Lazar (2016) and Wasserstein, Schirm, and Lazar (2019) are good places to start as these address the most common forms of misunderstanding and malpractice concerning the use of p-values, as well as giving advice on how to avoid them.

Coming back to RD designs, the findings of this thesis are probably not limited to the specific estimators I constructed. There are several reasonable ways to extend on the findings presented which I addressed shortly in the main text. I am going to elaborate more on some of those in the following.

Further Modification of the MI Estimators One possible modification to the MI3 and MI3$_+$ Estimators would be to construct them in a more hierarchical way. By hierarchical I mean, to not only include significance at a level of 95% in the estimators, but to also include significance at 90% and 99%. I propose one such estimator in section 9.4. To obtain useful information about its distribution one would probably need to use a larger simulation, as the density estimates for the highly significant false positives would be very imprecise for a small number of batches. It would also be possible to change the priorities of the estimators, giving more weight to positivity or other factors.

Approximation of the Power Function Given more time and computational resources it would be possible to add a third dimension to this simulation. Where this thesis only addressed the case of the null hypothesis being correct, it would also be interesting to see how the MI

estimators behave if there really is a discontinuity in the dgp. Simulating b data sets as shown before (each consisting of m batches of n observations generated by the same dgp) for different τ could be used to approximate the power function of the MI estimators. This could be interesting as even though the null $\tau = 0$ is technically incorrect in many of these scenarios, the MI estimators might still lead to an inflated number of rejections. These could be studied further to obtain even better knowledge on the effects of covert multiple testing.

Changes in the position of the Threshold If the choice of c, meaning the position of the threshold, is not fixed by theory another way to game the system in RD analysis might be to perform this analysis for many choices of c. As $c = 0$ is not special in the context of the homoskedastic dgp used in the main part of this thesis, one could calculate the underlying OLS estimators for a multitude of c and only report the findings that perform best in a publishing context. This further increases the problems found in this thesis and could be studied in more detail in another simulation.

Linear or Polynomial Data Generating Processes Another modification could be made to the dgps used in this thesis. Instead of assuming that X_i itself has no direct influence on Y_i, there are many possible specifications one could use to generate data. Examples include linear or polynomial processes of varying specification. This would enable a similar analysis as presented in this thesis. To keep the unbiased nature of the estimators, one would have to include only estimators of equal or higher order K than the polynomial order of the dgp. There are many more possible dgps which could require different estimators, but due to the vast nature of possibilities I will refrain from naming specific examples.

Nonparametric Discontinuity Estimators In this thesis I focused on parametric approaches to estimating discontinuities. From a theoretical standpoint it is often more informative to take a closer look at observations near the threshold, since individuals there might be more closely comparable. Therefore, it might be unwise to include observations far away from the threshold when searching for a discontinuity. This motivates the use of estimators such as the Nadaraya-Watson estimator[10] or local-linear estimator. Due to the nature of the dgp used in my analysis, these estimators would perform optimally for bandwidths that would counteract the local approach. By this I mean, that cross-validation would choose a bandwidth that is large enough for these estimators to become equal to their corresponding global counterparts. Choosing a different setting, for example a non-linear dgp, might be an interesting possibility to extend the analysis presented in this thesis. One could employ methods such as cross validation to obtain a good estimate for which bandwidth to choose in further analyses. It would also be possible to use a dishonest cross-validation criterion to create a different kind of non-parametric MI estimator that chooses its bandwidth to create estimates that fit a specific narrative, creating further possibilities for dishonest behaviour.

[10] Figure 9 shows an estimate created by the Nadaraya-Watson Estimator with a rectangular kernel and a bandwidth of 0.1.

7 Bibliography

References

Alexander A. Aarts et al. (2015). "PSYCHOLOGY. Estimating the reproducibility of psychological science". In: *Science (New York, N.Y.)* 349.6251, aac4716. DOI: 10.1126/science.aac4716.

Dunn, Olive Jean (1961). "Multiple Comparisons Among Means". In: *Journal of the American Statistical Association* 56.293, p. 52. ISSN: 01621459. DOI: 10.2307/2282330.

Guido W. Imbens and Thomas Lemieux (2008). "Regression discontinuity designs: A guide to practice". In: *Journal of Econometrics* 142.2, pp. 615–635. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2007.05.001.

Hayashi, Fumio (2000). *Econometrics*. Princeton: Princeton University Press. ISBN: 0691010188.

Lee, David S. and Thomas Lemieux (2010). "Regression Discontinuity Designs in Economics". In: *Journal of Economic Literature* 48.2, pp. 281–355. ISSN: 0022-0515. DOI: 10.1257/jel.48.2.281.

Shonkwiler, Ronald W. and Franklin Mendivil (2009). *Explorations in Monte Carlo methods*. Undergraduate texts in mathematics. Dordrecht and New York: Springer. ISBN: 978-0-387-87836-2.

Thistlethwaite, Donald L. and Donald T. Campbell (1960). "Regression-discontinuity analysis: An alternative to the ex post facto experiment". In: *Journal of Educational Psychology* 51.6, pp. 309–317. ISSN: 0022-0663. DOI: 10.1037/h0044319.

Wasserstein, Ronald L. and Nicole A. Lazar (2016). "The ASA Statement on p -Values: Context, Process, and Purpose". In: *The American Statistician* 70.2, pp. 129–133. DOI: 10.1080/00031305.2016.1154108.

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar (2019). "Moving to a World Beyond " $p < 0.05$"". In: *The American Statistician* 73.sup1, pp. 1–19. DOI: 10.1080/00031305.2019.1583913.

8 Schriftliche Versicherung

Ich versichere hiermit, dass ich die vorstehende Bachelorarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass die vorgelegte Arbeit noch an keiner anderen Hochschule zur Prüfung vorgelegt wurde und dass sie weder ganz noch in Teilen bereits veröffentlicht wurde. Wörtliche Zitate und Stellen, die anderen Werken dem Sinn nach entnommen sind, habe ich in jedem einzelnen Fall kenntlich gemacht.

Bonn, den .06.2020 _____

Jakob Ralph Jürgens

9 Appendix

List of Figures in the Appendix

List of Tables in the Appendix

9.1 Additional General Figures

Figure 6: Maximum discontinuities estimated by different polynomial OLS specifications

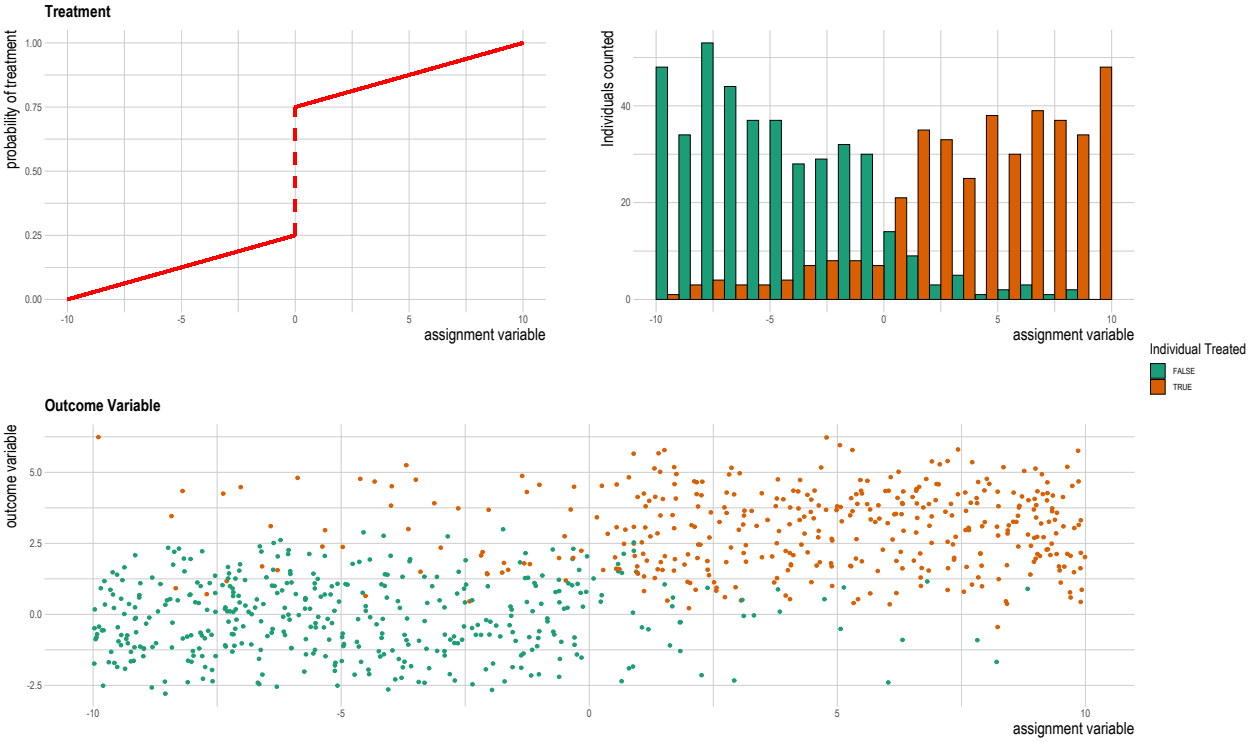

Figure 7: Visualisation for Fuzzy Regression Discontinuity

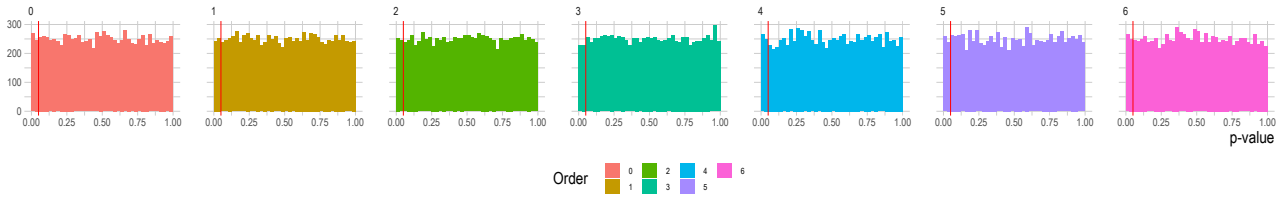

Figure 8: Histograms for p-values by order of polynomial

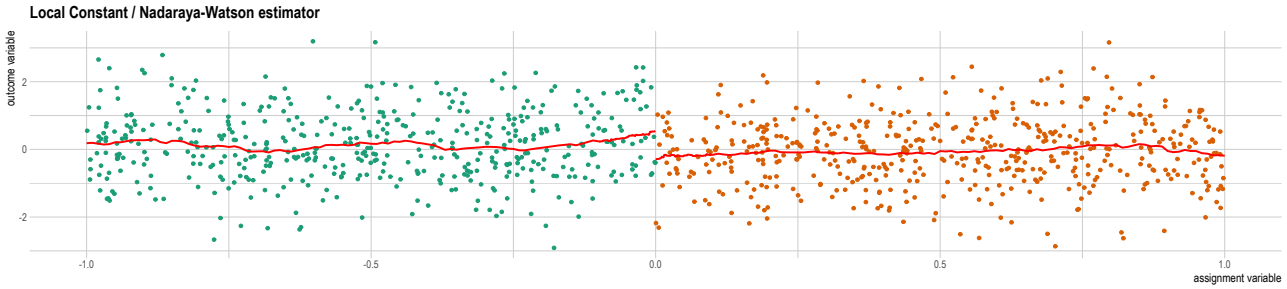

Figure 9: Maximum Discontinuity estimated using the Nadaraya-Watson Estimator

9.2 Additional Figures for MI Estimators

Figure 10: Which Estimators do the MI Estimators choose?

Figure 11: Which Estimators do the MI Estimators choose? (Only for significant results)

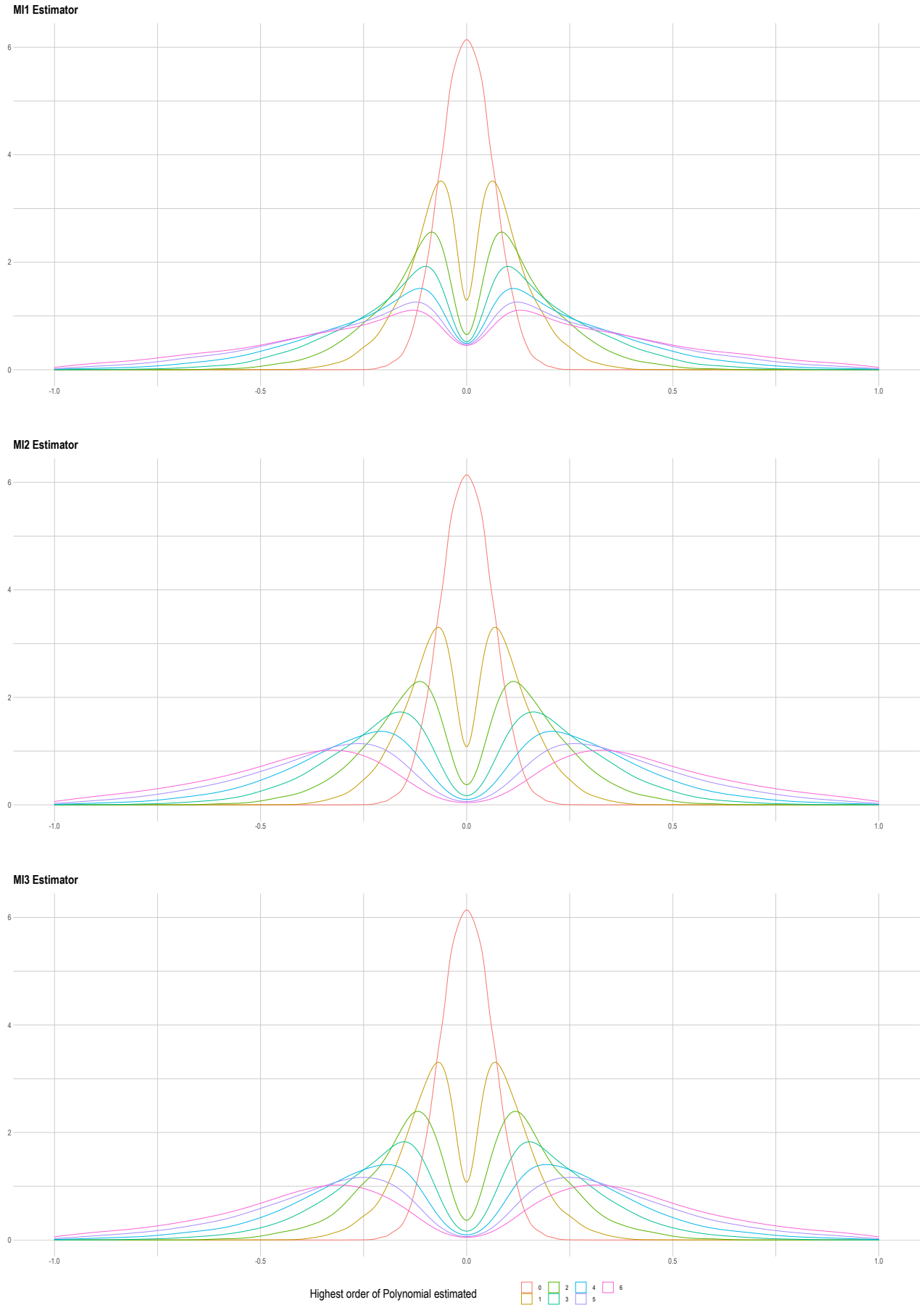

Figure 12: Estimated Densities for different orders of highest Polynomial estimated

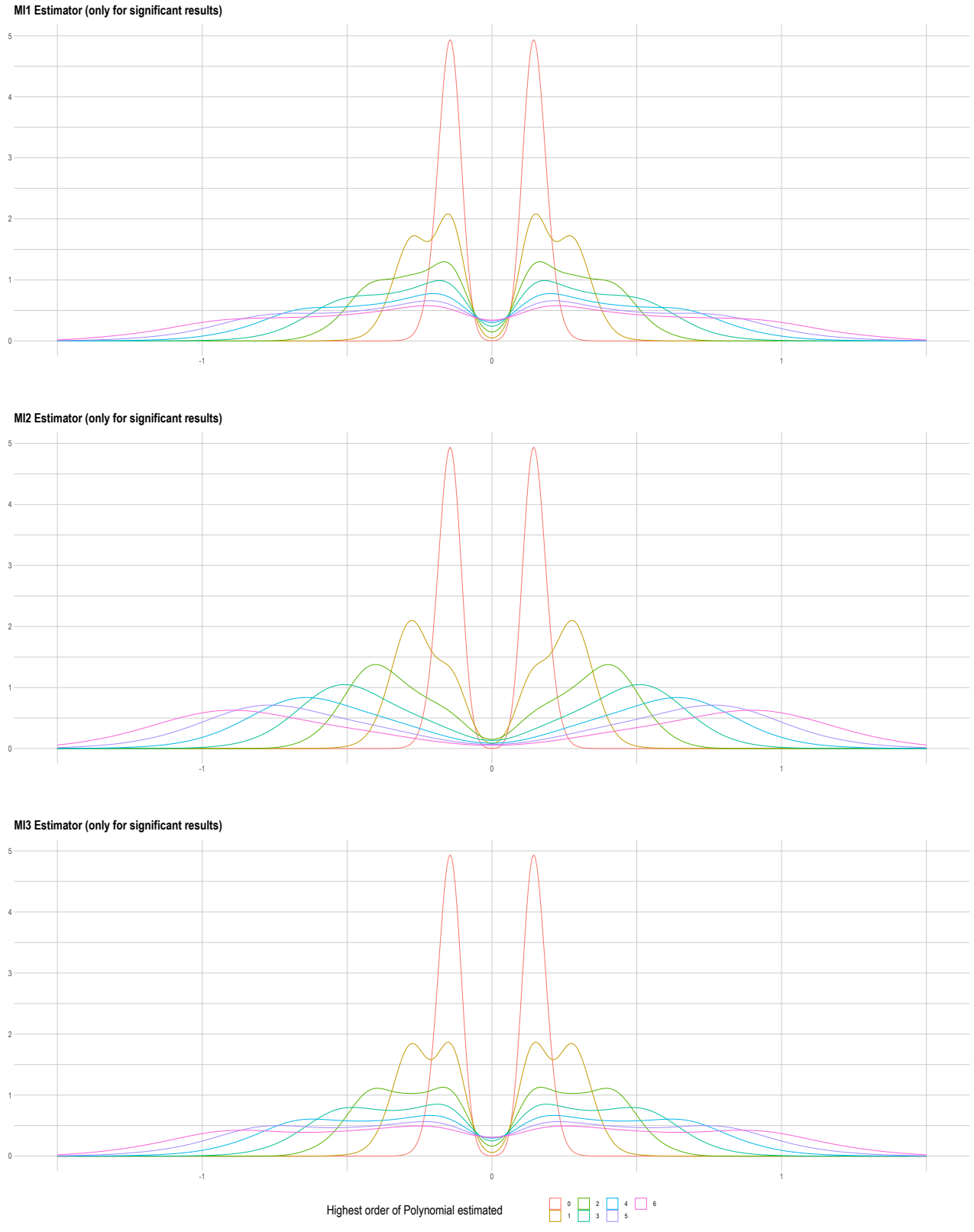

Figure 13: Estimated Densities for different orders of highest Polynomial estimated (only for significant estimates)

In both cases I use that, due to the setup, the density has to be symmetrical to obtain a better estimate.

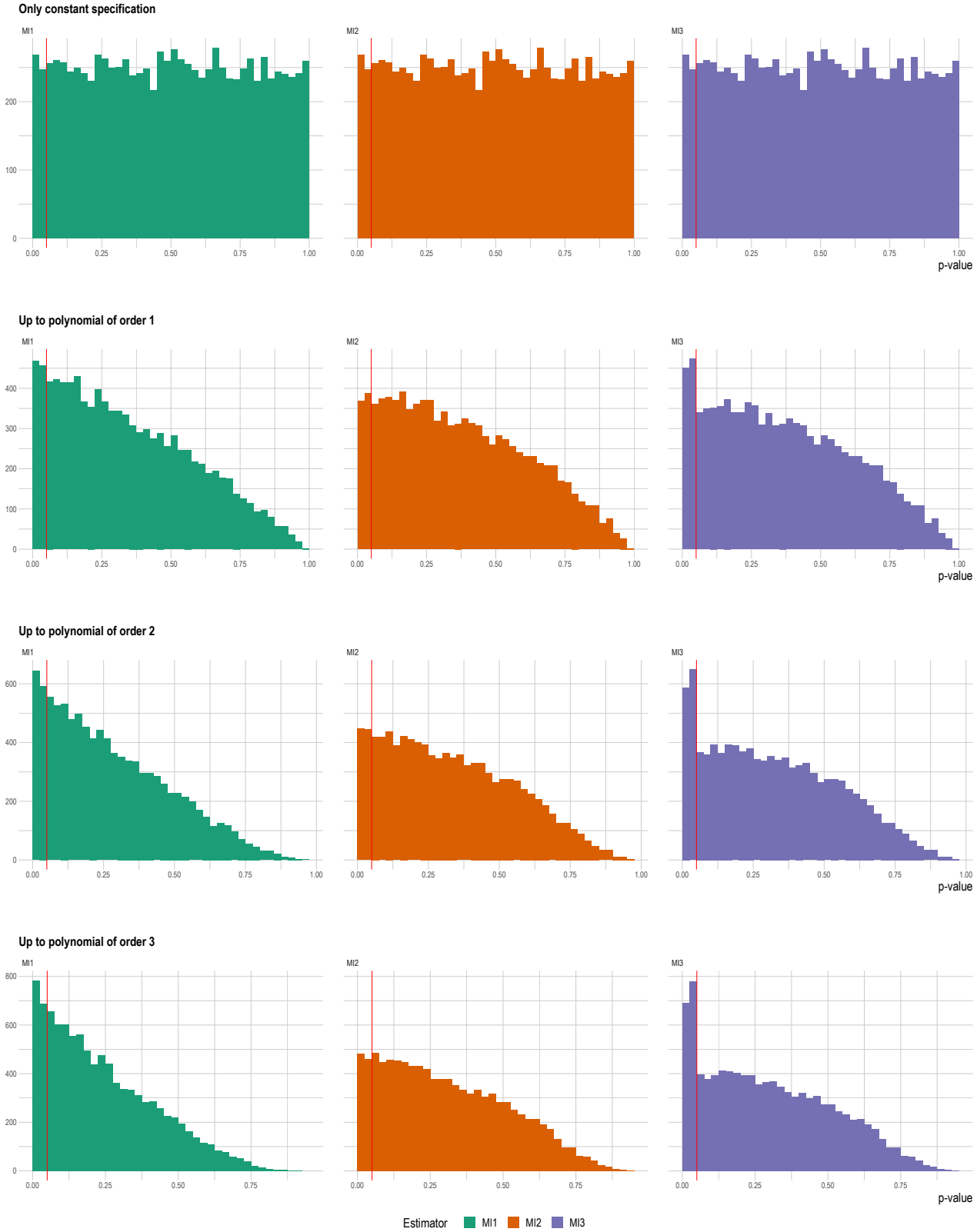

Figure 14: Histograms of p values for different orders of highest Polynomial estimated (0-3)

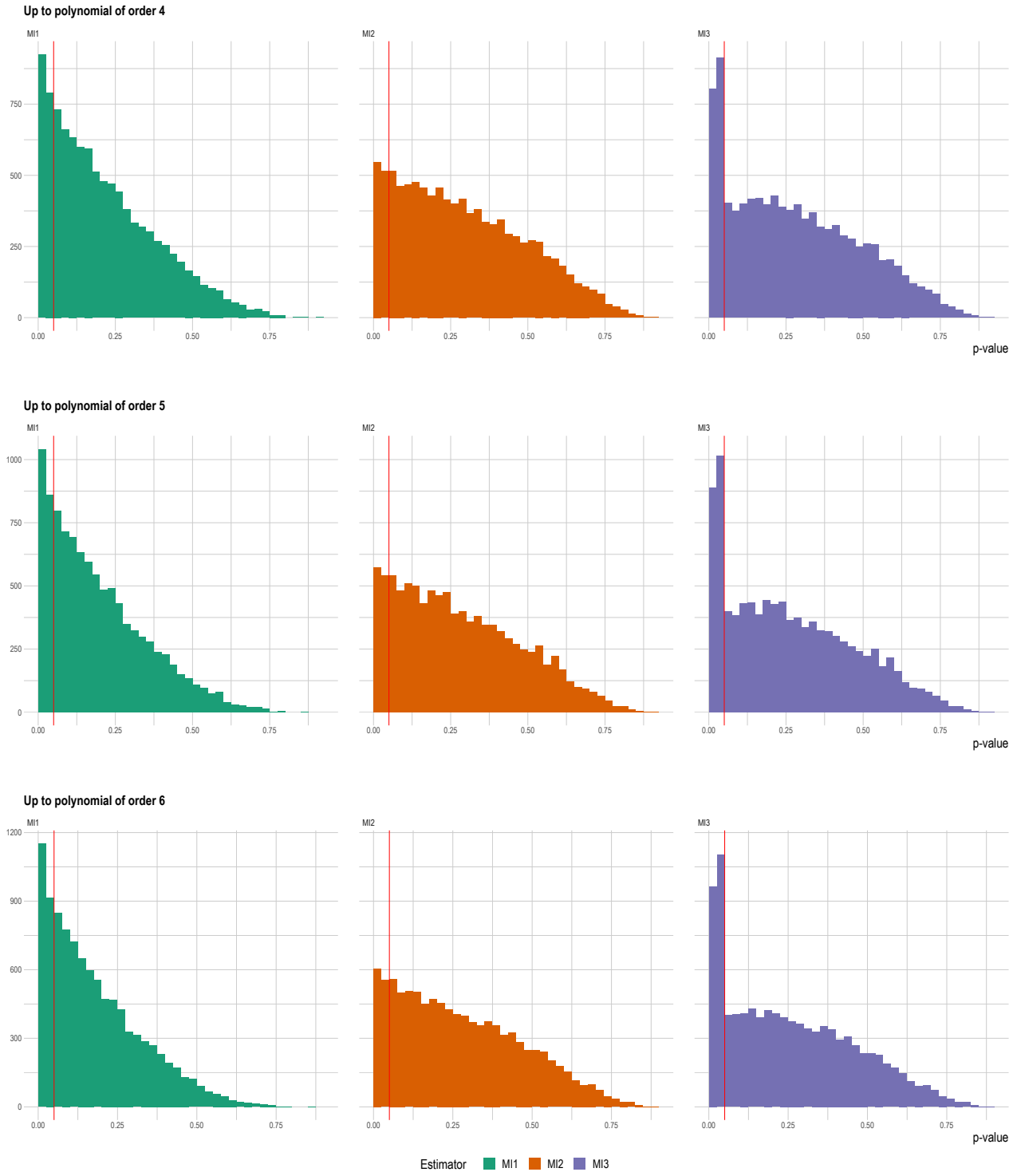

Figure 15: Histograms of p values for different orders of highest Polynomial estimated (4-6)

9.3 Additional Figures for One-sided MI Estimators

Figure 16: Which Estimators do the One-Sided MI Estimators choose?

Figure 17: Which Estimators do the One-Sided MI Estimators choose? (Only for significant results)

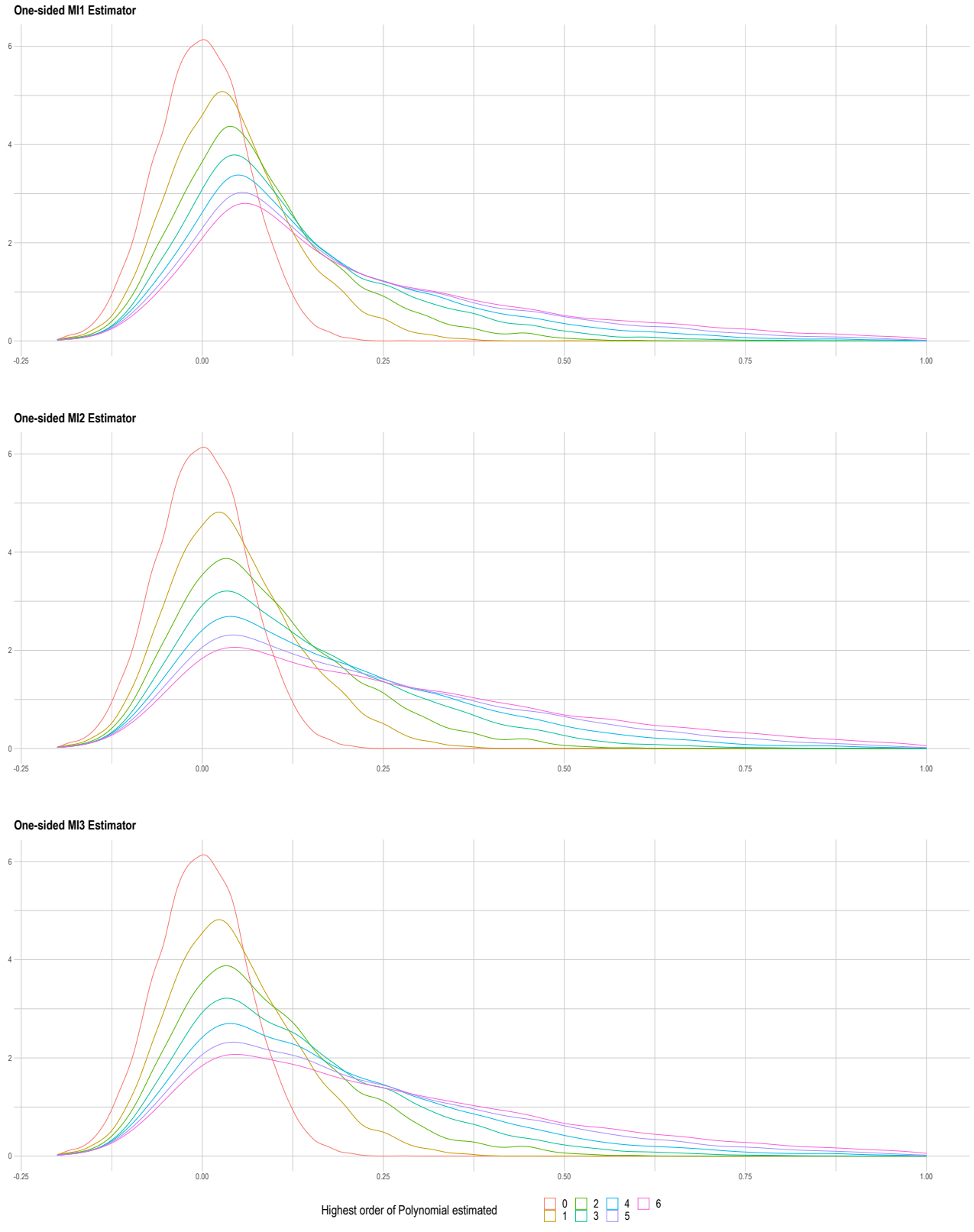

Figure 18: Estimated Densities for different orders of highest Polynomial estimated

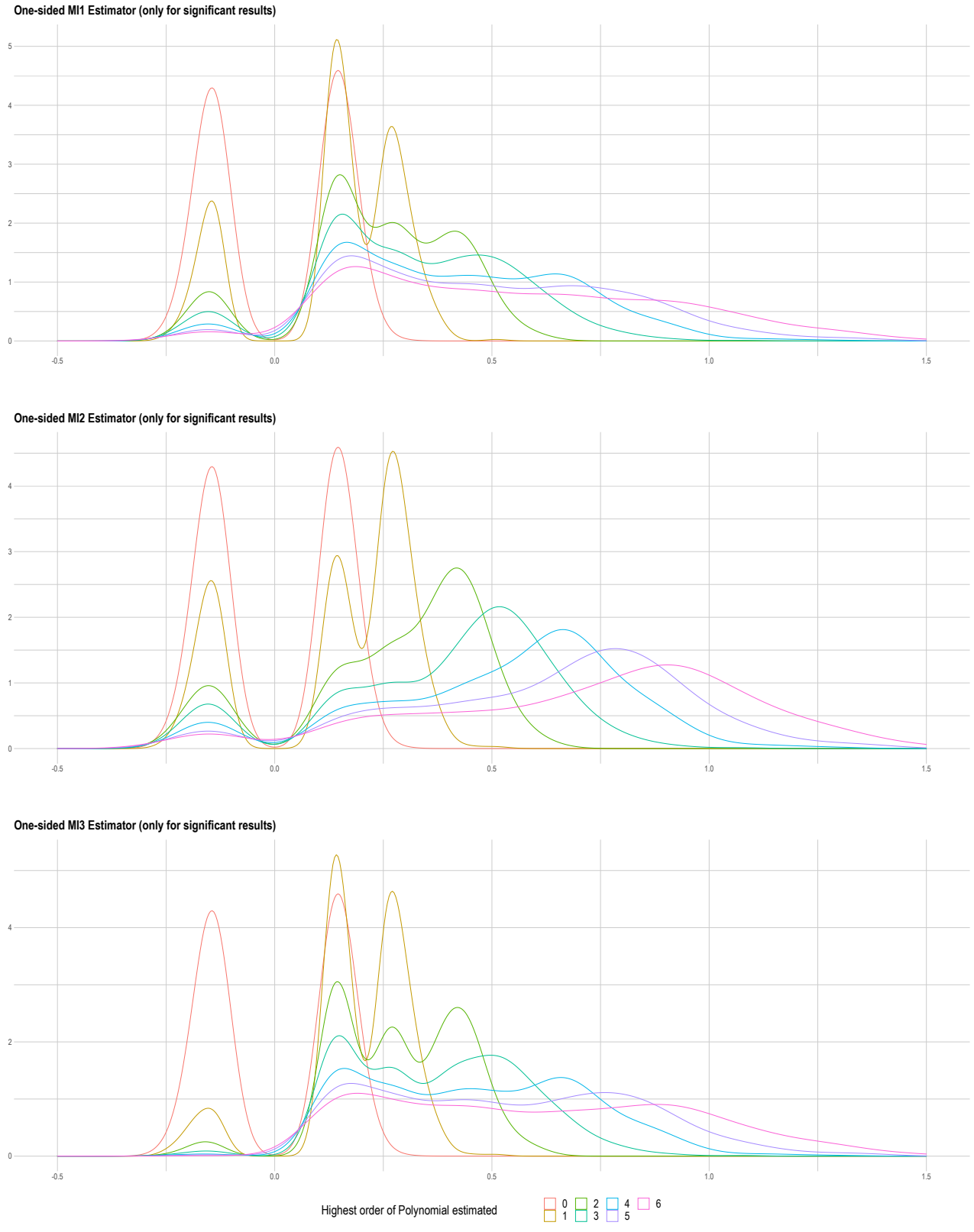

Figure 19: Estimated Densities for different orders of highest Polynomial estimated (only for significant estimates)

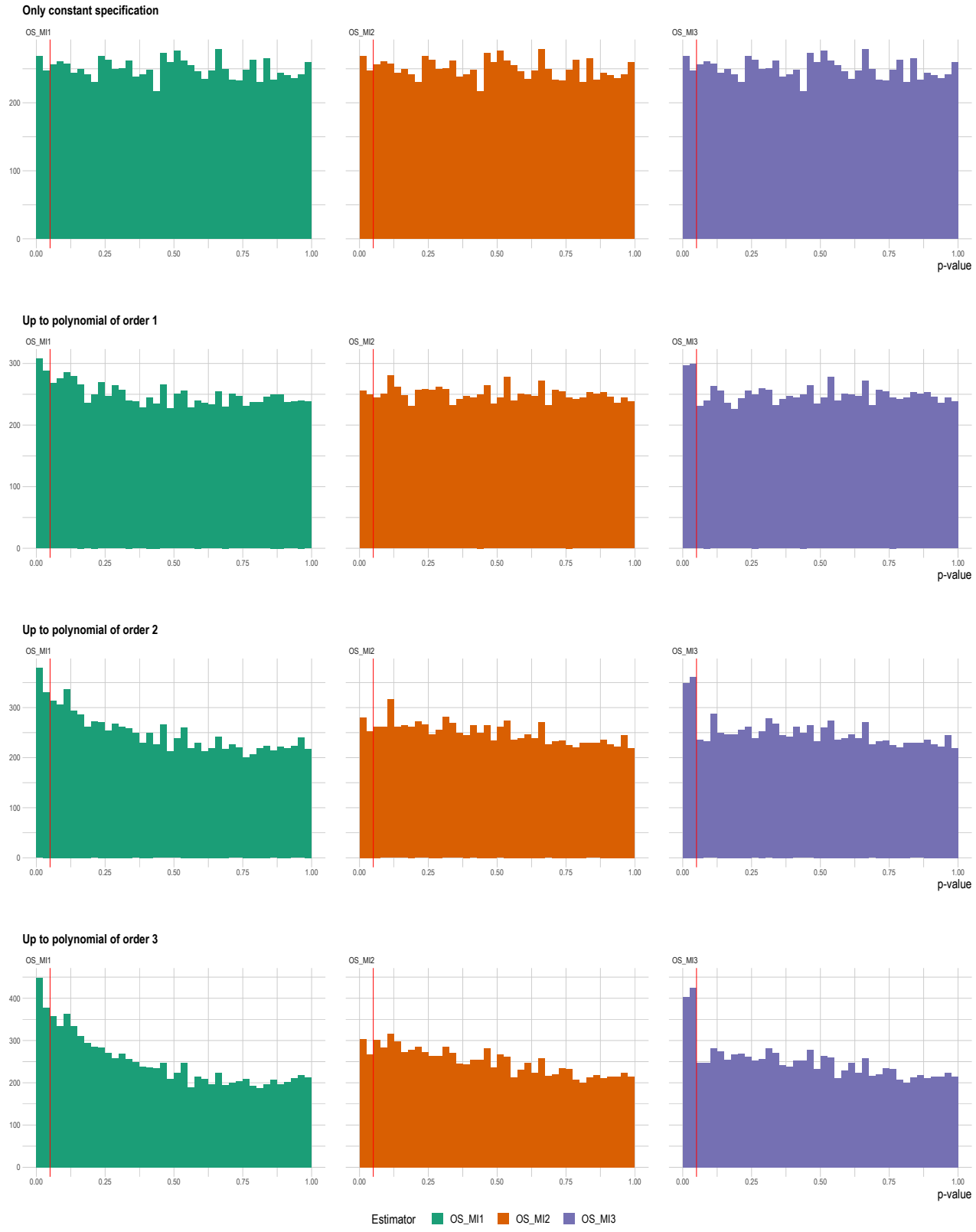

Figure 20: Histograms of p values for different orders of highest Polynomial estimated (0-3)

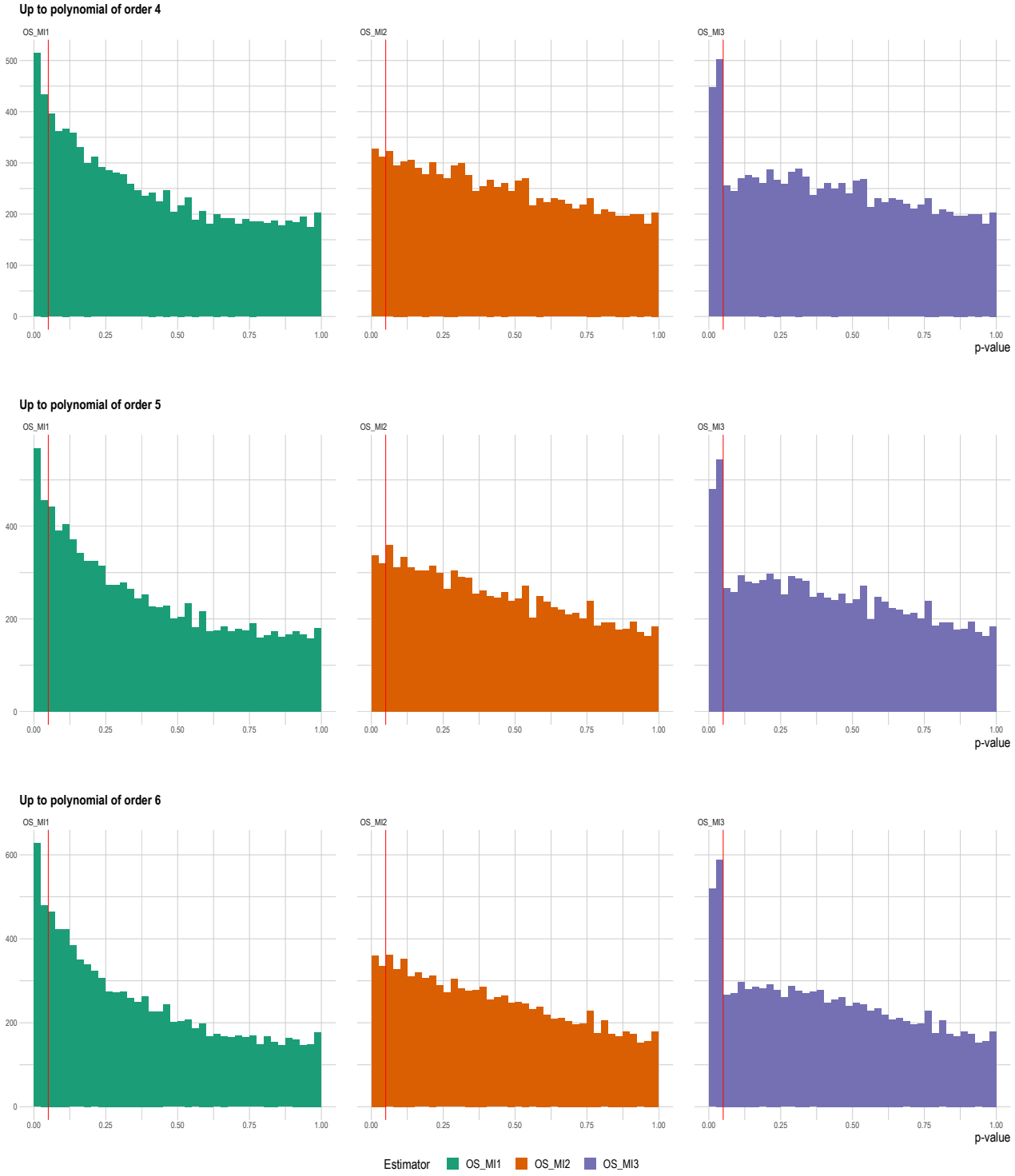

Figure 21: Histograms of p values for different orders of highest Polynomial estimated (4-6)

9.4 Hierarchical MI3

$$\mathbf{MI3}_{Hir}\left(\{(\hat{\tau}_k, p_k)\,; k = 1,\ldots,K\}\right) = \begin{cases} (\hat{\tau}_h, p_h), & \text{if } \{p_k \,|\, p_k \leq 0.1 \text{ for } k = 1,\ldots,K\} = \emptyset \\ (\hat{\tau}_i, p_i), & \text{if } \{p_k \,|\, p_k \leq 0.1 \text{ for } k = 1,\ldots,K\} \neq \emptyset \\ & \text{and } \{p_k \,|\, p_k \leq 0.05 \text{ for } k = 1,\ldots,K\} = \emptyset \\ (\hat{\tau}_j, p_j), & \text{if } \{p_k \,|\, p_k \leq 0.05 \text{ for } k = 1,\ldots,K\} \neq \emptyset \\ & \text{and } \{p_k \,|\, p_k \leq 0.01 \text{ for } k = 1,\ldots,K\} = \emptyset \\ (\hat{\tau}_l, p_l), & \text{otherwise} \end{cases} \tag{19}$$

$$\text{where} \quad l = \underset{k\in\{1,\ldots,K\} \text{ with } p_k \leq 0.01}{\arg\max} |\hat{\tau}_k|, \quad j = \underset{k\in\{1,\ldots,K\} \text{ with } p_k \leq 0.05}{\arg\max} |\hat{\tau}_k|,$$
$$i = \underset{k\in\{1,\ldots,K\} \text{ with } p_k \leq 0.1}{\arg\max} |\hat{\tau}_k| \quad \text{and} \quad h = \underset{k}{\arg\max} |\hat{\tau}_k|$$

9.5 Outline Estimation of Joint Distribution

In the following I will present an outline for estimation of the joint distribution for the discontinuity estimates of different polynomial specification. I am not going to elaborate on the necessary prerequisites for OLS estimation in the given context.

Let $\hat{\beta}_j \in \mathbb{R}^{2(j+1)}$ denote the vector of estimates created by the estimator with polynomial specification of order j. (The constant estimator is therefore denoted as $\hat{\beta}_0 \in \mathbb{R}^2$.) The following relationship is known from standard OLS theory and since there is no misspecification of the functional form (among other things) should hold here:

$$\sqrt{n}(\hat{\beta}_j - \beta_j) \to_d \mathcal{N}(0, \Sigma_j) \tag{20}$$

Since Σ_j is unknown we estimate it using

$$\hat{\Sigma}_j = (X_j^T X_j)^{-1} \hat{\sigma}_j^2 \tag{21}$$

with the usual notation.

To test the significance of these estimates we can also use the usual approach. Let $\hat{\beta}_{i,j}$ denote the i-th entry of $\hat{\beta}_j$. Assuming that the estimate for the discontinuity is always the second entry, this means that $\hat{\beta}_{2,j} := \hat{\tau}_j$. As a similar convention let $\hat{\sigma}^2_{i,j}$ denote the i-th diagonal element of $\hat{\Sigma}_j$ meaning the estimated variance of $\hat{\beta}_{i,j}$.

If we now stack the vectors $\hat{\beta}_j \quad \forall j = 0, 1, \ldots, J$, we obtain

$$\sqrt{n}\left(\hat{\beta}_{J,stacked} - \beta_{J,stacked}\right) = \sqrt{n}\left(\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_J\end{pmatrix} - \begin{pmatrix}0\\ 0\\ \vdots\\ 0\end{pmatrix}\right) \to_d \mathcal{N}\left(0, \begin{pmatrix}\Sigma_0 & \Sigma_{0,1} & \ldots & \Sigma_{0,J}\\ \Sigma_{1,0} & \Sigma_1 & \ldots & \Sigma_{1,J}\\ \vdots & \vdots & \ddots & \vdots\\ \Sigma_{J,0} & \Sigma_{J,1} & \ldots & \Sigma_J\end{pmatrix}\right) \tag{22}$$

Where all remaining components $\Sigma_{k,l}$ can be estimated.

$$\hat{\tau}_{J,stacked} := \begin{pmatrix}\hat{\beta}_{2,0}\\ \hat{\beta}_{2,1}\\ \vdots\\ \hat{\beta}_{2,J}\end{pmatrix} = A_\tau \begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_J\end{pmatrix} \to_d \mathcal{N}(0, \Sigma_\tau) \tag{23}$$

where $A_\tau \in \mathbb{R}^{(J+1)\times(J+1)(J+2)}$ is a matrix that consists only of zeroes, except for one entry in each row corresponding to the discontinuity estimate that is equal to 1. For the j-th row, this entry is in the k-th column, where $k = j^2 - j + 2$. Therefore we obtain the theoretical result, that

$$\Sigma_\tau = A_\tau \begin{pmatrix}\Sigma_0 & \Sigma_{0,1} & \ldots & \Sigma_{0,J}\\ \Sigma_{1,0} & \Sigma_1 & \ldots & \Sigma_{1,J}\\ \vdots & \vdots & \ddots & \vdots\\ \Sigma_{J,0} & \Sigma_{J,1} & \ldots & \Sigma_J\end{pmatrix} A_\tau^T \quad \approx \quad \hat{\Sigma}_\tau = A_\tau \begin{pmatrix}\hat{\Sigma}_0 & \hat{\Sigma}_{0,1} & \ldots & \hat{\Sigma}_{0,J}\\ \hat{\Sigma}_{1,0} & \hat{\Sigma}_1 & \ldots & \hat{\Sigma}_{1,J}\\ \vdots & \vdots & \ddots & \vdots\\ \hat{\Sigma}_{J,0} & \hat{\Sigma}_{J,1} & \ldots & \hat{\Sigma}_J\end{pmatrix} A_\tau^T \tag{24}$$

Approximation for large number of observations. This result could be used to construct less conservative confidence intervals for the type of multiple testing analysed in this thesis. To do so, it is useful to observe that

$$\max_j |\hat{\beta}_{i,j}| \leq t \iff |\hat{\beta}_{i,j}| \leq t \quad \forall j = 0, 1, \ldots, J \tag{25}$$

With $\hat{\Sigma}_\tau$ as an estimate of Σ_τ, we can use the properties of this distribution to calculate a value of t, such that

$$P(\max_j |\hat{\beta}_{2,j}| \leq t) = P(|\hat{\tau}_{J,stacked}| \leq t\vec{1}_{J+1}) = 1 - \alpha, \quad \text{where } \vec{1}_{J+1} = \begin{pmatrix}1\\ 1\\ \vdots\\ 1\end{pmatrix} \in \mathbb{R}^{J+1} \tag{26}$$

Giving us a confidence interval for the given confidence level of α.

9.6 Additional Figures and Tables for Heteroskedasticity

Diamond-Process The following Tables and Figures were created for the process with a diamond-like shape described in section 5.1.

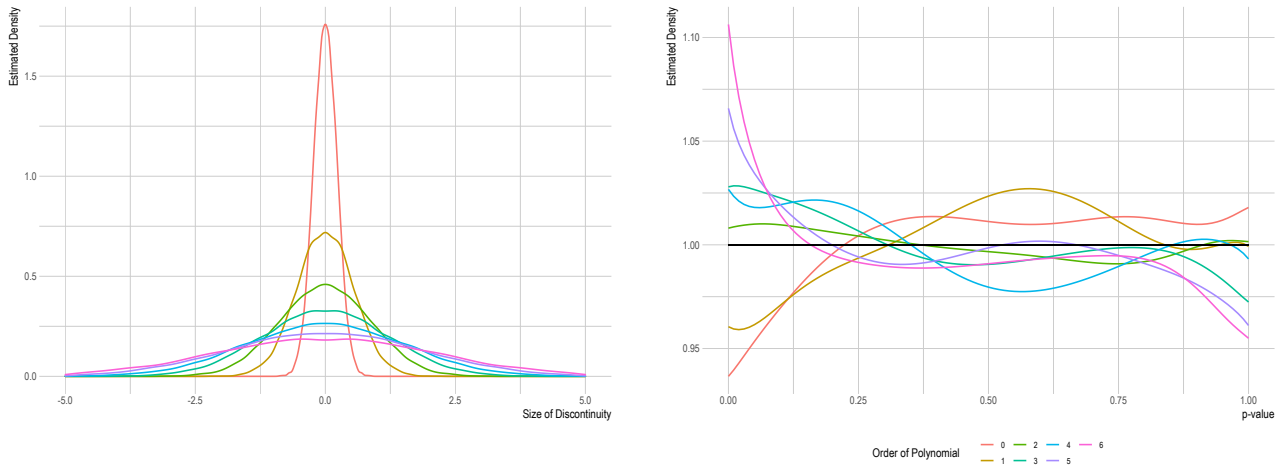

Figure 22: Estimated Densities for Discontinuities and p-values by order of Polynomial (Diamond-process)

K	0	1	2	3	4	5	6
correlation coefficient	-0.949	-0.947	-0.946	-0.946	-0.942	-0.936	-0.93

Table 6: Correlation between $|\hat{\tau}|$ and their corresponding p-values for different K (Diamond-process)

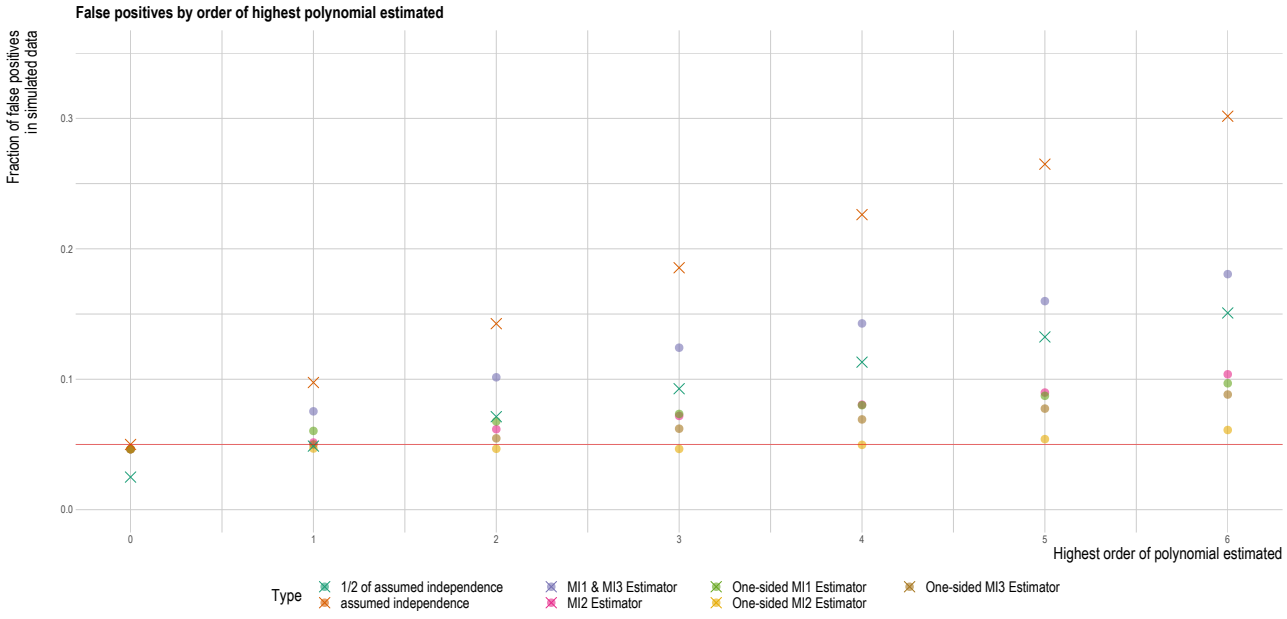

Figure 23: Fraction of False Positives by Maximum Order of Polynomial estimated (Diamond-process)

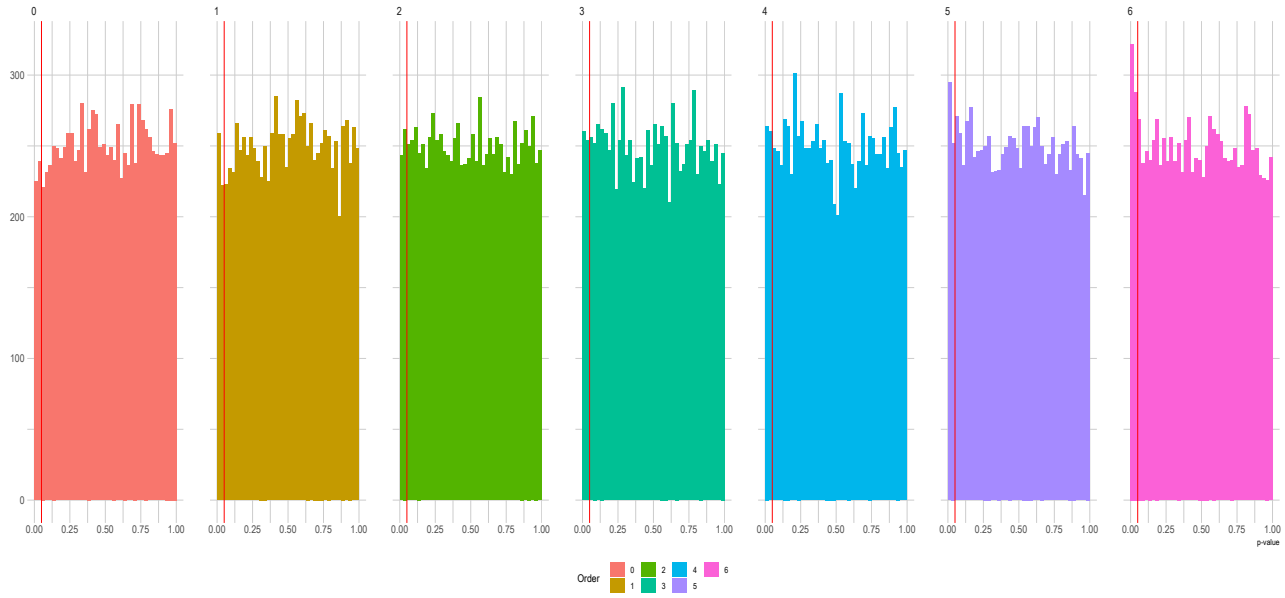

Figure 24: Histograms for p-values by order of polynomial (Diamond-process)

	0	1	2	3	4	5	6
0	1	0.474	0.146	0.0643	0.0377	0.0286	0.0222
1	0.783	1	0.55	0.252	0.149	0.101	0.0715
2	0.481	0.828	1	0.64	0.357	0.225	0.156
3	0.347	0.603	0.872	1	0.694	0.403	0.266
4	0.268	0.472	0.686	0.894	1	0.721	0.447
5	0.213	0.384	0.562	0.732	0.909	1	0.75
6	0.178	0.322	0.471	0.615	0.766	0.92	1

Table 7: Correlation between $|\hat{\tau}|$ for different K (purple) and p-values of different K (green) (Diamond-process)

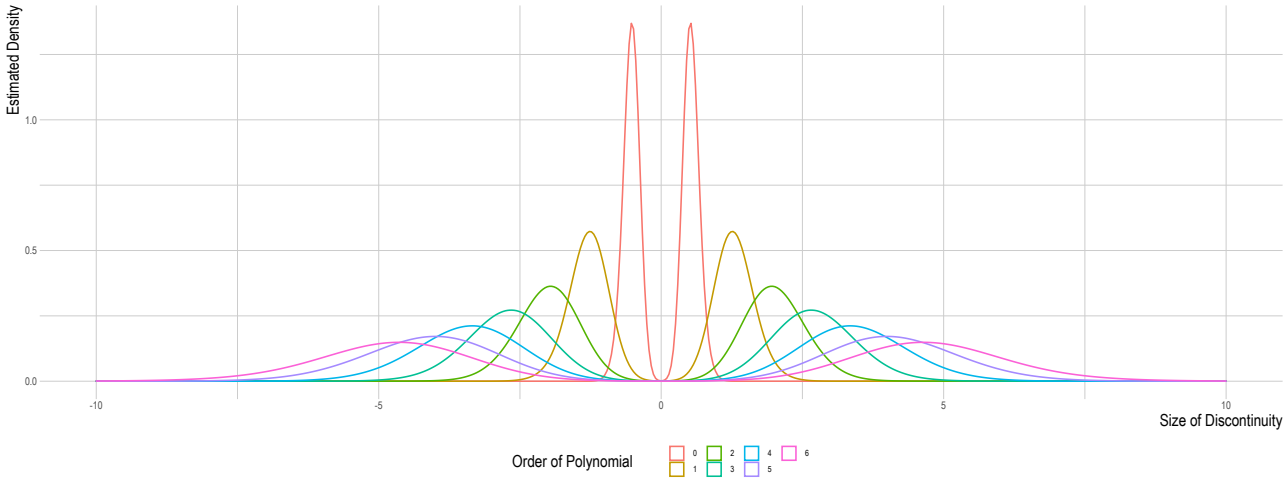

Figure 25: Estimated Densities for significant Discontinuity Estimates (Diamond-process)

Figure 26: Which Estimators do the MI Estimators choose? (Diamond-process)

Figure 27: Which Estimators do the MI Estimators choose? (Only for significant results) (Diamond-process)

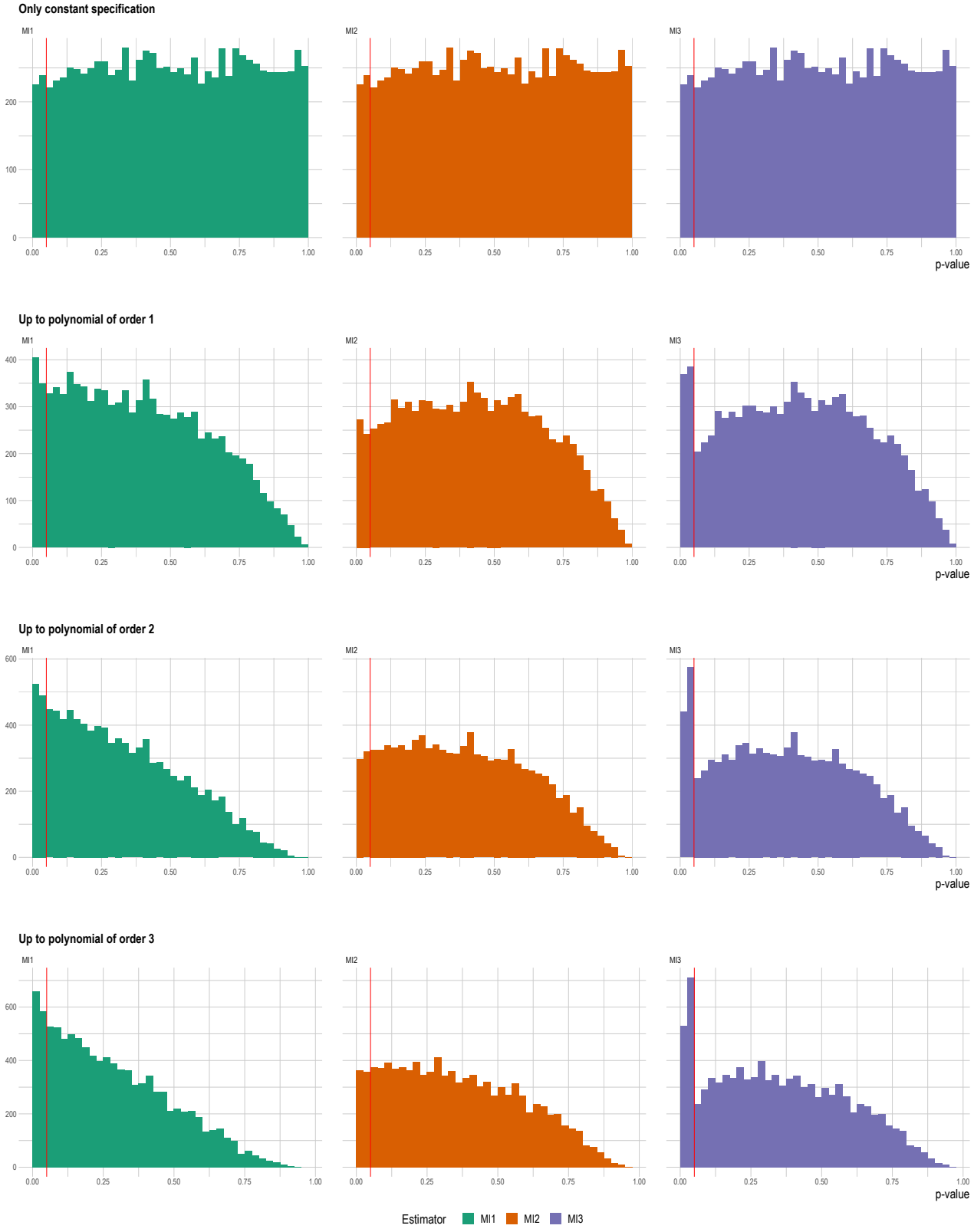

Figure 28: Histograms of p values for different orders of highest Polynomial estimated (0-3) (Diamond-process)

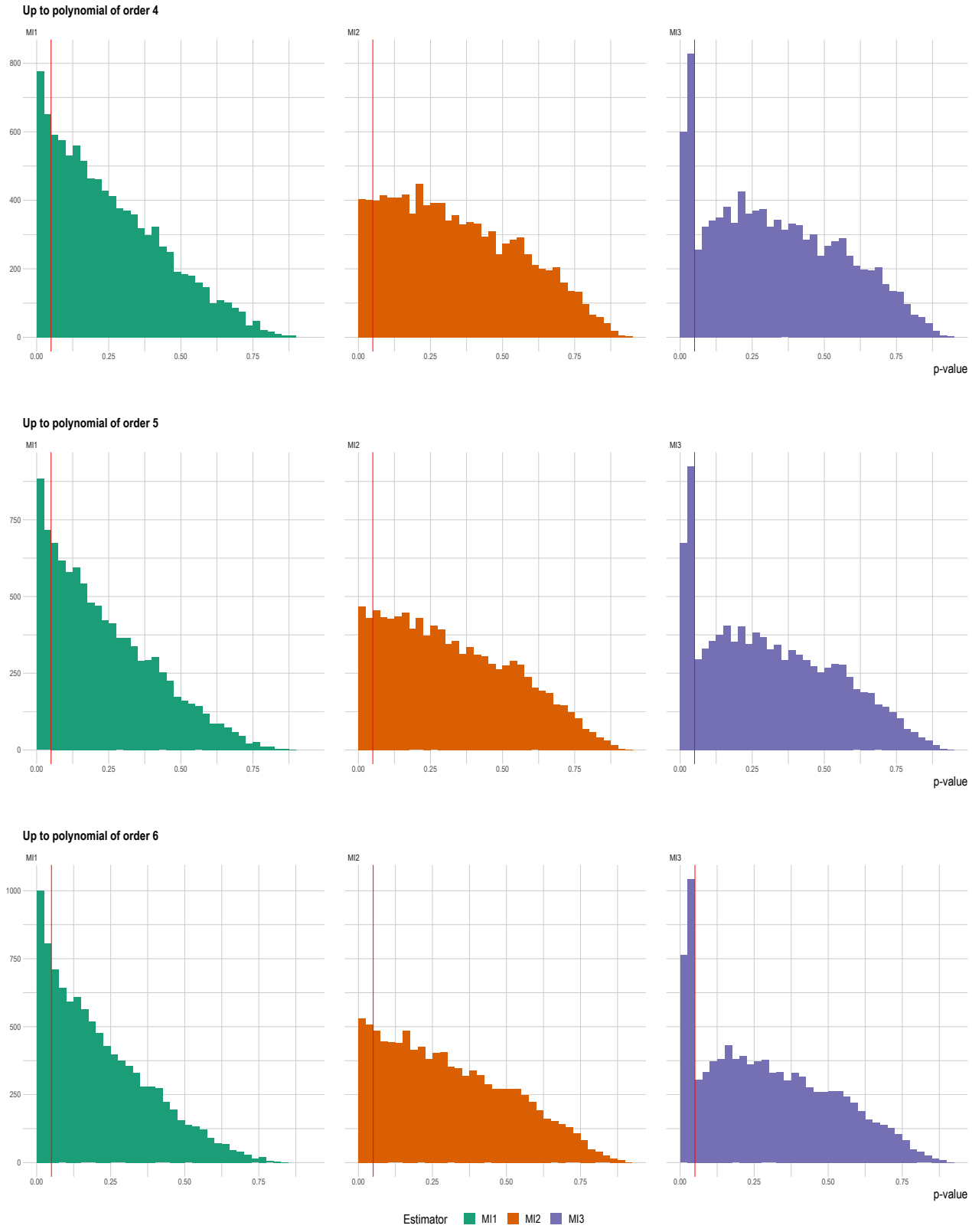

Figure 29: Histograms of p values for different orders of highest Polynomial estimated (4-6) (Diamond-process)

Hourglass-Process The following Tables and Figures were created for the process with an hourglass-like shape described in section 5.1.

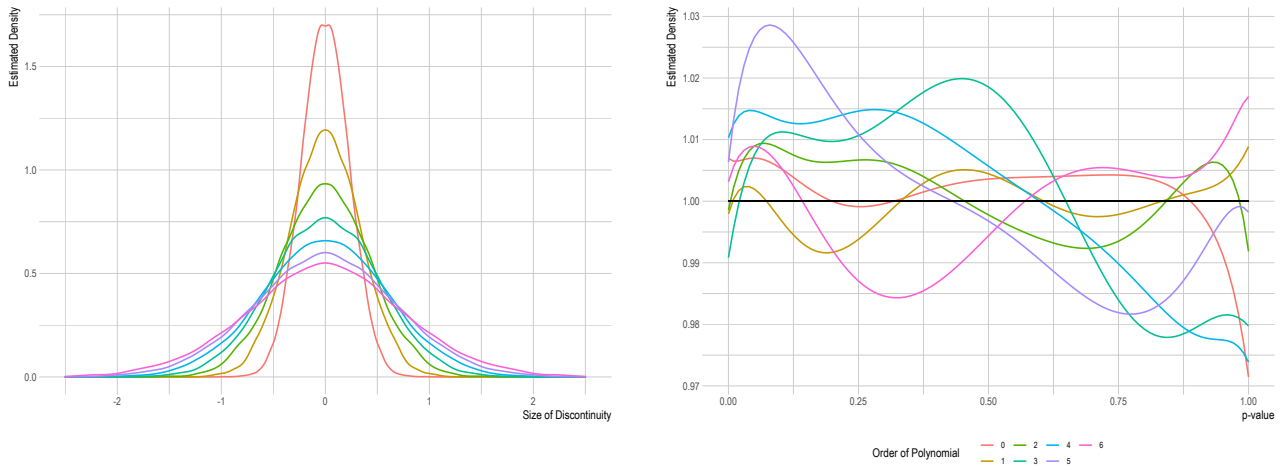

Figure 30: Estimated Densities for Discontinuities and p-values by order of Polynomial (Hourglass-process)

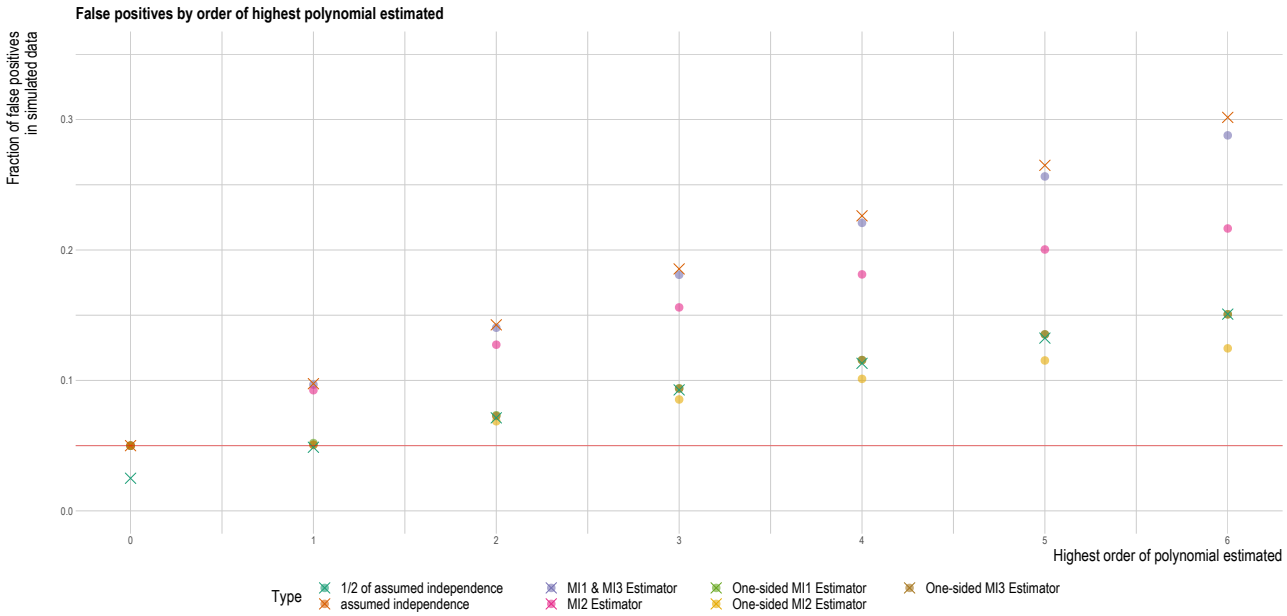

Figure 31: Fraction of False Positives by Maximum Order of Polynomial estimated (Hourglass-process)

	0	1	2	3	4	5	6
0	1	0.014	-0.0167	0.00748	-0.000477	-0.016	0.00258
1	0.0437	1	0.00567	0.00282	0.00545	0.0032	0.00687
2	0.0344	0.102	1	0.00618	0.00817	-0.00444	0.00219
3	0.0485	0.0888	0.171	1	0.0232	0.0293	0.0101
4	0.0227	0.0618	0.122	0.207	1	0.0472	0.0152
5	0.0271	0.0683	0.127	0.202	0.226	1	0.033
6	0.0249	0.045	0.107	0.173	0.209	0.283	1

Table 8: Correlation between $|\hat{\tau}|$ for different K (purple) and p-values of different K (green) (Hourglass-process)

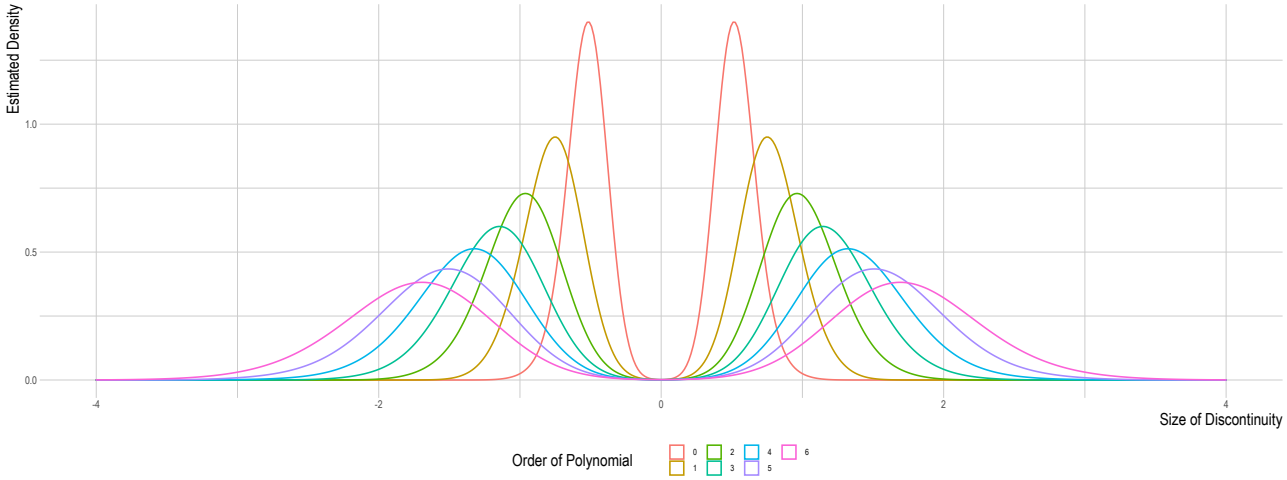

Figure 32: Estimated Densities for significant Discontinuity Estimates (Hourglass-process)

9.7 Additional Figures and Tables for Outliers

Outliers close to c The following Tables and Figures were created for the process with outliers close to c in section 5.2.

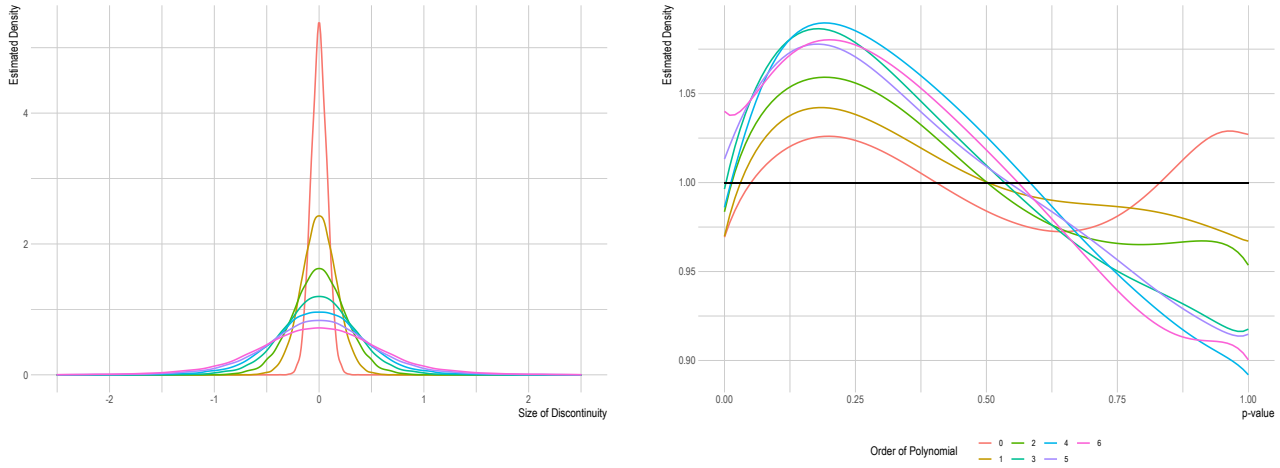

Figure 33: Estimated Densities for Discontinuities and p-values by order of Polynomial (outliers close to c)

K	0	1	2	3	4	5	6
correlation coefficient	-0.933	-0.909	-0.882	-0.853	-0.822	-0.798	-0.778

Table 9: Correlation between $|\hat{\tau}|$ and their corresponding p-values for different K (outliers close to c)

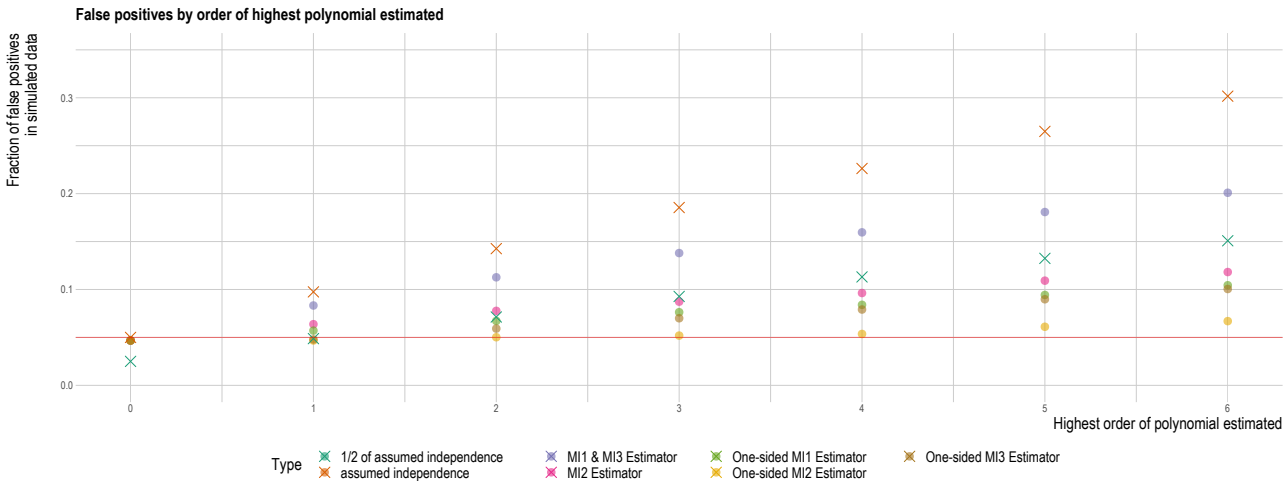

Figure 34: Fraction of False Positives by Maximum Order of Polynomial estimated (outliers close to c)

K/K	0	1	2	3	4	5	6
0	1	0.252	0.0751	0.0346	0.0249	0.0259	0.0234
1	0.613	1	0.376	0.16	0.0976	0.0692	0.0458
2	0.392	0.737	1	0.444	0.235	0.138	0.101
3	0.288	0.541	0.804	1	0.512	0.284	0.183
4	0.227	0.42	0.631	0.844	1	0.553	0.321
5	0.183	0.341	0.518	0.695	0.869	1	0.58
6	0.146	0.285	0.438	0.587	0.736	0.886	1

Table 10: Correlation between $|\hat{\tau}|$ for different K (purple) and p-values of different K (green) (outliers close to c)

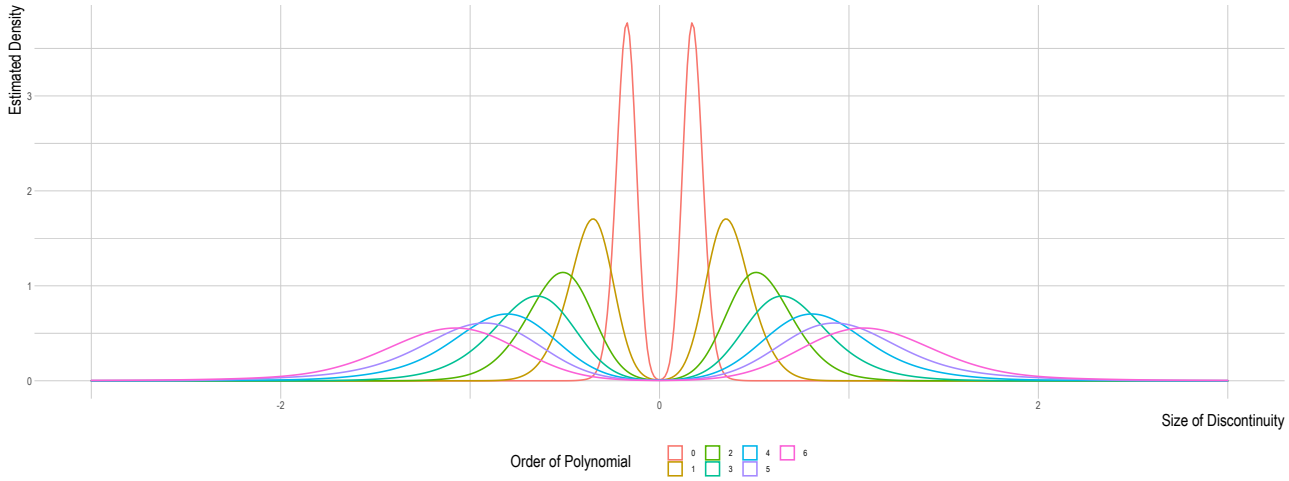

Figure 35: Estimated Densities for significant Discontinuity Estimates (outliers close to c)

Figure 36: Which Estimators do the MI Estimators choose? (outliers close to c)

Figure 37: Which Estimators do the MI Estimators choose? (Only for significant results) (outliers close to c)

Outliers far away from c The following Tables and Figures were created for the process with outliers far away from c described in section 5.2.

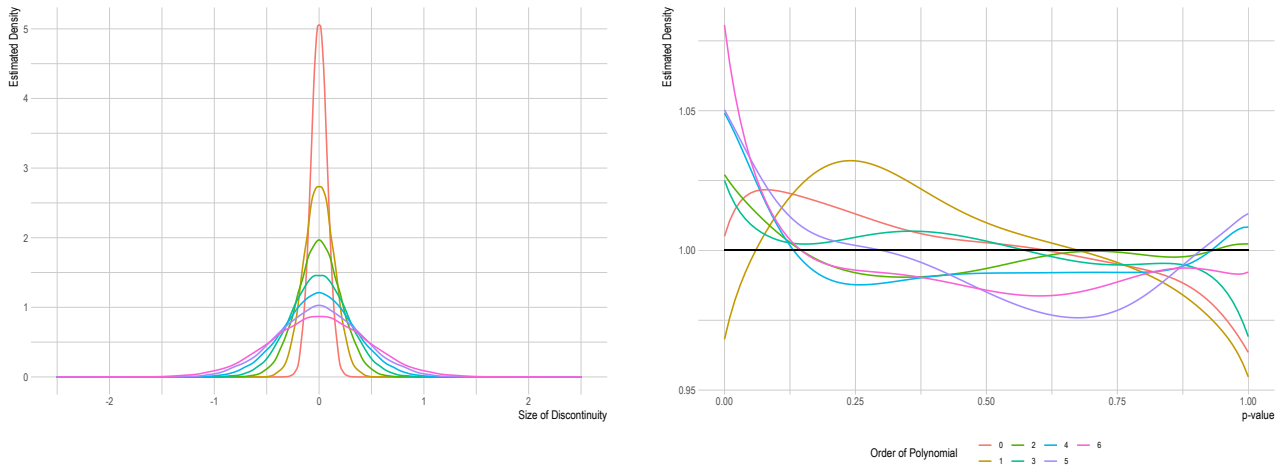

Figure 38: Estimated Densities for Discontinuities and p-values by order of Polynomial (outliers far from c)

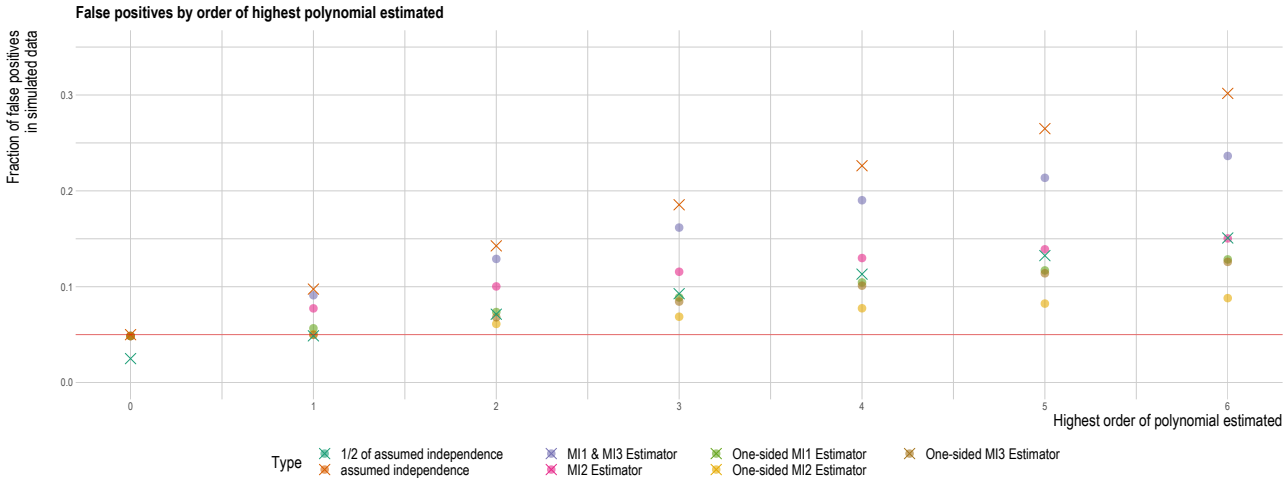

Figure 39: Fraction of False Positives by Maximum Order of Polynomial estimated (outliers far from c)

K/K	0	1	2	3	4	5	6
0	1	0.103	0.0579	0.0325	0.00502	0.009	0.0165
1	0.374	1	0.2	0.114	0.0708	0.0539	0.0522
2	0.269	0.552	1	0.323	0.193	0.147	0.106
3	0.211	0.419	0.656	1	0.392	0.259	0.185
4	0.165	0.327	0.522	0.714	1	0.443	0.297
5	0.141	0.281	0.434	0.597	0.751	1	0.49
6	0.116	0.236	0.362	0.505	0.637	0.779	1

Table 11: Correlation between $|\hat{\tau}|$ for different K (purple) and p-values of different K (green) (outliers far from c)

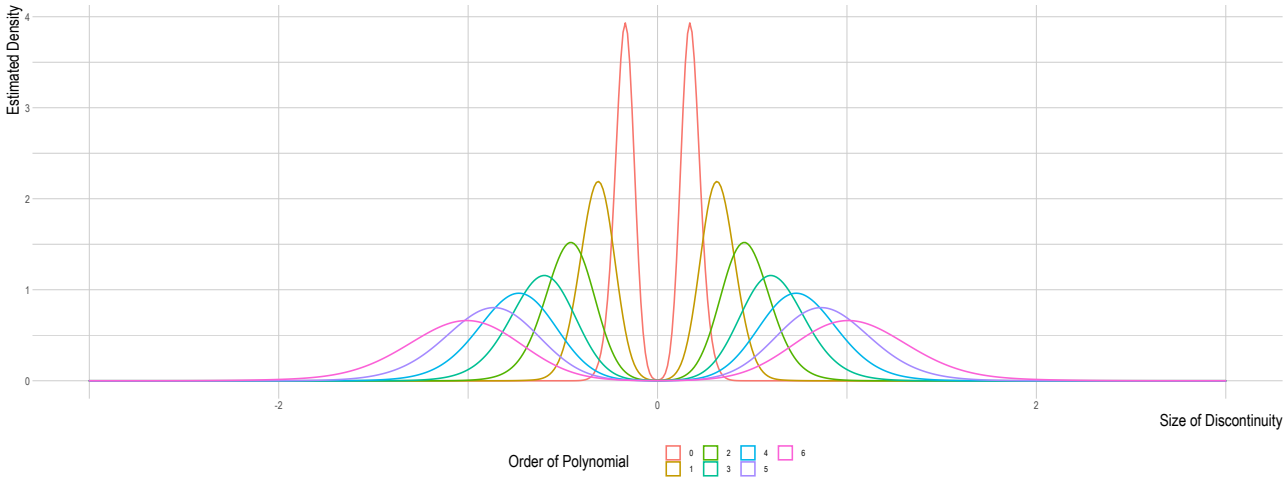

Figure 40: Estimated Densities for significant Discontinuity Estimates (outliers far from c)